SUELEM CRISTINA ALVES

COMPARAÇÃO DE MÉTODOS PARA DEFINIÇÃO DO NÚMERO ÓTIMO DE GRUPOS EM ANÁLISE DE AGRUPAMENTO

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
2012

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

T
A474c
2012

Alves, Suelem Cristina, 1986-
Comparação de métodos para definição do número ótimo de grupos em análise de agrupamento / Suelem Cristina Alves. – Viçosa, MG, 2012.
x, 62f. : il. ; 29cm.

Inclui anexo.
Inclui apêndices.
Orientador: Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Análise por agrupamento. 2. Simulação (Computadores). 3. Modelos não-lineares (Estatística). 4. Curva de crescimento. I. Universidade Federal de Viçosa. II. Título.

CDD 22. ed. 519.53

SUELEM CRISTINA ALVES

COMPARAÇÃO DE MÉTODOS PARA DEFINIÇÃO DO NÚMERO ÓTIMO DE GRUPOS EM ANÁLISE DE AGRUPAMENTO

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: 02 de fevereiro de 2012.

Moysés Nascimento

Cláudio José Borela Espeschit

Sebastião Martins Filho
(Coorientador)

Luiz Alexandre Peternelli
(Orientador)

“De fato, é Deus que desperta em vós a vontade e a ação, conforme a sua benevolência.” (Bíblia; Filipenses 2, 13)

A Deus, minha força.

À minha família, fonte de amor e incentivo.

Às mulheres da minha família, exemplo de força e coragem.

Ao meu namorado Walass, a quem Deus escolheu para estar do meu lado.

DEDICO

AGRADECIMENTOS

Agradeço primeiramente a Deus, que nunca deixou de me amparar e me guiar em todos os meus caminhos. E a constante intercessão de Nossa Senhora por todas as minhas necessidades.

À minha família, pelo apoio e amor sempre constantes, que me impulsionaram ao longo dessa jornada. As palavras de ânimo e incentivo foram fundamentais em todas as etapas. Obrigada por tudo!

Ao meu namorado Walass, pelo carinho, paciência e compreensão em todos os momentos. Obrigada por sempre cuidar de mim (nós), querendo me fazer mais feliz e por ser verdadeiro presente de Deus em minha vida.

Ao meu orientador Luiz Alexandre Peternelli, por me conduzir pelos caminhos da pesquisa, sempre acreditando em mim e me incentivando a seguir em frente. Seu estímulo foi essencial em todas as horas.

Ao "Povo da Capela", especialmente o GOU Morada do Espírito Santo, que me acolheu e chamou a "avançar para águas mais profundas" (Lc 5, 4). O companheirismo e as orações foram fundamentais para essa jornada se tornar mais leve e agradável de ser percorrida.

Aos meus amigos de perto e de longe; do mestrado, graduação e república, pelos momentos de estudo e descontração partilhados.

A todos os professores do Departamento de Estatística da UFV, pelos conhecimentos transmitidos e pela solicitude e paciência em auxiliar.

Ao professor Paulo Luiz Souza Carneiro (UESB) pela concessão dos dados utilizados nesse trabalho.

Aos funcionários do DET pela amizade e ajuda prestada.

Aos participantes da banca, que atenderam tão solicitamente ao convite.

Ao programa CAPES/REUNI pela concessão da bolsa de estudos.

A todos que de alguma forma contribuíram para essa conquista.

E novamente a Deus, a quem não me canso de agradecer constantemente, por tudo que me dá e por ter colocado todas essas pessoas em meu caminho, para que eu pudesse crescer e realizar mais esse sonho. Sem elas, a alegria da vitória não seria a mesma.

BIOGRAFIA

SUELEM CRISTINA ALVES, filha de Maria Joaquina Neta, nasceu em 28 de fevereiro de 1986, em Divinópolis, MG.

Em 2004 ingressou no curso de Matemática da Universidade Federal de Viçosa.

Em agosto de 2009 iniciou o mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 02 de fevereiro de 2012.

SUMÁRIO

RESUMO

ALVES, Suelem Cristina, M. Sc., Universidade Federal de Viçosa, fevereiro de 2012. **Comparação de métodos para definição do número ótimo de grupos em análise de agrupamento**. Orientador: Luiz Alexandre Peternelli. Coorientadores: Sebastião Martins Filho e Cosme Damião Cruz.

Estudos envolvendo análise de agrupamento hierárquico encontram um problema na hora de determinar o número ótimo de grupos, devido à falta de critérios objetivos. Pesquisas que envolvem o ajuste de modelos não-lineares a dados de crescimento ou de sobrevivência, cujo interesse principal é saber quantas curvas são necessárias para descrever o comportamento dos indivíduos analisados, utilizam dessa técnica. Como forma de auxiliar essa decisão, alguns pesquisadores recorrem aos índices BSS (Between-group Sum of Squares), SPRSQ (Semi-partial R-Squared), RMSSTD (Root Mean Square Standard Deviation), RS (R-Squared) e ao método de Mojena. Entretanto, não se sabe qual deles é a melhor escolha para determinação desse valor. A comparação dessas estatísticas foi o objetivo desse trabalho. Toda a metodologia utilizou o método de Ward para fazer o agrupamento das observações, o modelo de von Bertalanffy para o ajuste das curvas, e uma função própria, baseada na lei dos cossenos e na ideia do Método da Máxima Curvatura Modificado, para calcular o número de grupos indicado pelos índices. No capítulo 1 foi feito o estudo de caso real. O conjunto de dados possuía sete curvas de crescimento animal, que formavam três grupos. Após o agrupamento das estimativas dos parâmetros e o cálculo das estatísticas, foi constatado que apenas o índice SPRSQ apontou o número de grupos correto. Usando uma função que re-escalona o eixo dos índices de acordo com o eixo do número de grupos, para melhorar os resultados obtidos, apenas o RMSSTD não indicou o valor esperado. O capítulo 2 descreve o uso da simulação para descobrir qual das estatísticas citadas possuía maior porcentagem de acerto quanto à determinação do número ótimo de grupos em dois cenários. No primeiro, as observações provinham de uma única curva geradora e no outro, os indivíduos pertenciam a três populações diferentes. Para o caso de uma única curva, o índice RS foi o que apontou o número ótimo de grupos na maioria dos casos. Para o cenário onde se possuía três populações diferentes, o método de Mojena foi o que acertou o

número de grupos mais vezes. Nesses cenários, o uso da função que re-escalona os eixos não mostrou eficiência para melhorar a porcentagem de acertos dos índices. De modo geral, os índices RS e SPRSQ mostraram-se os mais indicados para auxiliar na determinação do número ótimo de grupos.

ABSTRACT

ALVES, Suelem Cristina, M. Sc., Universidade Federal de Viçosa, February, 2012. **Comparison of methods for defining the optimal number of groups in cluster analysis.** Adviser: Luiz Alexandre Peternelli. Co-advisers: Sebastião Martins Filho and Cosme Damião Cruz.

Studies that use hierarchical cluster analysis have a problem in determining the optimal number of groups due to lack of objective criteria. Researches involving the adjustment of nonlinear models to data on growth or survival, in which the main interest is to determine how many curves are needed to describe the behavior of the individuals analyzed, use this technique. Some researchers use indices BSS (Between-group Sum of Squares), SPRSQ (Semi-partial R-Squared), RMSSTD (Root Mean Square Standard Deviation), RS (R-Squared) and Mojena method, as a means of assistance in this decision. However, it is not known which one is the best choice to determine that value. The comparison of these statistics was the aim of this study. The entire methodology used the Ward's method to cluster the observations, the von Bertalanffy model to fit the curves, and a specific function, based on the law of cosines and the idea of the Modified Maximum Curvature Method, to calculate the number of groups indicated by the indices. In chapter 1, a real case study was developed. The data set had seven animal growth curves, forming three groups. After grouping the parameter estimates and the calculation of statistics, it was found that only the index SPRSQ pointed to the correct number of groups. Using a function to re-scale the axis of the indices according to the axis of the number of groups, to improve the results obtained, only RMSSTD did not indicate the expected value. Chapter 2 describes the use of simulation to find out which of the statistics mentioned had the highest percentage of accuracy in determining the optimal number of groups in two cases. In the first one, the observations came from a single generator curve and, in the other, the individuals belonged to three different populations. In the case of a single curve, the RS index pointed to the optimal number of groups in most cases. For the case in which there were three different populations, the Mojena method was the one that indicated the right number of groups more often. In these cases, the use of the function that re-scales the axes did not show efficiency to

improve the percentage of correct indices. In general, the indices RS and SPRSQ were the most appropriate to assist in determining the optimal number of groups.

1 INTRODUÇÃO GERAL

Análise de agrupamentos reúne uma gama de técnicas e algoritmos, cujo objetivo é dividir um grupo de observações em grupos homogêneos de acordo com algum critério de similaridade (BUSSAB et al., 1990). É também conhecida como análise de conglomerados, classificação ou *cluster*, e o resultado, após sua utilização, é que elementos pertencentes a um mesmo grupo serão mais similares entre si, enquanto que os pertencentes a grupos diferentes deverão ser heterogêneos entre si em relação a determinadas características (MINGOTI, 2005).

Nos últimos anos essa ferramenta vem sendo amplamente utilizada em estudos de diversas áreas como: psicologia, pesquisa de mercado, ecologia, geografia, ergonomia, geoquímica, entre outras (MINGOTI, 2005). Após a escolha de uma medida de similaridade ou dissimilaridade para se agrupar as observações, utiliza-se alguma técnica de agrupamento hierárquico ou não-hierárquico para a construção dos *clusters*. A partir daí, constrói-se um dendrograma que auxiliará na escolha do número ideal de grupos a ser utilizado. Segundo Khattree e Naik (2000) e Mingoti (2005), não existe uma resposta exata para se escolher esse valor. No entanto, alguns critérios podem auxiliar nessa decisão. Dentre eles estão os índices BSS (Between-group Sum of Squares), SPRSQ (Semi-partial R-Squared), RMSSTD (Root Mean Square Standard Deviation), RS (R-Squared) (SHARMA, 1996; KHATTREE; NAIK, 2000), algumas vezes usados com o Método da Máxima Curvatura Modificado. O método de Mojena (MOJENA, 1977) também tem sido utilizado.

Entretanto, a literatura não apresenta como proceder para melhor fazer a escolha do número de *clusters*, qual o melhor índice ou aquele que deve ser utilizado de acordo com alguma característica específica do conjunto de dados. Em alguns casos, a análise visual do dendrograma aliada a certo conhecimento do pesquisador é que muitas vezes prevalece na escolha do número ótimo de grupos a ser trabalhado, o que caracteriza a subjetividade na decisão.

Nesse trabalho, procurou-se avaliar qual estatística citada conseguia melhor determinar o número ótimo de grupos. Os valores obtidos pelos índices em cada passo do agrupamento foram colocados em um gráfico. O número ótimo foi encontrado a partir do cálculo do maior ângulo formado com as semirretas obtidas pela união desses pontos.

Em um primeiro momento foi feito um estudo com curvas de crescimento com dados reais, cujo número de grupos já era previamente determinado. O segundo passo foi fazer simulações, para confrontar os resultados obtidos com o encontrado anteriormente.

Com base nisso, o presente trabalho teve como objetivos: 1) agrupar curvas de crescimento ajustadas de um conjunto de dados reais, em que o número de grupos já era determinado e avaliar qual índice ou método indicava o valor correto para o caso estudado; e 2) comparar as diversas estatísticas citadas por meio da simulação de dois cenários, procurando encontrar, nas análises feitas, aquela que melhor indica o número ótimo de *clusters*, tendo como base um número de grupos pré-fixado.

2 REFERENCIAL TEÓRICO

2.1 Modelos não-lineares

A regressão é uma técnica estatística muito usada na análise de dados, pois permite averiguar e modelar a relação entre as variáveis dependentes e as independentes. Em muitos casos os modelos lineares não explicam satisfatoriamente o comportamento das observações em estudo. Isto motivou os pesquisadores a encontrar uma nova classe de modelos, os não-lineares, onde não são satisfeitas as condições de linearidade dos parâmetros.

O processo de crescimento dos seres vivos possui a característica de rápido desenvolvimento na idade inicial que, com o passar do tempo, tem um declínio na taxa de crescimento, estabilizando na idade adulta. Assim, as funções curvilíneas assintóticas podem representar melhor a relação entre as variáveis. Desse modo, os modelos não-lineares explicam bem o processo de crescimento.

Um modelo é chamado não-linear quando ele não é linear em relação aos parâmetros, ou seja, se há alguma derivada parcial em relação aos parâmetros que é função de algum parâmetro desconhecido, e não pode ser linearizado por meio de transformações.

Considere a equação abaixo, que representa a relação entre a variável resposta e a variável independente:

$$y_i = f(x_i, \boldsymbol{\theta}) + \varepsilon_i \,,$$

onde y_i representa a observação da variável dependente; i=1,...,n, $f(x_i, \boldsymbol{\theta})$ é a função de regressão, ou função resposta; x_i representa a observação da variável independente; $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_p]'$ é um vetor de parâmetros p-dimensional desconhecidos; ε_i representa o efeito do erro aleatório não observável, ε_i ~ NIID (0, σ^2).

Assim sendo, para saber se o modelo acima é não-linear, é necessário descobrir se a derivada $\frac{\partial f}{\partial \boldsymbol{\theta}}$ depende de $\boldsymbol{\theta}$.

Outras vantagens dos modelos não-lineares, em relação aos lineares, é que possuem bom ajuste aos dados com menos parâmetros (parcimônia), e que podem ser interpretados biologicamente (FREITAS, 2005). Esse fato auxilia na interpretação e conhecimento do fenômeno em estudo.

Considere como exemplo o modelo não-linear de von Bertalanffy (1957):

$$y_i = \beta_1\left(1 - \beta_2 e^{-\beta_3 x_i}\right)^3 + \mathrm{e}_i \,,$$

em que $\beta_1$representa o peso adulto, ou peso assintótico do animal; β_3 representa a taxa de maturidade, ou velocidade de crescimento; e β_2 não possui interpretação prática, sendo uma constante de integração.

Para descrever bem o processo de crescimento, o modelo não-linear escolhido deve apresentar pequenos desvios no ajuste em relação aos dados observados, e os parâmetros estarem de acordo com o fenômeno em estudo. Outra característica desejável é que o modelo tenha facilidade de convergência, já que para estimação dos parâmetros são utilizados métodos iterativos, sendo o de Gauss-Newton o mais usado. Em alguns casos o método pode convergir lentamente, ou nem convergir (REGAZZI, 2003).

2.2 Curvas de crescimento

A análise do crescimento de certo animal pode ser realizada pelo estudo de sua curva de crescimento, que deve apresentar uma boa relação entre o peso e a idade. Tais curvas ajudam no conhecimento do ganho de peso e maturidade animal que influenciam na quantidade e qualidade da carne, interferindo diretamente nos custos de produção e no lucro obtido nessa atividade.

Os modelos não-lineares vêm sendo preferíveis para descrever o comportamento do crescimento animal por duas características em especial. A primeira é que estes se ajustam melhor aos dados, descrevendo mais corretamente o crescimento observado, já que o processo de desenvolvimento de seres vivos é caracterizado por uma fase de rápido crescimento que vai diminuindo, até se estabilizar na idade adulta. A segunda é que a maioria dos parâmetros possui uma interpretação biológica, o que é interesse do pesquisador. Essa particularidade auxilia, por exemplo, a identificar numa população de animais aqueles que são mais pesados em idades mais jovens (FREITAS, 2005).

Os parâmetros frequentemente encontrados nos modelos não-lineares são: peso assintótico, ou peso adulto, geralmente representado por β_1; taxa de maturidade, que indica a precocidade do animal, parâmetro β_3; e a constante de integração, β_2, que não possui interpretação biológica. Os outros parâmetros que possam vir a existir

são constantes matemáticas, que auxiliam na determinação da forma da curva, como o β_4. Os modelos que apresentam esse parâmetro possuem ponto de inflexão variável, cuja localização é determinada pelo mesmo, enquanto que nos demais modelos ou o ponto de inflexão é fixo, ou não o possuem (SOUZA, 2010).

Para que um modelo de regressão não-linear descreva de forma satisfatória a relação peso-idade, Fitzhugh Jr. (1976) coloca que os seguintes pontos devem ser atendidos: interpretação biológica dos parâmetros, "alta qualidade" de ajuste e facilidade de convergência. Sobre o último aspecto, sua importância se deve ao fato de que os modelos não-lineares precisam de métodos iterativos para a estimação dos parâmetros.

A respeito do estudo de curvas de crescimento de ovinos, Sarmento et al. (2006) ajustaram os modelos de Gompertz, von Bertalanffy, Brody, Logístico e Richards a dados de idade-peso de cordeiros da raça Santa Inês, verificando que os dois primeiros são os mais indicados para descrever as curvas de crescimento dos grupos avaliados.

Malhado et al. (2009) analisou dentre os mesmos modelos anteriores quais melhores se ajustavam aos dados de cruzamentos de ovinos da raça Dorper com Morada Nova, Rabo Largo e Santa Inês. O Logístico foi preferível devido ao maior percentual de convergência e menor desvio médio absoluto.

Oliveira et al. (2009) comparou o coeficiente de determinação ajustado (R^2) e o percentual de convergência das curvas de Brody, von Bertalanffy, Gompertz e Logístico em dados coletados de machos e fêmeas de caprinos Anglonubianos. O modelo de Brody foi considerado inferior aos demais. Devido ao maior percentual de convergência do modelo Logístico, associado a um alto valor de R^2, esse foi o escolhido pelos autores. Entretanto os outros dois modelos não apresentaram inadequação.

Com animais dessa mesma raça, oriundos de rebanhos de elite e comercial, Malhado et al. (2008) procurou encontrar o modelo não-linear que melhor se ajustava aos dados dentre Logístico, Brody, Richards, Gompertz e von Bertalanffy. O último se mostrou o mais adequado.

O modelo de von Bertalanffy foi estudado junto com outros seis no ajuste de peso-idade de oito espécies: camarão, rã, coelho, ovino, caprino, bovino, suíno e frango (FREITAS, 2005). O autor chegou à conclusão de que apenas para o camarão

a adequação desse modelo aos dados não foi apropriada, considerando os critérios de convergência, coeficiente de determinação e interpretabilidade dos parâmetros.

Silveira (2010), usando os mesmos dados de ovinos cruzados com Dorper utilizados no presente trabalho, chegou a conclusão que os modelos de Richards e von Bertalanffy foram os que apresentaram os melhores ajustes para esses conjuntos. Aliado a isso, mesmo quando escolhido outro modelo para melhor caracterização dos dados, o modelo de von Bertalanffy não se apresentou inadequado nos casos encontrados na literatura.

Como era necessário definir uma única curva para ajuste de cada conjunto de animais para poder se fazer a comparação pelo agrupamento, outras características do modelo citado foram analisadas. Seu alto percentual de convergência e bom ajuste aos diversos tipos de conjuntos de dados relacionados ao crescimento animal, além de possuir dois parâmetros com fácil interpretação biológica, foram outros fatores que contribuíram para a escolha pela curva de von Bertalanffy no presente estudo.

O modelo de von Bertalanffy (VON BERTALANFFY, 1957) é descrito da seguinte maneira:

$$y_i = \beta_1\left(1 - \beta_2 e^{-\beta_3 x_i}\right) + \mathrm{e}_i,$$

onde β_1 representa o peso adulto ou peso assintótico do animal; e β_3 a taxa de maturidade, ou velocidade de crescimento. O parâmetro β_2 é uma constante de integração, sem interpretação biológica. y_i é o peso do animal na idade x_i e e_i representa o efeito do erro aleatório.

2.3 Análise de agrupamento

A análise de agrupamento é utilizada para se obter grupos homogêneos, por algumas técnicas que possibilitem reunir as observações em um determinado número de grupos, de modo que exista grande homogeneidade dentro de cada grupo e heterogeneidade entre eles (JONHSON; WICHERN, 1992; CRUZ; REGAZZI, 1997).

Para se iniciar o processo de agrupamento, determina-se uma medida de proximidade, que irá ajudar a decidir até que ponto dois elementos podem ser considerados similares (MINGOTI, 2005). Pode-se escolher entre as medidas de similaridade ou de dissimilaridade (KHATTREE; NAIK, 2000; MINGOTI, 2005; SHARMA, 1996). A opção por cada uma dessas medidas varia, entre outros

aspectos, com o tipo das variáveis, caso sejam qualitativas, quantitativas, binárias ou multicategóricas (MINGOTI, 2005; CRUZ; CARNEIRO, 2006).

Com relação às medidas de similaridade, quanto maior o valor, mais semelhantes são as observações, enquanto que nas medidas de dissimilaridade um alto valor indica que as observações são menos parecidas. Para medidas de similaridade existem os coeficientes de correlação e coeficientes de associação (SHARMA, 1996). Entretanto, a maioria dos algoritmos trabalha com o conceito de dissimilaridade, ou seja, de distância (MARDIA et al., 1997).

Existem várias técnicas para a construção dos *clusters*, dentre as hierárquicas e não- hierárquicas. Nas não-hierárquicas, como o método das *k*-médias, o número de grupos já é pré-especificado pelo pesquisador (MINGOTI, 2005), o que não ocorre nas hierárquicas. Estas últimas podem ser divididas em aglomerativas ou divisivas.

Nos métodos divisivos todos os objetos pertencem inicialmente ao mesmo grupo, que vai sendo dividido, até que cada observação forme um grupo individualmente (JOHNSON; WICHERN, 1992). Contudo, os métodos aglomerativos são os mais usados e apresentados na literatura. Neles, cada elemento é inicialmente um grupo isolado, que a cada passo do agrupamento vão se unindo, formando um único grupo ao final.

Sharma (1996) e Mingoti (2005) colocam que os métodos hierárquicos aglomerativos mais populares são: o método do centróide; o método do vizinho mais próximo (ou da ligação simples); o método do vizinho mais distante (ou da ligação completa); o método da média das distâncias e o método de Ward. Após o emprego do método escolhido para o agrupamento, um dendrograma é formado.

2.3.1 Dendrograma

O agrupamento é feito usando todas as variáveis disponíveis e, para melhor análise das junções, representado de maneira bidimensional por um dendrograma, que é um diagrama bidimensional em forma de árvore. Nele é possível ver as partições ou fusões feitas em cada nível do processo de agrupamento (JOHNSON; WICHERN, 1992). No eixo *x* são colocados os objetos numa ordem conveniente, enquanto que o eixo *y* representa as distâncias do agrupamento, ou o nível de similaridade ou dissimilaridade (MARDIA et al., 1997; MINGOTI, 2005).

Os ramos das árvores informam a ordem das (*n*–1) ligações, em que o primeiro nível representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos se juntem.

No geral, os dendrogramas apresentam estruturas de agrupamentos de objetos homogêneos. Entretanto, não existe uma resposta exata de como proceder para escolher o número final de grupos (KHATTREE; NAIK, 2000; MINGOTI, 2005), também conhecido como ponto de corte do dendrograma.

2.3.2 Medidas de dissimilaridade

Dentre medidas de dissimilaridade conhecidas, as que mais se destacam são a distância euclidiana, a distância euclidiana média e a distância de Mahalanobis, devido a sua maior utilização (KHATTREE; NAIK, 2000; CRUZ; CARNEIRO, 2006). Para essas medidas, quanto menor seus valores, mais similares são os elementos comparados (MINGOTI, 2005).

Entretanto, outros tipos de distância também são encontrados na literatura (BUSSAB et al., 1990; MARDIA, et al., 1997; MINGOTI, 2005).

Seja X_{ij} a observação no i-ésimo indivíduo para a j-ésima característica, então a distância euclidiana entre o par de indivíduos i e i' é dada por:

$$d_{ii'} = \sqrt{\sum_{j} (X_{ij} - X_{i'j})^2} \ .$$

Como a distância euclidiana sempre aumenta com o acréscimo de variáveis, a distância euclidiana média tem sido usada de forma alternativa (CRUZ; CARNEIRO, 2006). Sua fórmula é dada por:

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

sendo *n* o número de características estudadas.

Quando são usadas essas distâncias, a escala afeta o valor obtido, especialmente no caso de serem quantificadas em diferentes medidas. Desse modo, Cruz e Carneiro (2006) recomendam a padronização dos dados da seguinte forma:

$$x_{ij} = \frac{X_{ij}}{S(X_j)},$$

onde $S(X_j)$ é o desvio-padrão associado à j-ésima característica.

A distância de Mahalanobis é usada quando há um grau de correlação significativo entre os dados estudados. Para seu cálculo, é considerada a média das X_{ij} observações, sendo descrita por:

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

Em que $\boldsymbol{\delta}$ é a matriz gerada a partir das diferenças entre médias de dois progenitores i e i' para uma dada característica j; e $\boldsymbol{\psi}$ é a matriz de variâncias e covariâncias residuais (CRUZ; REGAZZI, 1997).

2.4 Método de Ward

Dentre os possíveis métodos de agrupamentos existentes, um dos mais conhecidos e utilizados é o método de Ward (WARD, 1963). Mingoti (2005) o define como "Mínima Variância", pois tem como objetivo unir as observações em grupos no qual a soma de quadrados é a menor possível, em cada passo do agrupamento (SHARMA, 1996).

Segundo Khattree e Naik (2000), o método de Ward é o mais estatístico dos métodos. A formação dos grupos é feita pela maximização da homogeneidade dentro do grupo. Para isso, o método de Ward procura minimizar a soma de quadrados dentro do grupo, que também é conhecida como a soma de quadrados do erro (ESS) (SHARMA, 1996).

O método afirma que em qualquer estágio de uma análise, a perda de informações que resulta do agrupamento de indivíduos em grupos pode ser medida pela soma total do quadrado dos desvios de todos os pontos em torno da média do grupo para o qual estão contidos. No princípio, têm-se *n* grupos, ou seja, um grupo para cada vetor do componente da base de dados.

Em cada passo dentro da análise, a união de todos os pares possíveis do grupo é considerada e os dois grupos cuja fusão resulte no menor incremento do erro na soma dos desvios quadráticos são combinados.

A cada etapa repete-se o procedimento considerando-se todas as possíveis uniões de grupos, e é escolhido o agrupamento que causa o menor aumento no erro interno do grupo.

Desse modo, se AB é um grupo obtido pela junção do grupo A e B, onde $\overline{Y}_A$, $\overline{Y}_B$ e $\overline{Y}_{AB}$ são os vetores de médias dos grupos A, B e AB, respectivamente, então:

$$ESS_A = \sum_{i=1}^{n_A} (Y_i - \overline{Y}_A)'(Y_i - \overline{Y}_A),$$

$$ESS_B = \sum_{i=1}^{n_B} (Y_i - \overline{Y}_B)'(Y_i - \overline{Y}_B) \quad \text{e}$$

$$ESS_{AB} = \sum_{i=1}^{n_{AB}} (Y_i - \overline{Y}_{AB})'(Y_i - \overline{Y}_{AB}).$$

Em que $\overline{Y}_{AB} = \frac{(n_A\overline{Y}_A + n_B\overline{Y}_B)}{n_A + n_B}$; e n_A, n_B, e $n_{AB} = n_A + n_B$ são o número de indivíduos em A, B e AB, nessa ordem.

O Método de Ward une dois grupos A e B que minimizam o acréscimo em ESS, também chamado de incremento da soma de quadrados (I_{AB}), definido como:

$$I_{AB} = ESS_{AB} - (ESS_A + ESS_B).$$

Ou ainda,

$$I_{AB} = \frac{(n_A.n_B)}{n_A + n_B}(\overline{Y}_A - \overline{Y}_B)'(\overline{Y}_A - \overline{Y}_B).$$

Assim, pela última equação, minimizar o aumento na ESS é equivalente a minimizar a distância entre grupos.

2.5 Estatísticas para encontrar o número ótimo de grupos

Dentre algumas formas para auxiliar na determinação do número ótimo de grupos, existem os índices RMSSTD (Root-mean-square Standard Deviation), BSS (Between-group Sum of Squares), SPRSQ (Semipartial R-Square) e RS (R-Square) (KHATTREE; NAIK, 2000; MINGOTI, 2005; SHARMA, 1996), e o método de Mojena (MOJENA, 1977).

A estatística BSS é a distância entre grupos, usada no Método de Ward, e mede a homogeneidade de grupos unidos.

Para a definição do RS e da SPRSQ, considere as somas abaixo:

i) A soma de quadrados total corrigida, que é dada por:

$$SST_c = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})'(X_{ij} - \bar{X}),$$

em que X_{ij} é o vetor de medidas observadas para o j-ésimo elemento amostral do i-ésimo grupo; $\bar{X}$ é o vetor de médias global; n_i o número de elementos do i-ésimo grupo; e g é o número de grupos formados na partição.

ii) A soma de quadrados total entre grupos, que possui a equação:

$$SSB = \sum_{i=1}^{g} n_i(\bar{X}_{i.} - \bar{X})'(\bar{X}_{i.} - \bar{X}),$$

onde $\bar{X}_{i.}$ é o vetor de médias do i-ésimo grupo; $\bar{X}$, n_i e g são como definidos acima.

A SPRSQ, ou correlação semiparcial, mede a perda de homogeneidade por juntar dois grupos. Esse índice é obtido por:

$$SPRSQ = \frac{BSS}{SST_c},$$

em que SST_c e BSS estão definidos acima.

O RS, conhecido também como coeficiente de determinação, avalia a heterogeneidade do agrupamento formado em um determinado passo. Um valor pequeno significa que os agrupamentos obtidos não são muito diferentes um do outro, já um valor grande representa que os agrupamentos formados a um determinado passo são bastante heterogêneos. Varia entre zero e um.

É calculado pela fórmula:

$$RS = \frac{SSB}{SST_c},$$

em que SSB e SST_c foram equacionados acima.

O índice RMSSTD, ou raiz quadrada do desvio padrão médio, é usado para calcular a homogeneidade do agrupamento (SHARMA, 1996), ou seja, quanto menor seu valor, mais homogêneo é o novo grupo formado.

Esse índice é encontrado pela fórmula:

$$RMSSTD = \sqrt{\frac{1}{p}\sum_{j=1}^{p} s_j^2},$$

em que p é o número de variáveis e s_j^2 são as variâncias para as p variáveis.

2.5.1 O método de Mojena

Para determinação do número ótimo de grupos, Mojena (1977) propôs um método baseado no tamanho relativo dos níveis de fusão no dendrograma. A recomendação é escolher o número de grupos no estágio *j* que primeiro atender a seguinte equação: $\alpha_j > \theta_k$,

em que α_j é o valor de distâncias dos níveis de fusão correspondentes aos estágios *j* (j=1,...,*n*) e θ_k é o valor referencial de corte.

θ_k é dado por: $\theta_k = \bar{\alpha} + k\hat{\sigma}_\alpha$,

sendo $\bar{\alpha}$ a média e $\hat{\sigma}_\alpha$ o desvio-padrão dos valores de α; e *k* é uma constante.

Mojena sugere valores de *k* em torno de 2,75 e 3,50. Entretanto, Milligan e Cooper (1985, apud FARIA, 2009) indicam o valor de *k* = 1,25 para a definição do número ótimo de grupos.

2.5.2 O método da Máxima Curvatura Modificado

O Método da Máxima Curvatura foi proposto por Smith (1938, apud LESSMAN; ATKINS, 1963) para calcular o número ótimo de parcelas em experimentos. Posteriormente foi modificado por Lessman e Atkins (1963).

O Método da Máxima Curvatura, em sua fórmula original, relaciona o coeficiente de variação (CV) e o tamanho da amostra, de acordo com a equação:

$$CV = \frac{a}{X^b}$$

em que *a* e *b* são constantes apropriadas, CV é o coeficiente de variação por unidade básica e X é o número de unidades básicas.

Uma primeira modificação desse método foi feita por Lessman e Atkins (1963). Em estudos com sorgo, eles fizeram o cálculo do coeficiente de variação e estimaram os valores de *a* e *b*, por meio do ajuste de uma equação aos dados. Ao derivarem a equação proposta em relação a X, encontraram as tangentes nos vários pontos da curva. As duas tangentes sucessivas com maior ângulo θ de interseção definem a região de curvatura máxima, onde a taxa de mudança do CV é a maior em relação aos aumentos de X.

Entretanto, Meier e Lessman (1971), procurando consertar um viés que o método modificado por Lessman e Atkins (1963) possuía em relação a menores

valores de X, propuseram uma mudança em relação à antiga fórmula para se encontrar o valor do número ótimo de parcelas (X_c).

Desse modo, com a nova modificação proposta por Meier e Lessman (1971), a equação final do Método da Máxima Curvatura Modificado que permite obter o valor de X_c, ou o tamanho ótimo de grupos, é dada por:

$$X_c = \left[\frac{a^2 b^2 (2b+1)}{b+2}\right]^{\frac{1}{(2b+2)}} .$$

onde X_c é o ponto de máxima curvatura da trajetória do CV em função do aumento do número de grupos (X), e *a* e *b* são como definidos anteriormente.

Faria (2009) fez uma adaptação do Método da Máxima Curvatura Modificado (MMCM) para descobrir o número ótimo de grupos de acordo com a trajetória dos índices RMSSTD e RS. A autora usou para esse valor o que o MMCM indicava para número ótimo de parcelas.

Após o cálculo dos valores dos índices, constrói-se um gráfico que representa o comportamento dos mesmos em função do número de grupos. Em seguida, é ajustada uma equação para encontrar os valores das constantes *a* e *b*. Com esse resultado em mãos é possível determinar o X_c, que corresponde ao número ótimo de grupos.

2.5.3 Função usada para determinar o número ótimo de grupos

A função desenvolvida para o cálculo do número ótimo de grupos baseou-se na ideia do Método da Máxima Curvatura Modificado (MEIER; LESSMAN, 1971) que diz que o ponto onde as duas tangentes sucessivas possuem o maior ângulo será o ponto de máxima curvatura e, desse modo, o número ótimo de parcelas. Nesse trabalho, esse número ótimo representa o número de grupos. O *script* da função, desenvolvido no R (R DEVELOPMENT CORE TEAM, 2010), encontra-se no Apêndice A.

É sabido dos estudos em trigonometria que em um triângulo o maior lado se opõe ao maior ângulo. Assim, ao unir os pontos obtidos pelo cálculo dos índices, foram formados triângulos e calculados os ângulos dos pontos em questão pela lei dos cossenos (deduzida no Apêndice B). A lei dos cossenos é usada em triângulos que não são retângulos e é dedutível a partir das relações trigonométricas e do

Teorema de Pitágoras. Para determinar o valor da medida do lado *a* no triângulo abaixo, por exemplo, temos a seguinte lei de formação:

$$a^2 = b^2 + c^2 - 2bc\cos\theta.$$

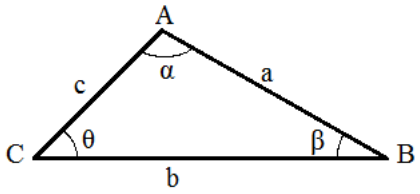

Figura 1 – Triângulo ACB, de lados *a*, *b*, *c* e ângulos α, β e θ, de onde se deduz a lei dos cossenos.

Essa relação é válida para qualquer lado, substituindo-se os valores pertinentes dos lados e do ângulo. Como o interesse era a medida do ângulo, foi usada a função inversa do cosseno para se chegar ao resultado.

Desse modo, o ponto que possuía o maior ângulo replementar foi o escolhido como ponto para determinar o número ótimo de grupos.

Uma ilustração é dada pela Figura 2, relativa ao gráfico do índice RMSSTD para o conjunto de dados reais usado no capítulo 1. Do lado esquerdo encontra-se o gráfico do índice feito com os valores calculados após a análise de agrupamento. A figura do lado direito mostra os triângulos formados pela junção de todos os pontos (triângulos $1\hat{2}3$, $2\hat{3}4$, $3\hat{4}5$ e $4\hat{5}6$). No caso do ponto três, não foi usado o ângulo replementar, mas o ângulo calculado pela lei dos cossenos.

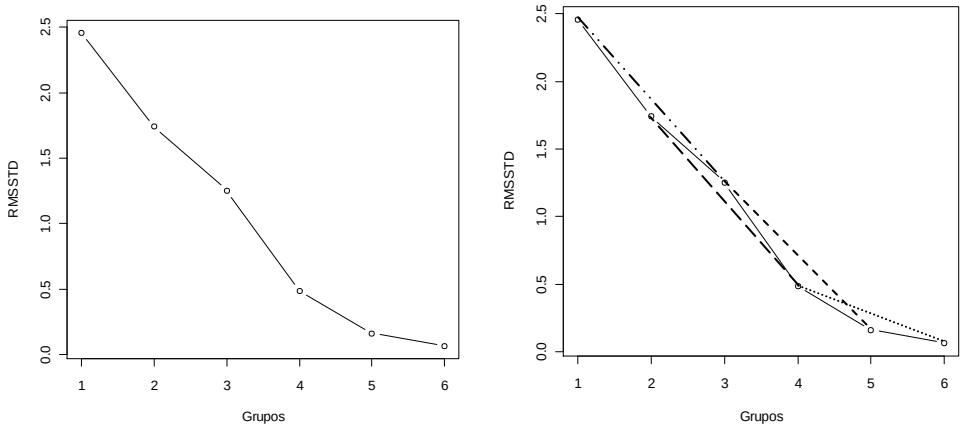

Figura 2 – Gráfico do índice RMSSTD, usando o conjunto de dados do capítulo 1, e dos triângulos formados pela junção dos pontos relativos ao número de grupos indicado pelo índice, respectivamente.

REFERÊNCIAS

BUSSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 87p.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. Vol. 2. Viçosa: UFV, 2006. 585p.

CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. 2 ed. Viçosa: UFV, 1997. 390p.

FARIA, P.N. **Avaliação de métodos para determinação do número ótimo de** *clusters* **em estudos de divergência genética entre acessos de pimenta**. 2009. xi, 54p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2009.

FREITAS, A.R. Curvas de Crescimento na Produção Animal. **Revista Brasileira de Zootecnia**, Viçosa, v.34, p.786-795, May./Jun., 2005.

FITZHUGH Jr., H.A. Analysis of growth curves and strategies for altering their shape. **Journal of Animal Science**, Champaign, v. 42, n. 4, p.1036-1051, Apr., 1976.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 3rd ed. New Jersey: Englewood Cliffs, 1992. xiv, 642p.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS software**. New York: John Wiley and Sons, 2000.

LESSMAN, K. J.; ATTKINS, R. E. Optimum plot size and relative efficiency of lattice designs for grain sorghum yield tests. **Crop Science**, Madison, v. 3, n. 5, p. 477-481, Nov./Dec., 1963.

MALHADO, C.H.M.; CARNEIRO, P.L.S.; CRUZ, J.F.; OLIVEIRA, D.F.; AZEVEDO, D.M.M.R.; SARMENTO, J.L.R. Curvas de crescimento para caprinos as raça Anglo-Nubiana criados na caatinga: rebanho de elite e comercial. **Revista Brasileira de Saúde e Produção Animal**. Salvador, v. 9, n.4, p. 662-671, out./dez. 2008.

MALHADO, C.H.M.; CARNEIRO, P.L.S.; AFFONSO, P.R.A.M.; SOUZA Jr., A.A.O.; SARMENTO, J.L.R. Growth curves in Dorper sheep crossed with the local Brazilian breeds, Morada Nova, Rabo Largo, and Santa Inês. **Small Ruminant Research**. v. 84, p.16-21, Jun. 2009.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis**. New York: Academic Press, 1997.

MEIER, V.D.; LESSMAN, K.J. Estimation of optimum field plot shape and size for testing yield in *Crambe abyssinica* Hochst. **Crop Science**, Madison, v. 11, n. 5, p. 648-650, Sep./Oct., 1971.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005. 297p.

MOJENA, R. Hierarquical grouping methods and stopping rules: an evaluation. **The Computer Journal**, London, 20(4), p. 359-363, 1977.

OLIVEIRA, D.F.; CRUZ, J.F.; CARNEIRO, P.L.S.; MALHADO, C.H.M.; RONDINA, D.; FERRAZ, R.C.N.; TEIXEIRA NETO, M.R. Desenvolvimento ponderal e características de crescimento de caprinos da raça Anglonubiana criados em sistema semi-intensivo. **Revista Brasileira de Saúde Produção Animal**, Salvador, v.10, p.256-265, abr./jun., 2009.

R DEVELOPMENT CORE TEAM (2010). **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (Acesso em 2010)

REGAZZI, A.J. Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão não-linear. **Revista Ceres**, Viçosa, v. 50, n.287, p. 9-26, jan./fev. 2003.

SARMENTO, J.L.R.; REGAZZI, A.J.; SOUZA, W.H.; TORRES, R.A.; BREDA, F.C.; MENEZES, G.R.O. Estudo da curva de crescimento de ovinos Santa Inês. **Revista Brasileira de Zootecnia**, Brasília, v. 35, n. 2, p. 435-442, 2006.

SHARMA, S. **Applied multivariate techniques**. New York: John Wiley and Sons, 1996.

SILVEIRA, F.G. **Classificação multivariada de modelos de crescimento para grupos genéticos de ovinos de corte.** 2010. xi, 61p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2010.

SOUZA, L.A. **Avaliação do crescimento de ovinos da Raça Morada Nova sob modelos não lineares convencionais e alternativos.** 2010. 53p. Dissertação (Mestrado em Zootecnia) – Universidade Estadual do Sudoeste da Bahia, Itapetinga, 2010.

VON BERTALANFFY, L. Quantitative laws in metabolism and growth. **The Quarterly Review of Biology**, Chicago, v. 32, n. 3, p. 217-231, Sep. 1957.

WARD, J.H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v.58, p.236-244, Mar. 1963.

CAPÍTULO 1

COMPARAÇÃO DE MÉTODOS NA DETERMINAÇÃO DO NÚMERO ÓTIMO DE GRUPOS EM UM CONJUNTO DE CURVAS DE CRESCIMENTO AJUSTADAS A DADOS REAIS

RESUMO

O objetivo desse estudo foi avaliar o desempenho dos índices RMSSTD (Root Mean Square Standard Deviation), RS (R-Squared), SPRSQ (Semi-partial R-Squared) e BSS (Between-group Sum of Squared), e do método de Mojena em determinar o número ótimo de grupos de um conjunto de dados cujo número de grupos já era previamente definido e igual a três. As observações provinham de cinco experimentos, totalizando sete cruzamentos ou raças, em que se fez a medição do ganho de peso em caprinos e ovinos. Foi feito o ajuste dos dados de pesagem para cada animal de acordo com o modelo de von Bertalanffy. A seguir foi feita a média das estimativas dos parâmetros associados aos animais dentro de cada conjunto, encontrando-se, assim, sete curvas. Utilizou-se o método de Ward para fazer a análise de agrupamento usando essas médias. Com base nesse resultado foi aplicado o método de Mojena e calculado o valor dos índices citados em cada passo do agrupamento. Usando uma função própria, segundo a lei dos cossenos, e baseada na ideia do Método da Máxima Curvatura Modificado, foi determinado o número de *clusters* que cada um dos índices indicava. O Método da Máxima Curvatura Modificado original também foi aplicado a seguir, a título de comparação com os resultados encontrados anteriormente. O índice que informou corretamente o número de grupos foi o SPRSQ, sendo que o BSS e o RMSSTD foram os que indicaram um número de grupos com maior discrepância da realidade. Quando usada uma função que coloca o valor dos índices na mesma escala do número de grupos, apenas o RMSSTD não apontou o número correto de *clusters*. Ao usar o Método da Máxima Curvatura Modificado nenhum dos índices recomendou corretamente o número de grupos inicial, permanecendo o BSS com maior diferença.

Palavras-chave: análise de agrupamento; von Bertalanffy.

1 INTRODUÇÃO

O estudo do crescimento animal é de grande interesse de pesquisadores e produtores, pois ao se saber de características próprias do ganho de peso e a rapidez com que isso acontece pode-se aumentar a lucratividade do produtor; ou se descobrir cruzamentos ou raças que são superiores de acordo com determinado atributo (SOUZA; BIANCHINI SOBRINHO, 1994).

Ao se fazer o ajuste das curvas de crescimento dos animais, muitas vezes tem-se interesse em se descobrir aqueles que podem ser unidos em um mesmo grupo de acordo com as características estudadas. O procedimento usado para tal finalidade é a análise de agrupamento. Essa técnica tem como objetivo dividir um conjunto de observações em grupos mais homogêneos de acordo com algum critério de similaridade (BUSSAB *et al.*, 1990).

No resultado final da análise de agrupamento temos que elementos pertencentes a um mesmo grupo sejam mais similares entre si, enquanto que os pertencentes a grupos diferentes deverão ser heterogêneos entre si em relação a determinadas características (MINGOTI, 2005; SHARMA, 1996).

Nos últimos anos essa ferramenta vem sendo amplamente utilizada em diversas áreas, como: psicologia, pesquisa de mercado, ecologia, geografia, ergonomia, geoquímica, entre outras (MINGOTI, 2005). Após a escolha de uma medida de similaridade ou dissimilaridade para se agrupar as observações, utiliza-se alguma técnica de agrupamento para a construção dos *clusters*. A partir daí, um dendrograma é construído para auxiliar na escolha do número ideal de grupos a ser utilizado.

Segundo Mingoti (2005) não existe uma resposta exata para se escolher esse valor, mas alguns critérios podem auxiliar nessa decisão. Dentre eles estão os índices BSS (Between-group Sum of Squared), RS (R-Squared), SPRSQ (Semi-partial R-Squared) e RMSSTD (Root Mean Square Standard Deviation) (KHATTREE; NAIK, 2000; SHARMA, 1996; MINGOTI, 2005). Maia et al. (2009) fizeram uso dos três primeiros em um estudos com cultivares de bananeiras. Tomaz (2009) também empregou essas estatísticas quando comparou o resultado obtido pelo teste de logrank e o pelo encontrado na análise de agrupamento com curvas de sobrevivência. Já Faria (2009) utilizou o RS e o RMSSTD associados ao Método da Máxima Curvatura em estudos com pimenta.

O método de Mojena (MOJENA, 1977) também tem sido usado. Silva et al. (2011) utilizaram esse método para encontrar o número de agrupamentos final de 12 genótipos de arroz. O mesmo foi feito por Nassi et al. (2003) para descobrirem a origem de um genótipo de ameixa, a partir de características morfo-fenológicas observadas em outras variedades.

Entretanto, a literatura não indica como proceder para se fazer a melhor escolha, qual é o melhor índice ou qual deles deve ser utilizado de acordo com alguma característica específica do conjunto de dados. Em alguns casos, a análise visual do dendrograma, aliada a certo conhecimento do pesquisador, é que muitas vezes prevalece na escolha do número ótimo de grupos a serem trabalhados, o que caracteriza a subjetividade na decisão (KHATTREE; NAIK, 2000).

O presente trabalho procurou avaliar se os índices citados ou o método de Mojena indicava corretamente o número ótimo de grupos em um conjunto de dados no qual essa informação era previamente conhecida.

2 MATERIAIS E MÉTODOS

2.1 Dados

Os dados utilizados nesse estudo foram cedidos pela Universidade Estadual do Sudoeste da Bahia e são referentes a pesagens de ovinos deslanados e caprinos, criados no Nordeste brasileiro, provenientes de cinco conjuntos de dados, conforme caracterizados a seguir.

No primeiro caso, o experimento foi realizado na Estação Experimental de Jaguaquara, pertencente à Empresa Baiana de Desenvolvimento Agrícola S.A. (EBDA), localizada no município de Jaguaquara, BA, microrregião administrativa de Jequié, BA. Foram avaliados os pesos de ovinos dos cruzamentos Dorper x Morada Nova (DMN); Dorper x Rabo Largo (DRL) e Dorper x Santa Inês (DSI), no período de 2003 a 2005. Foram utilizados quatro animais DMN, nove animais DRL e doze animais DSI.

Para o segundo conjunto de dados, o experimento também foi realizado pela EBDA no mesmo local do primeiro, no período de outubro de 2006 a novembro de 2007. Os ovinos eram da raça Morada Nova (MN), totalizando vinte e nove animais.

Os animais do terceiro experimento são caprinos da raça Mambrina (MAM), criados na caatinga, no período mais seco do ano, na Estação Experimental de Caraíba, da EBDA, localizada no município de Uauá, somando dezenove animais.

No quarto caso, o experimento foi conduzido na Fazenda Rancho do Sol, em Tanhaçu, BA, no período de agosto de 2005 a abril de 2006. São 27 caprinos da raça Anglonubiano (AN).

O último conjunto se caracteriza por quarenta e cinco animais do cruzamento de ovinos Santa Inês x Texel (SIT) em um experimento realizado na Fazenda Provisão, localizada no município de Jequié, BA, no período de março de 2006 a agosto de 2007.

Em todos os casos, foram considerados apenas os animais cujo modelo apresentou convergência para as estimativas dos parâmetros; aqueles que não possuíam estimativas para valores de β_1 (peso assintótico) muito discrepantes, ou seja, muito diferente dos encontrados na prática, e os que possuíam mais de 200 dias no número de pesagens. Por todas as razões citadas, o número de animais utilizado foi menor do que o contido em cada conjunto de dados original.

Para a estimativa final da curva de crescimento de cada grupo de animais, foi utilizada a média dos parâmetros encontrados para cada indivíduo dentro de cada cruzamento ou raça.

Os grupos foram compostos da seguinte maneira: no Grupo 1 estavam os caprinos da raça Anglonubiano (AN) e Mambrina (MAM), por se tratarem de espécie diferente. O Grupo 2 possuía os cruzamentos de ovinos Santa Inês x Texel (SIT), Dorper x Rabo Largo (DRL) e Dorper x Santa Inês (DSI), por terem ganho de peso e peso final mais semelhantes. O Grupo 3 foi formado pelos ovinos da raça Morada Nova (MN) e o cruzamento dos ovinos Dorper x Morada Nova (DMN). A raça Morada Nova possui ganho de peso e peso adulto inferior, quando comparada com outras raças de ovinos. Seu cruzamento com Dorper também apresentou essa característica, conforme encontrado por Oliveira (2011).

Todas as análises e figuras foram obtidas por meio de comandos próprios realizados no software livre R (R DEVELOPMENT CORE TEAM, 2010).

2.2 Modelo da curva de crescimento

Os modelos não-lineares vêm sendo preferíveis para descrever o comportamento do crescimento animal por duas características em especial. A primeira é que estes se ajustam melhor aos dados, descrevendo mais corretamente o crescimento observado, já que o processo de desenvolvimento de seres vivos é caracterizado por uma fase de rápido crescimento que vai diminuindo, até se estabilizar na idade adulta. A segunda é que a maioria dos parâmetros possui uma interpretação biológica, o que é interesse do pesquisador. Essa particularidade auxilia, por exemplo, a identificar numa população de animais aqueles que são mais pesados em idades mais jovens (FREITAS, 2005).

Para que um modelo de regressão não-linear descreva de forma satisfatória a relação peso-idade, Fitzhugh Jr. (1976) coloca que os seguintes pontos devem ser atendidos: interpretação biológica dos parâmetros, “alta qualidade” de ajuste e facilidade de convergência. Sobre o último aspecto, sua importância se deve ao fato de que os modelos não-lineares precisam de métodos iterativos para a estimação dos parâmetros.

Na literatura, mesmo quando escolhido outro modelo para melhor caracterização dos dados, o modelo de von Bertalanffy não se mostrou inadequado

no ajuste dos casos encontrados. Inclusive em outros estudos que utilizaram os mesmos conjuntos de dados aqui apresentados.

Como era necessário definir uma única curva para ajuste de cada conjunto de animais para poder se fazer a comparação pelo agrupamento, outras características do modelo citado foram analisadas. Seu alto percentual de convergência e bom ajuste aos diversos tipos de conjuntos de dados relacionados ao crescimento animal, além de possuir dois parâmetros com fácil interpretação biológica, foram outros fatores que contribuíram para a escolha pela curva de von Bertalanffy no presente estudo.

O modelo de von Bertalanffy (VON BERTALANFFY, 1957) é descrito da seguinte maneira:

$$y_i = \beta_1\left(1 - \beta_2 e^{-\beta_3 x_i}\right) + \mathrm{e}_i,$$

onde β_1 representa o peso adulto ou peso assintótico do animal; e β_3 a taxa de maturidade, ou velocidade de crescimento. O parâmetro β_2 é uma constante de integração, sem interpretação biológica. y_i é o peso do animal na idade x_i e e_i representa o efeito do erro aleatório.

2.3 Determinação do número ótimo de grupos

Após o cálculo dos índices BSS, RMSSTD, RS e SPRSQ, para a determinação do número ótimo de grupos foi feita uma função própria, no ambiente de programação R (R DEVELOPMENT CORE TEAM, 2010), usando a lei dos cossenos e baseada na ideia do Método da Máxima Curvatura Modificado (MMCM) (MEIER; LESSMAN, 1971). Como o MMCM é muitas vezes empregado para auxiliar na escolha do número de *clusters*, ele foi posteriormente usado.

A expressão para o cálculo do ponto de máxima curvatura, X_c, é dada por:

$$X_c = \left[\frac{a^2 b^2 (2b+1)}{(b+2)}\right]^{\frac{1}{(2+2b)}}$$

onde *a* e *b* são constantes apropriadas, encontradas no ajuste da curva aos dados.

Faria (2009) adaptou o Método da Máxima Curvatura Modificado para descobrir o número ótimo de grupos pelos índices RMSSTD e RS, utilizando o X_c como sendo o ponto de máxima curvatura da trajetória desses índices. Com base nisso, o presente trabalho também utilizou essa metodologia, estendendo-a para os índices SPRSQ e BSS.

A função desenvolvida para o cálculo do número ótimo de grupos baseou-se na ideia do Método da Máxima Curvatura Modificado que diz que o ponto onde as duas tangentes sucessivas possuem o maior ângulo será o ponto de máxima curvatura e, desse modo, o número ótimo de parcelas. Para o presente trabalho, esse número ótimo representa o número de grupos.

Dos estudos em trigonometria, tem-se que em um triângulo o maior lado se opõe ao maior ângulo. Assim, ao unir os pontos obtidos pelo cálculo dos índices inicialmente citados, foram formados triângulos e calculados os ângulos dos pontos em questão pela lei dos cossenos (deduzida no Apêndice B). A lei dos cossenos é usada em triângulos que não são retângulos e é dedutível a partir das relações trigonométricas e do Teorema de Pitágoras. Para determinar o valor da medida do lado *a* no triângulo abaixo, por exemplo, temos a seguinte lei de formação:

$$a^2 = b^2 + c^2 - 2bc\cos\theta.$$

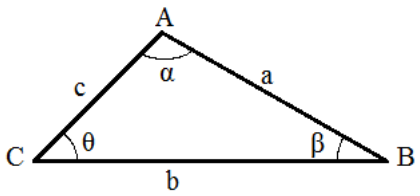

Figura 1 – Triângulo ACB, de lados *a*, *b*, *c* e ângulos α, β e θ, de onde se deduz a lei dos cossenos.

Essa relação é válida para qualquer lado, substituindo-se os valores pertinentes dos lados e do ângulo. Como o interesse era a medida do ângulo, foi usada a função inversa do cosseno para se chegar ao resultado.

Desse modo, o ponto que possuía o maior ângulo replementar foi o escolhido como ponto para determinar o número ótimo de grupos. O *script* desenvolvido no R (R DEVELOPMENT CORE TEAM, 2010) para a função encontra-se no Apêndice A.

A justificativa para criação de uma nova função que pudesse fornecer o número ótimo de grupos veio do fato de que o MMCM muitas vezes não converge na hora de encontrar as constantes *a* e *b*. Isso impossibilita encontrar a curva que descreve os índices e, consequentemente, o número ótimo de grupos.

As diferenças na escala do eixo *x* (eixo do número de grupos) e do eixo *y* (eixo do índice calculado) também podem acarretar em uma interpretação errônea do gráfico e até na determinação do número ótimo de grupos pela função que foi criada

com esse propósito. Buscando contornar tal problema, foi usada uma função para re-escalonar os valores encontrados inicialmente pelos índices de acordo com o número de grupos apontado. A função `rescala` encontra-se no livro de Peternelli e Mello (2011) e está no Anexo 1.

3 RESULTADOS E DISCUSSÃO

O ajuste das curvas de crescimento, usando o modelo de von Bertalanffy, feitas para cada um dos sete genótipos está na Figura 1. Por ela, pode ser observado que não somente o peso assintótico ajustado (beta 1) é fator determinante para que as curvas sejam consideradas pertencentes ao mesmo grupo, mas a taxa de crescimento (beta 3) também influencia. Isso é o esperado ao se fazer a análise de agrupamento onde várias características são medidas e o interesse é saber quais observações pertencem ao mesmo *cluster* (MINGOTI, 2005).

Os ovinos da raça Morada Nova e do seu cruzamento com Dorper apresentaram peso assintótico inferior aos demais cruzamentos. OLIVEIRA (2010) já havia encontrado resultado semelhante quando comparou DMN com DRL e DSI. A autora também verificou que o cruzamento DSI foi superior aos outros dois em relação ao ganho de peso e precocidade, sendo condizente com o resultado observado na Figura 1.

O cruzamento com Dorper favoreceu o desempenho dos animais da raça Morada Nova, como pode ser visto. Malhado et al. (2009) ressaltam a importância de pesquisas nesse campo para favorecer o desempenho dos ovinos e aumentar a qualidade final das carcaças.

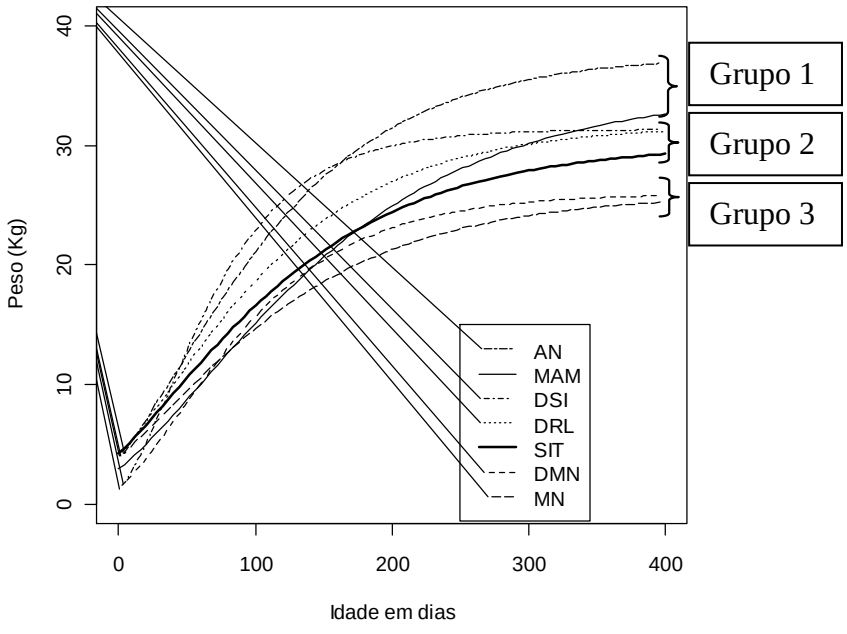

Figura 1 – Ajuste das curvas de crescimento pelo modelo de von Bertalanffy para cada um dos 7 genótipos. Grupo 1: AN e MAM; Grupo 2: DSI, DRL e SIT; Grupo 3: DMN e MN.

As médias das estimativas dos parâmetros calculadas para cada cruzamento ou raça são apresentadas na Tabela 1. Como o número de animais utilizados neste trabalho não foi o mesmo que o usado nos outros estudos que envolviam os mesmos conjuntos de dados (MALHADO et al., 2009; MALHADO et al., 2008a; MALHADO et al., 2008b; CARNEIRO et al., 2009; OLIVEIRA et al., 2009; OLIVEIRA, 2011; SILVEIRA, 2010; SOUZA, 2010), os valores encontrados não puderam ser rigorosamente comparados. Entretanto, estão de acordo com a realidade.

A justificativa para não terem sido utilizados animais com menos de 200 dias no número de pesagens vem do estudo de Alves et al. (2011). Os autores compararam o agrupamento de curvas ajustadas ao mesmo conjunto de dados de ovinos quando consideradas apenas as pesagens com menos de 200 dias com o agrupamento gerado por todas as pesagens que se possuía, que passavam dos 200 dias, encontrando diferenças significativas na configuração final dos grupos formados.

Tabela 1 – Média dos parâmetros estimados pelo modelo de von Bertalanffy para as curvas de crescimento dos sete genótipos estudados.

Grupo	Beta 1	Beta 2	Beta 3
AN	37,62894	0,5351360	0,011142815
MAM	34,55781	0,5618056	0,008474632
DSI	31,36189	0,6666369	0,019057750
DRL	31,72694	0,5004919	0,011269889
SIT	30,13104	0,4823850	0,009829800
DMN	25,99667	0,6266848	0,013967250
MN	25,94345	0,4698430	0,009969138

Alguns fatos relativos ao parto, ano de nascimento e sexo do animal foram considerados significativos por Malhado et al. (2009) na determinação do parâmetro β_3, que é a taxa de maturidade do animal, nos estudos com os ovinos DMN, DSI e DRL. Para o cruzamento SIT, Malhado et al. (2008b) encontrou diferença para esse parâmetro apenas quanto ao tipo de nascimento.

Carneiro et al. (2009) constatou que, mesmo em animais nascidos mais pesados, além da diferença genética, a alimentação com ou sem suplementos também tem grande influência no peso de caprinos ao longo de seu desenvolvimento. Isso pode resultar em animais com o peso adulto inferior (parâmetro beta 1) ao máximo que se pode obter com o genótipo criado.

As estimativas encontradas para os caprinos da raça Anglonubiana foram muito próximas das encontradas por Malhado et al. (2008a) para todos os parâmetros. Nesse estudo, os autores encontraram evidências de que animais mais precoces possuem menor probabilidade de atingirem pesos elevados à idade adulta.

O dendrograma utilizando o método de Ward aplicado a essas estimativas é mostrado na Figura 2. A altura representa os níveis aos quais foram feitas as fusões em cada passo do agrupamento. A formação de três grupos bastante distintos é observada desde o início, quando as alturas relativas aos passos do agrupamento são bem pequenas.

O primeiro grupo a ser formado foi o de DMN com MN, por apresentarem elementos bem diferentes dos demais quanto aos parâmetros. Já a segunda junção, do DRL com DSI, obteve a inclusão do SIT logo em seguida, confirmando a semelhança desses cruzamentos quanto aos parâmetros estimados. Por fim, mas com uma diferença não muito grande, foi formado o grupo com a união dos caprinos AN e MAM. Esses se assemelham mais aos cruzamentos do segundo grupo do que todos os ovinos entre si (Grupos 2 e 3). Isso se deve ao fato de os cruzamentos do Grupo 2 terem uma taxa de crescimento (beta 1) e peso assintótico (beta 3) superior aos ovinos da raça Morada Nova e de seu cruzamento com Dorper. A diferença apresentada pelo Grupo 3 é maior em relação a todos os outros, sendo a união desse grupo junto aos outros agrupamentos formados o último passo do agrupamento realizado.

Na Figura 3 estão os gráficos dos índices BSS e SPRSQ, respectivamente. Já na Figura 4, encontram-se os gráficos dos índices RS e RMSSTD, para o conjunto de dados analisados. Os índices BSS e SPRSQ medem a homogeneidade ao se fundir grupos, e o RMSSTD quando é formado um novo *cluster*. Para esses, o valor esperado é que seja pequeno, pois assim os agrupamentos formados serão mais homogêneos. Já o RS mede a heterogeneidade do grupo, sendo que seu valor varia de zero a um. O desejado para esse índice é que seu valor seja alto (SHARMA, 1996; KHATTREE; NAIK, 2000).

Ao utilizar a função desenvolvida para o cálculo do número ótimo de grupos em cada índice, os valores encontrados foram: BSS: 5, RMSSTD: 5, RS: 2, SPRSQ: 3. Quando calculado pelo Método da Máxima Curvatura Modificado foi obtido: BSS: 6, RMSSTD: 2, RS: 1, SPRSQ: 2. Os valores foram arredondados para o inteiro superior ao apontado pelo Método.

O método de Mojena (MOJENA, 1977) ao ser aplicado aos dados sugeriu um número de grupos igual a dois. Faria (2009) o descreve como sendo um procedimento "objetivo", pois se baseia nos níveis de fusões do dendrograma. O número de grupos é determinado por um valor referencial de corte, calculado em cada passo do agrupamento.

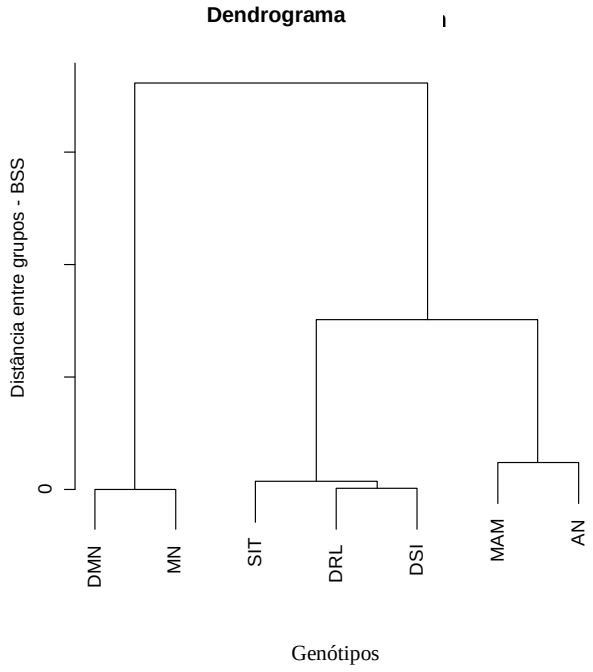

Figura 2 – Dendrograma obtido pelo método de Ward aplicado às médias das estimativas dos parâmetros, pelo modelo de von Bertalanffy, para as curvas de crescimento dos sete genótipos estudados (Tabela 1).

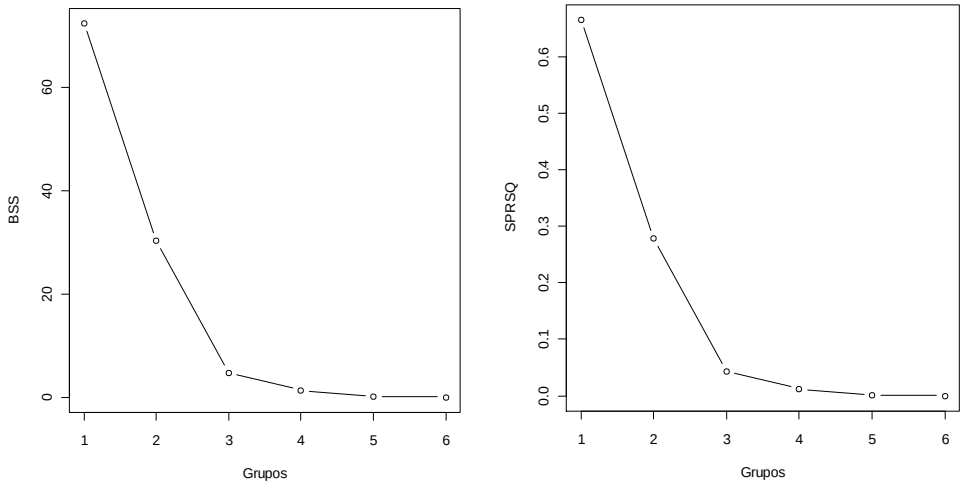

Figura 3 – Gráficos dos índices BSS e SPRSQ, respectivamente, obtidos após o agrupamento e o cálculo do número de grupos com a função criada.

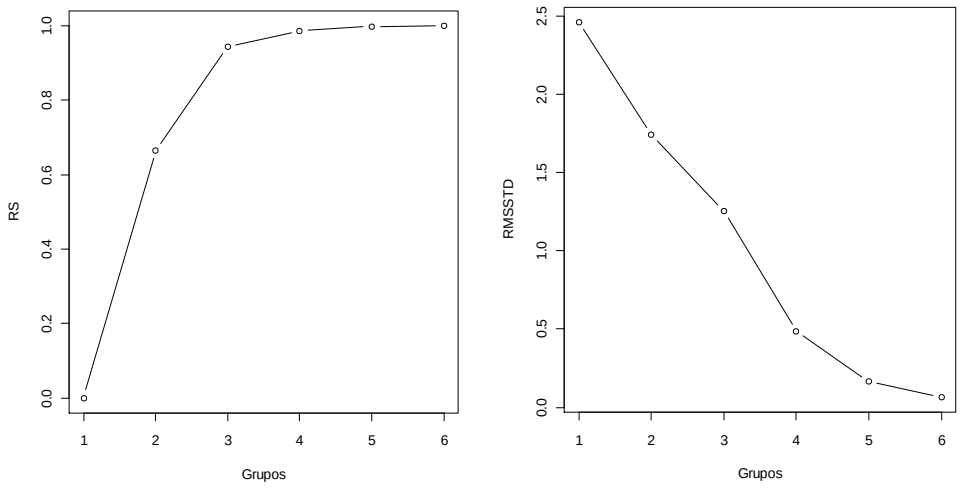

Figura 4 – Gráficos dos índices RS e RMSSTD, respectivamente, obtidos após o agrupamento e o cálculo do número de grupos com a função criada.

A análise visual, apesar de toda subjetividade envolvida no exame do gráfico (FARIA, 2009), muitas vezes é usada para tomada de decisão sobre o número de grupos. Mingoti (2005) indica, inclusive, a análise do gráfico do índice versus o passo do agrupamento. Ao observar um "ponto de salto" grande em relação aos demais, esse seria o momento de parada do algoritmo de agrupamento, indicando o número ideal de grupos. Entretanto, a metodologia escolhida para determinação do número ótimo de *clusters* deve ser a mesma durante todo o trabalho. No presente caso foi o resultado da função feita para calcular o número ótimo de grupos, que leva

em consideração o maior ângulo externo ao das semirretas feitas por dois pontos consecutivos.

Um dos objetivos de usar a função foi tirar a subjetividade presente na determinação do número de grupos. Caso a análise gráfica fosse utilizada para escolher o ângulo levaria ao erro, pois há diferença de escala entre o eixo *x* (número de grupos) e *y* (índice) nos gráficos apresentados pelo R e também por outros softwares. Isso pode ser observado pelo índice BSS, cujo número de grupos calculado pela função foi cinco. Entretanto, ao observar apenas o gráfico apresentado na Figura 3, devido à diferença de escala, a determinação do maior ângulo visualmente seria no ponto três, não correspondendo ao que é pelos cálculos. Esse fato pode ser observado na Figura 5, que apresenta uma ampliação do gráfico do índice BSS no ponto indicado como o de maior ângulo externo entre as duas retas.

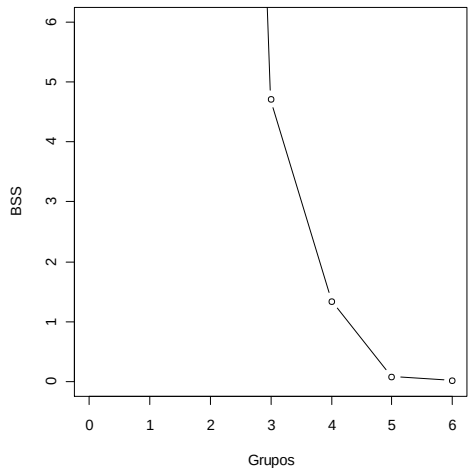

Figura 5 – Gráfico de parte do índice BSS, com enfoque no maior ângulo formado, relativo a 5 grupos.

Buscando contornar o fato de as escalas diferentes acarretarem em uma interpretação errônea, foi usada a função que re-escalona o eixo do índice, de acordo com o eixo dos grupos (PETERNELLI; MELLO, 2011).

A partir das novas medidas encontradas, foi usada novamente a função desenvolvida para calcular o número de grupos que cada índice indicava. Com essa modificação, os índices BSS e RS passaram a informar que o número ideal de *clusters* a ser adotado era três. Para o SPRSQ e RMSSTD os valores de 3 e 5, respectivamente, não foram alterados. Desse modo, a função de re-escalonamento auxiliou na determinação do número de grupos no caso estudado.

As Figuras 6 e 7 mostram os gráficos dos índices versus o número de grupos após o re-escalonamento. Exceto para o caso do índice RMSSTD, caso a análise visual fosse realizada, o maior ângulo externo observado é o equivalente ao ponto três. Desse modo, a função `rescala` poderia auxiliar na determinação do número de grupos para aqueles que ainda preferem a análise gráfica.

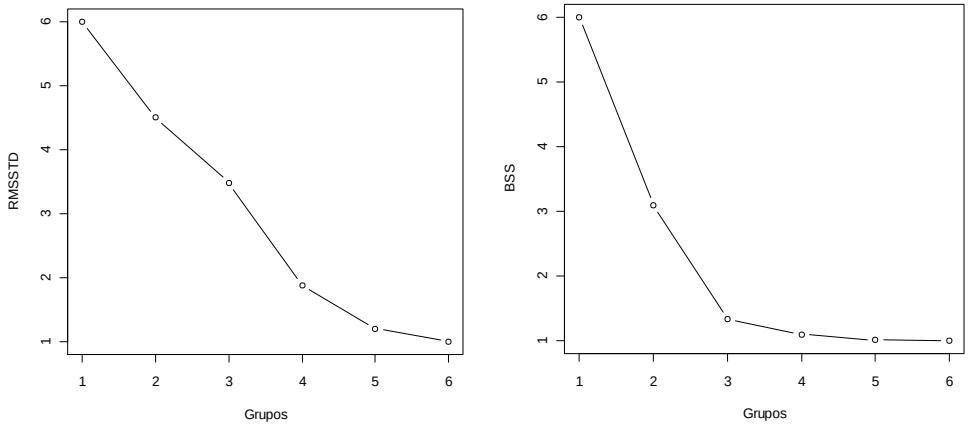

Figura 6 – Gráficos do índice RMSSTD e BSS, respectivamente, obtidos após o agrupamento e o cálculo do número de grupos com a função criada, e com o re-escalonamento do eixo do índice.

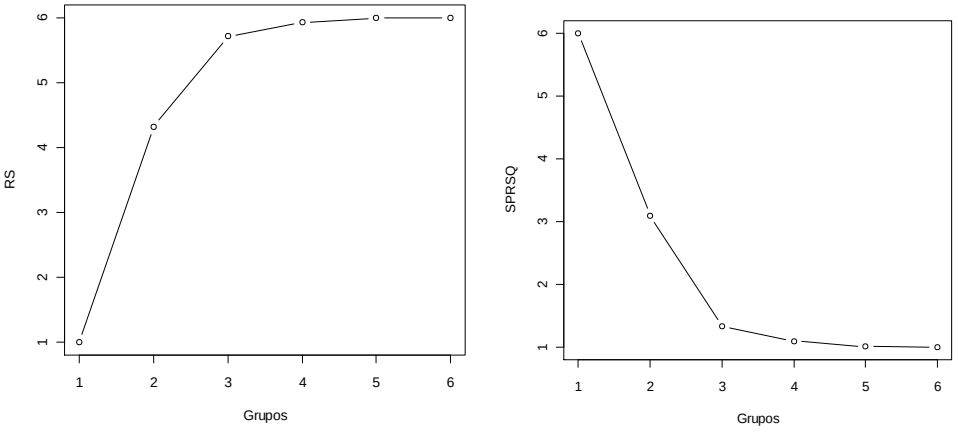

Figura 7 – Gráficos do índice RS e SPRSQ, respectivamente, obtidos após o agrupamento e o cálculo do número de grupos com a função criada, e com o re-escalonamento do eixo do índice.

4 CONCLUSÃO

A aplicação da metodologia proposta para avaliar as estatísticas usualmente utilizadas e o método de Mojena para determinar o número ótimo de grupos no caso estudado identificou que o índice SPRSQ inicialmente indicou o número correto de grupos.

O uso da função que re-escalona o valor do índice auxiliou a determinar o número ótimo de *clusters* no caso estudado para os índices RS, BSS e SPRSQ.

O Método da Máxima Curvatura Modificado não se mostrou adequado para determinação do número de grupos.

REFERÊNCIAS

ALVES, S.C.; PETERNELLI, L.C.; CARNEIRO, P.L.S. **Número de medições no ajuste de curvas de crescimento em ovinos**. 2011. Trabalho apresentado no X Encontro Mineiro de Estatística. São João del-Rei, out. 2011.

BUSSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 87p.

CARNEIRO, P.L.S.; MALHADO, C.H.M.; AFFONSO, P.R.A.M.; PEREIRA, D.G.; SUZART, J.C.C.; RIBEIRO JÚNIOR, M.; ROCHA, J.L. Curva de crescimento em caprinos, da raça Mambrina, criados na caatinga. **Revista Brasileira de Saúde e Produção Animal**, Salvador, v. 10, n.3, p.536-545, jul./set. 2009.

FARIA, P.N. **Avaliação de métodos para determinação do número ótimo de** *clusters* **em estudos de divergência genética entre acessos de pimenta**. 2009. xi, 54p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2009.

FITZHUGH Jr., H.A. Analysis of growth curves and strategies for altering their shape. **Journal of Animal Science**, Champaign, v. 42, n. 4, p.1036-1051, Apr., 1976.

FREITAS, A.R. Curvas de Crescimento na Produção Animal. **Revista Brasileira de Zootecnia**, Viçosa, v.34, p.786-795, May./Jun., 2005.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS software**. New York: John Wiley and Sons, 2000.

LESSMAN, K. J.; ATTKINS, R. E. Optimum plot size and relative efficiency of lattice designs for grain sorghum yield tests. **Crop Science**, Madison, v. 3, n. 6, p. 477-481, Nov./Dec., 1963.

MAIA, E.; SIQUEIRA, D.L.; SILVA, F.F.; PETERNELLI, L.A.; SALOMÃO, L.C.C. Método de comparação de modelos de regressão não-lineares em bananeiras. **Ciência Rural**, Santa Maria, v. 39, n.5, p.1380-1386, ago. 2009.

MEIER, V.D.; LESSMAN, K.J. Estimation of optimum field plot shape and size for testing yield in *Crambe abyssinica* Hochst. **Crop Science**, Madison, v. 11, n. 5, p. 648-650, Sep./Oct., 1971.

MALHADO, C.H.M.; CARNEIRO, P.L.S.; AFFONSO, P.R.A.M.; SOUZA Jr., A.A.O.; SARMENTO, J.L.R. Growth curves in Dorper sheep crossed with the local Brazilian breeds, Morada Nova, Rabo Largo, and Santa Inês. **Small Ruminant Research**. v. 84, p.16-21, Jun. 2009.

MALHADO, C.H.M.; CARNEIRO, P.L.S.; CRUZ, J.F.; OLIVEIRA, D.F.; AZEVEDO, D.M.M.R.; SARMENTO, J.L.R. Curvas de crescimento para caprinos as raça Anglo-Nubiana criados na caatinga: rebanho de elite e comercial. **Revista Brasileira de Saúde e Produção Animal**. Salvador, v. 9, n.4, p. 662-671, out./dez. 2008 a.

MALHADO, C.H.M.; CARNEIRO, P.L.S.; SANTOS, P.F; AZEVEDO, D.M.M.R.; SOUZA, J.C.; AFFONSO, P.R.M. Curva de crescimento em ovinos mestiços Santa Inês x Texel criados no Sudoeste do Estado da Bahia. **Revista Brasileira de Saúde e Produção Animal**. Salvador, v. 9, n.2, p. 210-218, abr./jun. 2008 b.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005. 297p.

MOJENA, R. Hierarquical grouping methods and stopping rules: an evaluation. **The Computer Journal**, London, 20(4), p. 359-363, 1977.

NASSI, M.O.; RUFFA, E.; ME, G.; LEPORI, G.; RADICATI, L. A contribution to the systematics of a piedmontese plum ecotype. **Plant Breeding**, Berlin, 122, p. 532-535, 2003.

OLIVEIRA, D.C. **Funções splines para estudo de curvas de crescimento em ovinos cruzados**. 2011. ix, 57p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2011.

PETERNELLI, L.A.; MELLO, M.P. **Conhecendo o R**: uma visão estatística. Viçosa: Editora UFV, 2011. 185 p.

R DEVELOPMENT CORE TEAM (2010). **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (Acesso em 2010)

SHARMA, S. **Applied multivariate techniques**. New York: John Wiley and Sons, 1996.

SILVA, E.F.; SILVA, V.A.C.; GUIMARÃES, J.F.R.; MOURA, R.R. Divergência fenotípica entre genótipos de arroz de terras altas. **Revista Brasileira de Ciências Agrárias**, Recife, v. 6, n. 2, p. 280-286, abr./jun., 2011.

SILVEIRA, F.G. **Classificação multivariada de modelos de crescimento para grupos genéticos de ovinos de corte.** 2010. xi, 61p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2010.

SOUZA, L.A. **Avaliação do crescimento de ovinos da Raça Morada Nova sob modelos não lineares convencionais e alternativos.** 2010. 53p. Dissertação (Mestrado em Zootecnia) – Universidade Estadual do Sudoeste da Bahia, Itapetinga, 2010.

SOUZA, J.C.; BIANCHINI SOBRINHO, E. Estimativa do peso de bovinos de corte, aos 24 meses, da raça Nelore, usando curvas de crescimento. **Revista Brasileira de Zootecnia**, Viçosa, v. 23, n. 1, p. 85-91, 1994.

TOMAZ, F.S.C. **Análise de agrupamento para a avaliação de identidade de modelos não-lineares em análise de sobrevivência**. 2009. x, 70p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2009.

VON BERTALANFFY, L. Quantitative laws in metabolism and growth. **The Quarterly Review of Biology**, Chicago, v. 32, n. 3, p. 217-231, Sep. 1957.

WARD, J.H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v.58, p. 236-244, Mar. 1963.

CAPÍTULO 2

O USO DE SIMULAÇÃO NA COMPARAÇÃO DE MÉTODOS PARA DETERMINAÇÃO DO NÚMERO ÓTIMO DE GRUPOS EM ANÁLISE DE AGRUPAMENTO

RESUMO

O presente trabalho teve como objetivo analisar quais dentre os índices RMSSTD (Root Mean Square Standard Deviation), RS (R-Squared), BSS (Between-group Sum of Squares) e SPRSQ (Semi-partial R-Squared) e o método de Mojena aquele que se sobressaía na determinação do número ótimo de grupos em análise de agrupamento. Para tanto, foram feitas simulações em dois tipos de cenários: um em que todas as observações vinham de uma mesma curva e no outro em que as observações eram originadas de três curvas diferentes. Após o ajuste do modelo de von Bertalanffy às observações, realizou-se o agrupamento pelo método de Ward a partir das estimativas dos parâmetros encontrados. Com esse resultado, foi feito o cálculo para os índices citados em cada passo do agrupamento e utilizado o método de Mojena. Para determinação do número de *cluster* que cada estatística indicava, utilizou-se uma função própria, segundo a lei dos cossenos, e baseada na ideia do Método da Máxima Curvatura Modificado. Como era determinado de antemão o número de grupos que cada cenário simulado conteria, obteve-se a porcentagem do número de acertos após a resposta apontada pelo índice em todas as simulações. Para o caso em que se havia uma única curva geradora, aquele que mais informou corretamente o número inicial de grupos foi o RS. Já quando o cenário possuía três curvas geradoras para os dados, o método de Mojena foi o que apresentou desempenho mais satisfatório, sendo que o BSS e o RMSSTD foram os que tiveram menor percentual de acertos em ambos os casos, indicando o número de grupos mais distante do real. Uma função que re-escalona o eixo dos índices de acordo com o eixo do número de grupos foi usada posteriormente, procurando melhorar os resultados obtidos. Entretanto, sua aplicação não trouxe resultados satisfatórios para determinação do número ótimo de grupos.

Palavras-chave: curvas de crescimento, modelo de von Bertalanffy.

1 INTRODUÇÃO

A análise de agrupamento envolve uma gama de técnicas e algoritmos, cuja finalidade é encontrar e separar objetos em grupos similares (BUSSAB et al., 1990; JONHSON; WICHERN, 1992; SHARMA, 1996; CRUZ; REGAZZI, 2001). Também conhecida como *cluster analysis*, é amplamente utilizada em diversas áreas da ciência. Tem como objetivo dividir os elementos da amostra em grupos de forma que aqueles pertencentes a um mesmo *cluster* sejam homogêneos de acordo com algumas características medidas e elementos pertencentes a grupos diferentes sejam heterogêneos em relação às mesmas variáveis (MINGOTI, 2005). Existem vários métodos de agrupamento que podem ser utilizados para unir as observações, mas de acordo com Khattree e Naik (2000) o método de Ward (WARD, 1963) é o mais estatístico dentre eles.

Após decidir qual medida de similaridade ou dissimilaridade vai ser usada para fazer o agrupamento e escolher o método, obtém-se o dendrograma (diagrama bidimensional em forma de árvore). Nele são representados os níveis de fusão dos passos do agrupamento (CRUZ; CARNEIRO, 2006). Entretanto, a falta de critérios objetivos para determinar o ponto de corte no dendrograma faz com que a escolha do número ótimo de grupos seja subjetiva, ficando muitas vezes a cargo da experiência do pesquisador.

Alguns índices são usados para auxiliar nessa decisão. Dentre os mais utilizados estão: RMSSTD (Root Mean Square Standard Deviation), RS (R-Squared), BSS (Between-group Sum of Squared) e SPRSQ (Semi-partial R-Squared) (SHARMA, 1996; KHATTREE; NAIK, 2000; MINGOTI, 2005); o método desenvolvido por Mojena (MOJENA, 1977) também tem sido objeto de interesse dos pesquisadores nos últimos tempos, por ser considerado um método "objetivo" (FARIA, 2009).

Em estudos com modelos não-lineares há interesse em descobrir quando o comportamento dos indivíduos pesquisados pode ser explicado por uma única curva ou não. Vários trabalhos com curvas de crescimento ou sobrevivência têm utilizado análise de agrupamento com esse intuito (CARNEIRO et al., 2007; MAIA et al., 2009; FARIA, 2009; TOMAZ, 2009; SOUZA, 2010; SILVEIRA, 2010; OLIVEIRA, 2011; ALVES et al.; 2011). Entretanto, a literatura ainda apresenta uma lacuna quanto à melhor forma de escolher o número ótimo de grupos a ser trabalhado.

Este trabalho teve como finalidade a comparação dos índices citados e o método de Mojena, buscando identificar qual possui a maior porcentagem de acerto quanto à determinação do número ótimo de grupos a partir de dados simulados com curvas de crescimento do modelo de von Bertalanffy (VON BERTALANFFY, 1957).

2 MATERIAIS E MÉTODOS

2.1 Simulação dos dados

Os dados usados nesse trabalho foram obtidos por meio de simulação no software livre R (R DEVELOPMENT CORE TEAM, 2010). A simulação dos dados foi baseada em um conjunto de observações reais, para que sua aplicação prática não ficasse comprometida. Os dados originais eram relativos a pesagens de ovinos e caprinos, em experimentos realizados no nordeste brasileiro.

A simulação procedeu do seguinte modo: foi escolhido o número *x* de dias entre as pesagens e qual seria o total de medições de acordo com o observado pelos dados. Devido a problemas iniciais na convergência do modelo, notado por um número pequeno de medições que é encontrado muitas vezes na prática, esse número foi aumentado posteriormente, resultando ao final em 11 medições, espaçadas em um intervalo de 32 dias. Alves et al. (2011) já haviam observado diferenças entre a estimativa feita no ajuste de curvas com poucos dias de medições e quando as pesagens do animal são realizadas por um período maior de tempo. Essa divergência pode comprometer a análise final, pois gera agrupamentos diferentes.

Dois cenários foram definidos. No cenário 1 foram geradas observações que decorriam de uma única curva. Desse modo todos os animais gerados pertenciam a uma única população. Para o Cenário 2, as observações provinham de três curvas diferentes, representando três grupos genéticos distintos. O modelo não-linear das curvas foi o de von Bertalanffy (VON BERTALANFFY, 1957), descrito por:

$$y_i = \beta_1\left(1 - \beta_2 e^{-\beta_3 x_i}\right) + \mathrm{e}_i,$$

onde β_1 representa o peso adulto ou peso assintótico do animal, e β_3 a taxa de maturidade, ou velocidade de crescimento. O parâmetro β_2 é uma constante de integração, sem interpretação biológica. y_i é o peso do animal na idade x_i e e_i representa o efeito do erro aleatório, $e_i \sim \mathrm{NID}\ (0; \sigma_e^2)$.

A partir das curvas obtidas no ajuste aos dados de pesagem de cada animal real, foi feita a curva média para cada conjunto. Julgou-se pertinente fazer uma variação nos parâmetros das curvas, já que dentro de um mesmo genótipo os animais possuem essa diferença, devido a características próprias de cada um, quando estimada a curva final baseada nos dados de medição. Para isso, os parâmetros foram

considerados seguindo uma distribuição normal, com média igual ao beta estimado e variância σ^2.

A determinação da variância para geração dos parâmetros foi obtida por meio da equação: $\hat{\sigma}^2 = \frac{\Sigma^2}{SQR}$, onde Σ^2 é a variância dos betas do grupo considerado, e SQR é a soma de quadrados residuais do grupo.

Também foi feita a simulação do erro aleatório presente em qualquer experimento. Ou seja, ao final cada animal possuía uma curva geradora diferente, mas advinda da curva média do grupo que pertencia; e cada uma das 11 pesagens possuía o acréscimo de um erro aleatório.

Após a criação dos dados representando o número de pesagem dos animais, fez-se a estimação dos parâmetros da curva para cada um, e em seguida a análise de agrupamento pelo método de Ward, conforme sugerido por Khattree e Naik (2000).

Após o resultado obtido pelo método de Ward, foi calculado o valor de cada índice, e usada uma função própria, baseada na lei dos cossenos e na ideia do Método da Máxima Curvatura Modificado (MEIER; LESSMAN, 1971), para o cálculo do número de grupos que cada um indicava. O método de Mojena também foi realizado. No total foram feitas 1000 simulações para cada cenário. Por fim, observou-se a porcentagem de acertos quanto ao número inicial de grupos.

Abaixo seguem os passos para a simulação e análise dos dados gerados. Todos eram repetidos a cada nova simulação, feita após o cálculo da curva média dos animais reais.

Passo 1: geração da variação dos parâmetros.

Passo 2: simulação do erro associado às pesagens.

Passo 3: "criação" de 40 animais com 11 pesagens cada, de acordo com o novo beta obtido no passo 1 e com a soma do erro do passo 2.

Passo 4: estimação dos parâmetros da curva de cada animal gerado, de acordo com as pesagens simuladas no passo 3.

Passo 5: agrupamento das estimativas do passo 4 pelo método de Ward.

Passo 6: cálculo das estatísticas e do número de grupos que cada uma indicava.

A função própria para calcular o número de grupos a ser trabalhado fundamentou-se na ideia do Método da Máxima Curvatura Modificado (MMCM)

(MEIER; LESSMAN, 1971) para calcular o número ótimo de parcelas em experimentos.

Faria (2009) fez uma adaptação do MMCM para descobrir o número ótimo de grupos de acordo com a trajetória dos índices RMSSTD e RS, usando para esse valor o indicado pelo método para número ótimo de parcelas.

A função desenvolvida usou o maior ângulo externo formado pelas duas retas consecutivas da união de cada ponto dos índices. O valor correspondente no eixo do número de grupos era o do número indicado de agrupamentos. O *script* dessa função, desenvolvido no R (R DEVELOPMENT CORE TEAM, 2010), encontra-se no Apêndice A.

Da trigonometria, tem-se que em um triângulo o maior lado se opõe ao maior ângulo. Desse modo, ao juntar os pontos obtidos pelo cálculo dos índices em cada passo do agrupamento, foram feitos triângulos e calculados os ângulos dos pontos em questão pela lei dos cossenos. A lei dos cossenos é usada em triângulos que não são retângulos e pode ser deduzida a partir das relações trigonométricas e do Teorema de Pitágoras. Para saber o valor do lado *a* no triângulo abaixo, por exemplo, temos a seguinte lei de formação:

$$a^2 = b^2 + c^2 - 2bc\cos\theta.$$

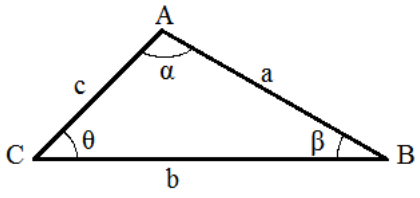

Figura 1 – Triângulo ACB, de lados *a*, *b*, *c* e ângulos α, β e θ, de onde se deduz a lei dos cossenos.

Essa relação é válida para qualquer lado, desde que se substituam os valores pertinentes. Como o procurado era o ângulo, foi usada a função inversa do cosseno para encontrar esse resultado. De tal modo que o ponto que possuía o maior ângulo replementar foi o escolhido como o do número ótimo de grupos.

A justificativa para criação de uma nova função que pudesse fornecer o número ótimo de grupos veio do fato de que o MMCM muitas vezes não converge na hora de encontrar as constantes *a* e *b*. Isso impossibilita encontrar a curva que descreve os índices e, consequentemente, o número ótimo de grupos. Para o caso das

simulações, isso possui o agravante de travar a execução do *script*, pois o retorno é uma mensagem de erro, e não um valor numérico que seria o desejado.

Uma ilustração é dada pela Figura 2, relativa ao gráfico do índice RMSSTD para o conjunto de dados reais que embasou esse trabalho. Do lado esquerdo encontra-se o gráfico do índice feito com os valores calculados após a análise de agrupamento. A figura do lado direito mostra os triângulos formados pela junção de todos os pontos (triângulos $1\hat{2}3$, $2\hat{3}4$, $3\hat{4}5$ e $4\hat{5}6$). No caso do ponto três, não foi usado o ângulo replementar, mas o ângulo calculado pela lei dos cossenos.

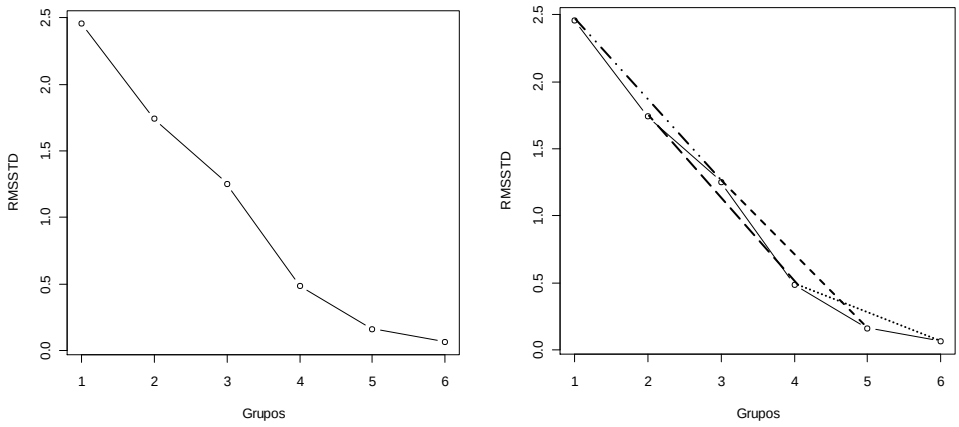

Figura 2 – Gráfico do índice RMSSTD, usando o conjunto de dados do capítulo 1, e dos triângulos formados pela junção dos pontos relativos ao número de grupos indicado pelo índice, respectivamente.

As diferenças na escala do eixo *x* (eixo do número de grupos) e do eixo *y* (eixo do índice calculado) também podem acarretar em uma interpretação errônea do gráfico e até na determinação do número ótimo de grupos pela função que foi criada com esse propósito. Buscando contornar tal problema, foi usada uma função para re-escalonar os valores encontrados inicialmente pelos índices de acordo com o número de grupos apontado. Essa função é descrita por Peternelli e Mello (2011) e está no Anexo 1.

3 RESULTADOS E DISCUSSÃO

Os resultados finais das simulações podem ser observados pelos gráficos abaixo com a porcentagem total de acertos que as estatísticas tiveram. O número de grupos começa no 2 porque a função criada não contempla o ponto 1, pois usa o ângulo entre as semirretas formadas pela junção dos pontos. Para o caso do ponto 1 não há esse ângulo.

Nas Figuras de 3 a 5 estão os gráficos das porcentagens (eixo *y*) do número de grupos (eixo *x*) indicado para cada índice e o método de Mojena, para o Cenário 2, quando os indivíduos vinham de três populações.

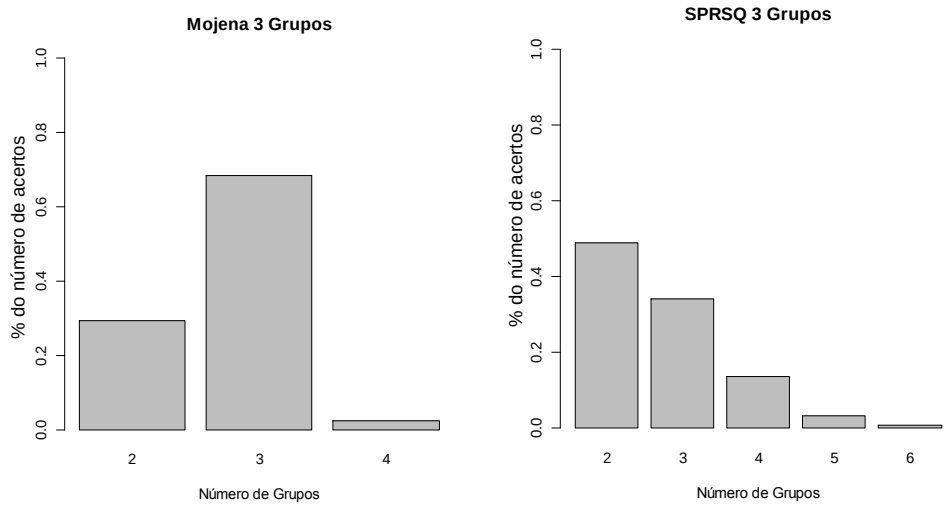

Figura 3 – Gráfico da porcentagem do número de grupos indicado pelo método de Mojena e pelo SPRSQ, respectivamente, no total de 1.000 simulações, quando se tinha 3 populações.

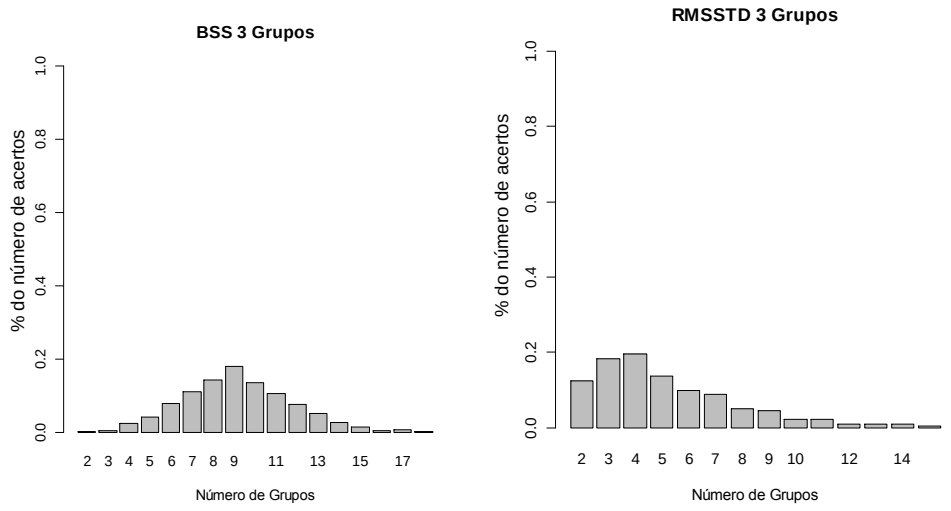

Figura 4 – Gráfico da porcentagem do número de grupos indicado pelo BSS e pelo RMSSTD, respectivamente, no total de 1.000 simulações, quando se tinha 3 populações.

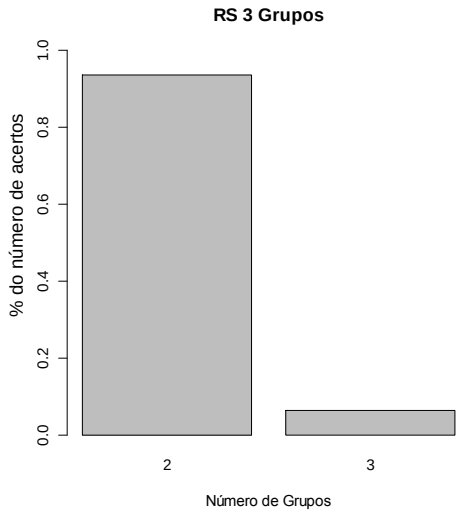

Figura 5 – Gráfico da porcentagem do número de grupos indicado pelo RS, no total de 1.000 simulações, quando se tinha 3 populações.

Nesse cenário, o método de Mojena foi o que mais indicou o número de grupos correto, em 68,3% dos casos. Sua utilização tem acontecido especialmente em trabalhos envolvendo comparação de genótipos de cultivares e os possíveis agrupamentos resultantes.

Silva et al. (2011) utilizaram o método de Mojena para encontrar o número de agrupamentos final de 12 genótipos de arroz. O mesmo foi feito por Nassi et al. (2003) para descobrirem a origem de um genótipo de ameixa, a partir de características morfo-fenológicas observadas em outras variedades.

Entretanto, Faria (2009) comparou o resultado do método de Mojena com o dos índices RMSSTD e RS associados ao Método da Máxima Curvatura Modificado e considerou que metodologia proposta para os índices indicou melhor o número de grupos dos dados de pimenta estudados. Em seu trabalho, a autora encontrou um número de grupos igual a sete por ambos os índices, e igual a três, por Mojena.

Em geral, o método de Mojena não indica altos valores para o número de grupos final. Isso pode ser um problema quando o número de indivíduos trabalhados é grande e há interesse em detectar mais grupos com especificidades inerentes.

As Figuras 6, 7 e 8 mostram os gráficos das porcentagens (eixo *y*) do número de grupos (eixo *x*) que cada índice e o método de Mojena indicou para o Cenário 1, caso em que todas as observações eram geradas por uma única curva. Para esse caso, como a função criada não contemplava o ponto 1, já que não era possível calcular o ângulo, o esperado é que a estatística indicasse 2 grupos criados.

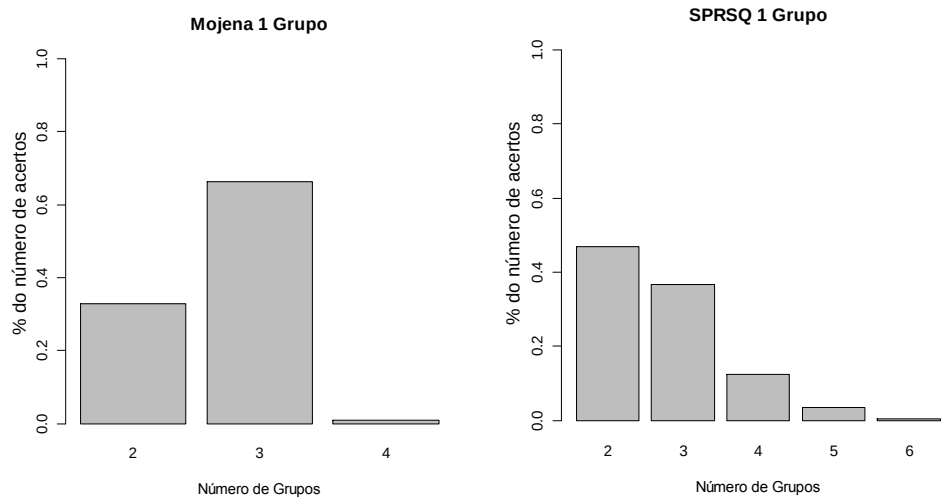

Figura 6 – Gráfico da porcentagem do número de grupos indicado pelo método de Mojena e pelo SPRSQ, no total de 1.000 simulações, quando se tinha 1 população.

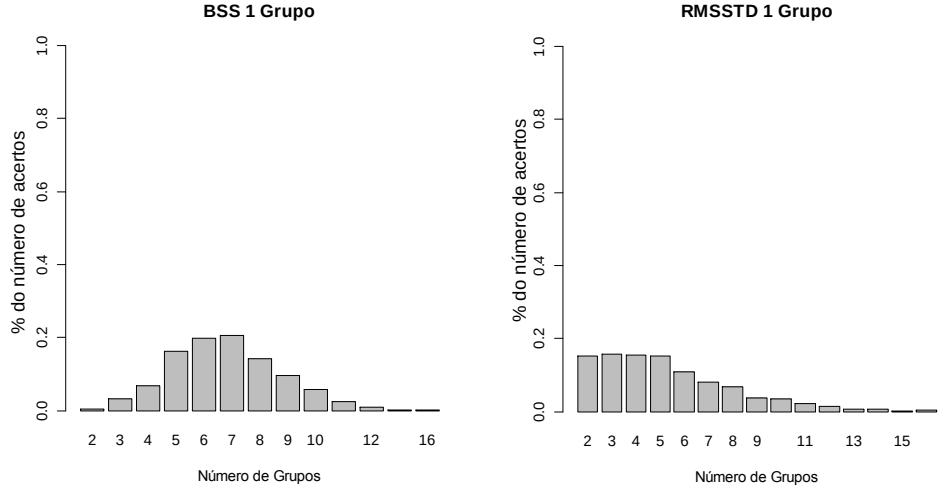

Figura 7 – Gráfico da porcentagem do número de grupos indicado pelo BSS e pelo RMSSTD, no total de 1.000 simulações, quando se tinha 1 população.

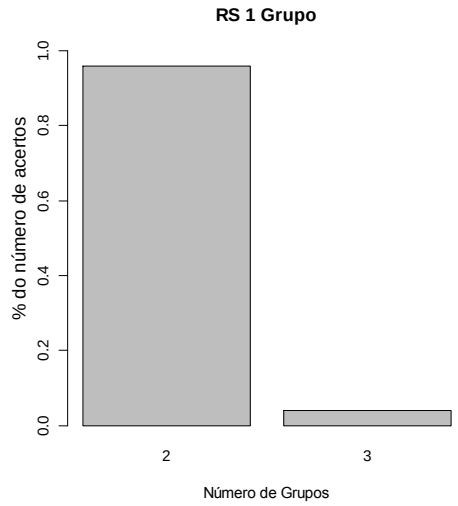

Figura 8 – Gráfico da porcentagem do número de grupos indicado pelo RS, no total de 1.000 simulações, quando se tinha 1 população.

Para esse cenário, o índice que melhor indicou o número ótimo de grupos foi o RS, com 93,6% de acertos, em seguida, o SPRSQ, com 48,9%.

Os índices RS e RMSSTD têm sido muito utilizados em trabalhos de biometria, para o auxílio na escolha do número ótimo de grupos (SILVEIRA, 2010). Faria (2009) usou essas estatísticas associadas ao Método da Máxima Curvatura em estudos com pimenta para encontrar o número ótimo de *clusters*.

Já o BSS é usado apenas quando o método de Ward é aplicado no agrupamento. Tomaz (2009) fez uso dessa estatística, juntamente com o RS e o SPRSQ em seu trabalho, como auxílio na determinação do número de grupos a ser trabalhados. O mesmos fez Maia et al. (2009) em um estudo com cultivares de bananeiras.

A análise visual, apesar de toda subjetividade envolvida no exame do gráfico (FARIA, 2009), muitas vezes é usada para tomada de decisão sobre o número de grupos. Mingoti (2005) indica, inclusive, a análise do gráfico do índice versus o passo do agrupamento. Ao observar um "ponto de salto" grande em relação aos demais, esse seria o momento de parada do algoritmo de agrupamento, indicando o número ideal de grupos. Essa metodologia, no entanto, não seria possível no uso de simulações, devido a sua impraticabilidade pelo número de gráficos gerados.

Um fato a ser considerado é que a análise gráfica pode levar ao erro, caso seja observado o ângulo formado pelas semirretas do índice, pois há diferença de escala entre o eixo *x* (número de grupos) e *y* (índice) nos gráficos apresentados pelo R e

também por outros softwares. Para contornar tal problema, foi usada a função `rescala` (PETERNELLI; MELLO, 2011).

As análises foram refeitas e os índices re-escalonados. Com esses novos valores em mãos, foi usada novamente a função desenvolvida para calcular o número de grupos que cada índice indicava. Os gráficos com as novas porcentagens do número de grupos apontado pelos índices como sendo o ideal encontram-se nas Figuras de 9 a 12.

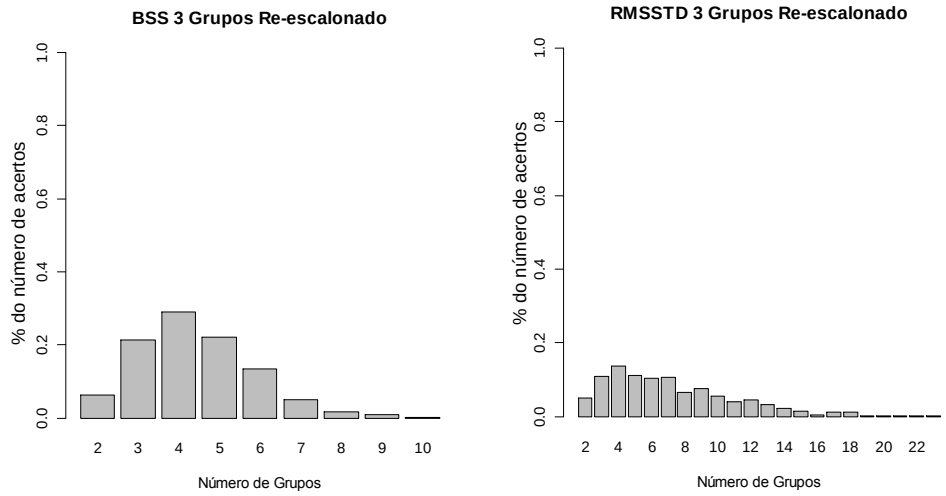

Figura 9 – Gráfico da porcentagem do número de grupos indicado pelo BSS e RMSSTD, respectivamente, no total de 1.000 simulações, quando se tinha 3 populações, com os valores re-escalonados.

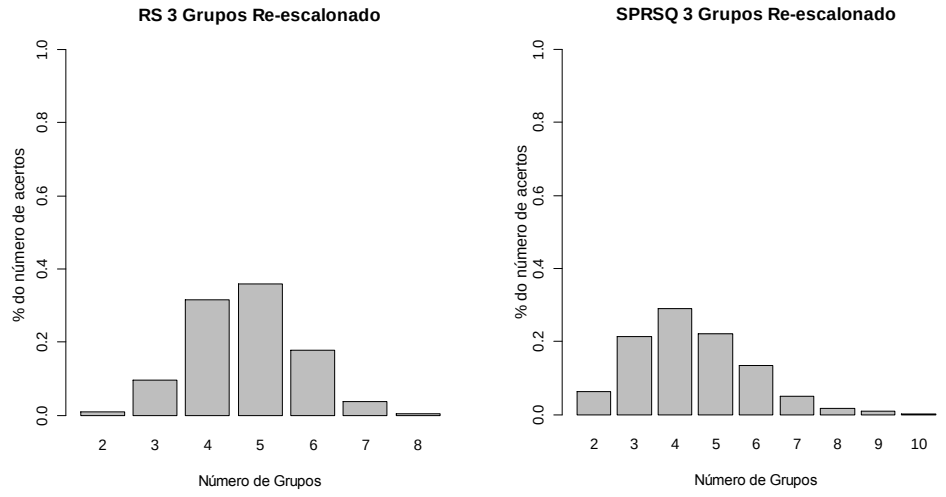

Figura 10 – Gráfico da porcentagem do número de grupos indicado pelo RS e SPRSQ, respectivamente, no total de 1.000 simulações, quando se tinha 3 populações, com os valores re-escalonados.

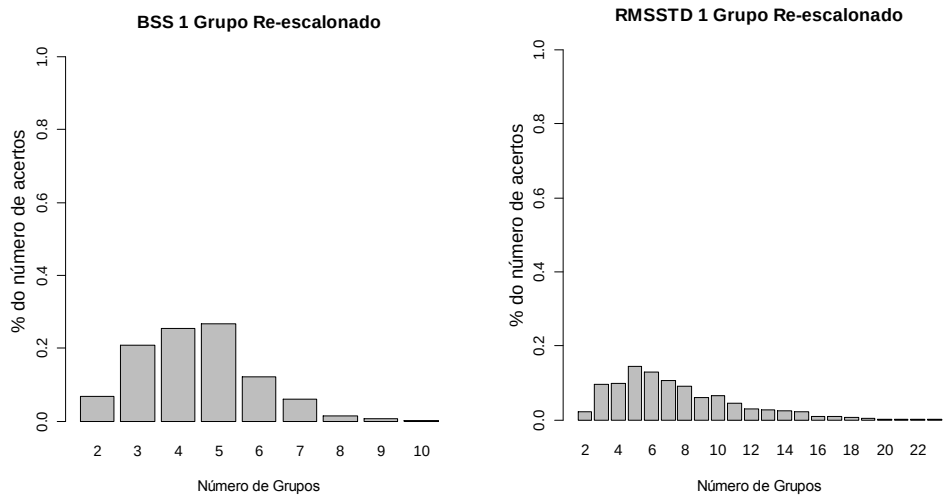

Figura 11 – Gráfico da porcentagem do número de grupos indicado pelo BSS e RMSSTD, respectivamente, no total de 1.000 simulações, quando se tinha 1 população, com os valores re-escalonados.

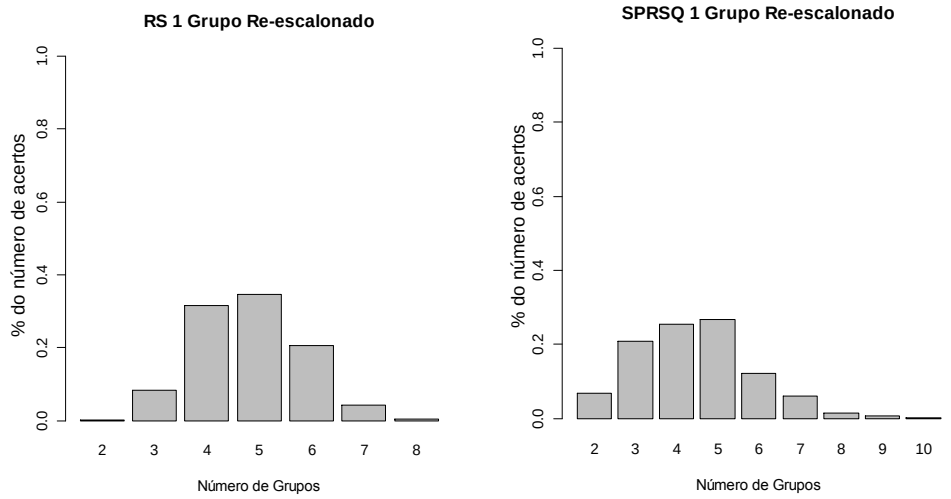

Figura 12 – Gráfico da porcentagem do número de grupos indicado pelo RS e SPRSQ, respectivamente, no total de 1.000 simulações, quando se tinha 1 população, com os valores re-escalonados.

No caso dos índices RS, SPRSQ e RMSSTD o número de grupos indicado como sendo o número ótimo aumentou nos dois cenários considerados, distando consideravelmente do real e do que era esperado. Nesse estudo, esses índices têm seus valores máximos pequenos, próximos de um para os dois primeiros e de três para o último. Esse fato pode indicar que, quando isso ocorre, a função `rescala`

tende a aumentar o valor final do número ótimo de grupos. Desse modo, seu uso não é aconselhado nesses casos.

Para o índice BSS, como seu valor máximo era muito maior que o do número de grupos, a função `rescala` beneficiou a indicação do número final de grupos a ser trabalhados. Entretanto, mesmo com a diminuição desse valor, a resposta final a respeito do número ótimo de grupos não foi satisfatória.

4 CONCLUSÃO

O uso da simulação para comparar as estatísticas usualmente utilizadas e o método de Mojena para determinar o número ótimo de grupos no Cenário 1, onde tinha-se apenas uma curva gerada, indicou que o índice RS foi o que apontou o número correto de grupos em 93,6% dos casos.

Para o Cenário 2, onde as observações foram geradas de três curvas, o método de Mojena foi o que melhor assinalou o número correto de grupos, em 68,3% das simulações.

Ao se utilizar a função que re-escalona o eixo do índice, o número de grupos indicado como ótimo não foi o próximo do real na maioria dos casos, para todos os índices nos dois cenários. Desse modo, essa função deve ser usada com cautela. Sua aplicação deve acontecer especialmente quando o índice apresentar o valor máximo muito maior do que o número de grupos máximo.

REFERÊNCIAS

ALVES, S.C.; PETERNELLI, L.C.; CARNEIRO, P.L.S. **Número de medições no ajuste de curvas de crescimento em ovinos**. 2011. Trabalho apresentado no X Encontro Mineiro de Estatística. São João del-Rei, out. 2011.

BUSSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 87p.

CARNEIRO, P.L.S.; MALHADO, C.H.M.; SOUZA JÚNIOR, A.A.O.; SILVA, A.G.S.S.; SANTOS, F.N.S.; SANTOS, P.F.; PAIVA, S.R. Desenvolvimento ponderal e diversidade fenotípica entre cruzamentos de ovinos Dorper com raças locais. **Pesquisa Agropecuária Brasileira**, Brasília, v. 42, p. 991-998, 2007.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. Vol. 2. Viçosa: UFV, 2006. 585p.

CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. 2 ed. rev. Viçosa: UFV, 2001. 390p.

FARIA, P.N. **Avaliação de métodos para determinação do número ótimo de *clusters* em estudos de divergência genética entre acessos de pimenta**. 2009. xi, 54p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2009.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 3rd ed. New Jersey: Englewood Cliffs, 1992. xiv, 642p.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS software**. New York: John Wiley and Sons, 2000.

MAIA, E.; SIQUEIRA, D.L.; SILVA, F.F.; PETERNELLI, L.A.; SALOMÃO, L.C.C. Método de comparação de modelos de regressão não-lineares em bananeiras. **Ciência Rural**, Santa Maria, v. 39, n.5, p.1380-1386, ago. 2009.

MEIER, V.D.; LESSMAN, K.J. Estimation of optimum field plot shape and size for testing yield in *Crambe abyssinica* Hochst. **Crop Science**, Madison, v. 11, n. 5, p. 648-650, Sep./Oct., 1971.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005. 297p.

MOJENA, R. Hierarquical grouping methods and stopping rules: an evaluation. **The Computer Journal**, London, 20(4), p. 359-363, 1977.

NASSI, M.O.; RUFFA, E.; ME, G.; LEPORI, G.; RADICATI, L. A contribution to the systematics of a piedmontese plum ecotype. **Plant Breeding**, Berlin, 122, p. 532-535, 2003.

OLIVEIRA, D.C. **Funções splines para estudo de curvas de crescimento em ovinos cruzados**. 2011. ix, 57p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2011.

PETERNELLI, L.A.; MELLO, M.P. **Conhecendo o R**: uma visão estatística. Viçosa: Editora UFV, 2011. 185 p.

R DEVELOPMENT CORE TEAM (2010). **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. (Acesso em 2010)

SHARMA, S. **Applied multivariate techniques**. New York: John Wiley and Sons, 1996.

SILVA, E.F.; SILVA, V.A.C.; GUIMARÃES, J.F.R.; MOURA, R.R. Divergência fenotípica entre genótipos de arroz de terras altas. **Revista Brasileira de Ciências Agrárias**, Recife, v. 6, n. 2, p. 280-286, abr./jun., 2011.

SILVEIRA, F.G. **Classificação multivariada de modelos de crescimento para grupos genéticos de ovinos de corte.** 2010. xi, 61p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2010.

SOUZA, L.A. **Avaliação do crescimento de ovinos da Raça Morada Nova sob modelos não lineares convencionais e alternativos.** 2010. 53p. Dissertação (Mestrado em Zootecnia) – Universidade Estadual do Sudoeste da Bahia, Itapetinga, 2010.

TOMAZ, F.S.C. **Análise de agrupamento para a avaliação de identidade de modelos não-lineares em análise de sobrevivência**. 2009. x, 70p. Dissertação

(Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa, 2009.

VON BERTALANFFY, L. Quantitative laws in metabolism and growth. **The Quarterly Review of Biology**, Chicago, v. 32, n. 3, p. 217-231, Sep. 1957.

WARD, J.H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, Alexandria, v.58, p.236-244, Mar. 1963.

CONSIDERAÇÕES FINAIS

O objetivo desse estudo foi comparar as estatísticas empregadas no auxílio da escolha do número ótimo de grupos a ser trabalhado em análise de agrupamento. O referencial teórico necessário aos dois capítulos foi primeiramente abordado.

No capítulo 1 foi feito o estudo de caso real, onde o número de grupos era pré-determinado e igual a três. Dentre os índices usados: BSS, RMSSTD, RS e SPRSQ e o método de Mojena, o que identificou corretamente o número ótimo de grupos foi o SPRSQ. O uso de uma função para re-escalonar os valores dos índices, de modo que ficassem na mesma proporção do número de grupos, contribuiu para que o BSS e o RS também passassem a informar corretamente o número ótimo de *clusters*. O RMSSTD foi o índice que apresentou pior desempenho em ambos os casos, identificando cinco grupos.

No segundo capítulo utilizou-se de simulação para comparar os índices BSS, RS, RMSSTD e SPRSQ e o método de Mojena quanto à determinação do número ótimo de grupos. Foi comparada a porcentagem de acertos das estatísticas em dois cenários. Naquele em que as observações eram originadas de uma única curva, o RS indicou o número de grupos corretamente na grande maioria dos casos. Para o cenário em que as observações provinham de três curvas geradoras, o método de Mojena indicou o número ótimo de grupos na maioria das vezes. Tanto o BSS quanto o RMSSTD mostraram-se insatisfatórios para indicar o número ótimo de *clusters*, pois além do baixo percentual de acertos, indicaram um valor muito maior na maioria dos casos.

Ao se usar a função `rescala` nas simulações, os índices apresentaram uma piora para acertar o número ótimo de grupos, inclusive distando mais do real. Exceto para o caso do BSS, em que houve ligeira melhora. Desse modo, o uso da função `rescala` deve ser feito com moderação.

De modo geral, os índices RS e SPRSQ se mostraram satisfatórios na indicação do número ótimo de grupos a ser trabalhados. Entretanto novos estudos, contemplando outros cenários e casos reais, devem ser feitos para efetivar tais constatações.

Do mesmo modo, o uso da função `rescala` deve ser melhor investigado para que sua aplicação ajude na determinação do número ótimo de grupos correto, não comprometendo a decisão final.

APÊNDICES

APÊNDICE A

Função criada para calcular o número de grupos, baseada na lei dos cossenos e no Método da Máxima Curvatura Modificado.

```
calcula.nc<-function(eixo.do.x,medida.calculada)
{
x<-eixo.do.x
y<-medida.calculada
nx<-length(x)
angulo1<-NULL
dif<-y[-nx]-y[-1]
for(i in 2:(nx-1))
 {
  a<-c(x[i],y[i])        # definindo as coordenadas do ponto a
  b<-c(x[i-1],y[i-1])
  c<-c(x[i+1],y[i+1])
  dab<-sqrt((b[1]-a[1])^2+(b[2]-a[2])^2)#distância entre a e b
  dac<-sqrt((c[1]-a[1])^2+(c[2]-a[2])^2)
  dbc<-sqrt((b[1]-c[1])^2+(b[2]-c[2])^2)

if(dab+dac-dbc<0.0001) {angulo1[i-1]<-180}
else{
alfar<-acos((dab^2+dac^2-dbc^2)/(2*dac*dab))  # lei cossenos
angulo1[i-1]<-360-(180*alfar/pi) # pegando ângulo replementar
    }
angulo<-rev(angulo1)
   }
perigo<-which((abs(dif[-1])-abs(dif[-(nx-1)]))<0)
angulo[perigo]<-360-angulo[perigo]
max.curv<-which.max(angulo)+1
return(max.curv)
}
```

APÊNDICE B

Dedução da lei dos cossenos.

Considere o triângulo ABC abaixo:

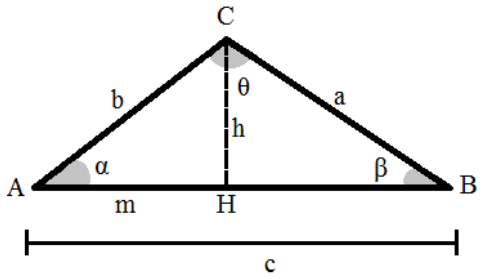

Figura 1 – Triângulo ACB, de lados *a*, *b*, *c* e ângulos α, β e θ, de onde se deduz a lei dos cossenos.

No triângulo BCH, pelo Teorema de Pitágoras, tem-se:

$a^2 = h^2 + (c - m)^2$ (I)

No triângulo ACH, também pelo Teorema de Pitágoras, tem-se:

$b^2 = h^2 + m^2 \quad \Rightarrow \quad h^2 = b^2 - m^2$ (II)

Substituindo (II) em (I), obtém-se:

$a^2 = b^2 - m^2 + (c - m)^2 \quad \Rightarrow \quad a^2 = b^2 + c^2 - 2.c.m$ (III)

Ainda no triângulo ACH, tem-se:

$\cos\alpha = \frac{m}{b} \quad \Rightarrow \quad m = b.\cos\alpha$ (IV)

Substituindo (IV) em (III), obtém-se:

$a^2 = b^2 + c^2 - 2.b.c.\cos\alpha$

APÊNDICE C

Função criada para simular o máximo de três populações de animais distintas.

```
gera.animais.dif<-
function(nsimTotal=1,n1=10,n2=12,n3=18,grupo,arquivo.do.x,
beta1.1,beta2.1,beta3.1, beta1.2,beta2.2,beta3.2,
beta1.3,beta2.3,beta3.3,
var.beta1.1,var.beta2.1,var.beta3.1,
var.beta1.2,var.beta2.2,var.beta3.2,
var.beta1.3,var.beta2.3,var.beta3.3, media.erro=0,
desvio.erro.1=1, desvio.erro.2=1, desvio.erro.3=1)

{
#nsimTotal = número de simulação a ser gerado. Em cada
#simulação serão gerados 'num.animais' indivíduos
#eixo.x<-arquivo.do.x
n<-length(x)
dp1<-desvio.erro.1
dp2<-desvio.erro.2
dp3<-desvio.erro.3

################################################################
################################################################
#iniciando cada simulação

for(nsim in 1:nsimTotal)
{

################### para população 1 ############
#criando o primeiro indivíduo só para iniciar
################################################
#definindo o modelo pop1
beta1<-rnorm(n1,beta1.1,sqrt(var.beta1.1))
beta2<-rnorm(n1,beta2.1,sqrt(var.beta2.1))
beta3<-rnorm(n1,beta3.1,sqrt(var.beta3.1))
betas.1<-
cbind(rep(beta1,each=n),rep(beta2,each=n),rep(beta3,each=n))
#grupo 1
media<-media.erro
modelo=beta1*(1-beta2*exp(-beta3%*%t(x)))^3
modelo.pop1<<-modelo

#definindo o modelo pop2
beta1<-rnorm(n2,beta1.2,sqrt(var.beta1.2))
beta2<-rnorm(n2,beta2.2,sqrt(var.beta2.2))
beta3<-rnorm(n2,beta3.2,sqrt(var.beta3.2))
betas.2<-
cbind(rep(beta1,each=n),rep(beta2,each=n),rep(beta3,each=n))
#grupo 2
media<-media.erro
modelo=beta1*(1-beta2*exp(-beta3%*%t(x)))^3
modelo.pop2<-modelo

#definindo o modelo pop3
```

```
beta1<-rnorm(n3,beta1.3,sqrt(var.beta1.3))
beta2<-rnorm(n3,beta2.3,sqrt(var.beta2.3))
beta3<-rnorm(n3,beta3.3,sqrt(var.beta3.3))
betas.3<-
cbind(rep(beta1,each=n),rep(beta2,each=n),rep(beta3,each=n))
#grupo 3
media<-media.erro
modelo=beta1*(1-beta2*exp(-beta3%*%t(x)))^3
modelo.pop3<-modelo

############################################################

k<-1
y<-round(modelo.pop1[k,]+rnorm(n,media,dp1),6)
dados<-data.frame(ind=k,x,y)

#agora fazendo para os outros indivíduos.
for (k in 1:n1)
{
y<-round(modelo.pop1[k,]+rnorm(n,media,dp1),6)
dados2<-data.frame(ind=k,x,y)
dados<-rbind(dados,dados2)
}
dados.pop1<<-(dados[-(1:(length(x))),])
rm(dados)
#}#finalizando cada simulação
################## fim pop 1 ##########################


################### para população 2 #################
#criando o primeiro indivíduo só para iniciar
k<-1
y<-round(modelo.pop2[k,]+rnorm(n,media,dp2),6)
dados<-data.frame(ind=k,x,y)

#agora fazendo para os outros indivíduos. No final vou
#eliminar o primeiro gerado acima
for (k in (n1+1):(n1+n2))
{
y<-round(modelo.pop2[k-n1,]+rnorm(n,media,dp2),6)
dados2<-data.frame(ind=k,x,y)
dados<-rbind(dados,dados2)
}

dados.pop2<-(dados[-(1:(length(x))),])
rm(dados)
#}#finalizando cada simulação
################## fim pop 2 ##########################

################### para população 3 #################
#criando o primeiro indivíduo só para iniciar
k<-1
y<-round(modelo.pop3[k,]+rnorm(n,media,dp3),6)
dados<-data.frame(ind=k,x,y)

#agora fazendo para os outros indivíduos.
```

```
for (k in (n1+n2+1):(n1+n2+n3))
{
y<-round(modelo.pop3[k-n1-n2,]+rnorm(n,media,dp3),6)
dados2<-data.frame(ind=k,x,y)
dados<-rbind(dados,dados2)
}
dados.pop3<-(dados[-(1:(length(x))),])
rm(dados)
################## fim pop 3 ##########################

#juntando os 3 arquivos de dados gerados. Um de cada população
betas<-rbind(betas.1,betas.2,betas.3)
dados<-rbind(dados.pop1,dados.pop2,dados.pop3)
dados<-cbind(dados,betas)
names(dados)<-c("ind","x","y","beta1","beta2","beta3")

write.table(dados,paste("dados.simulados/",grupo,nsim,".txt",
sep=""),quote=F,row.names=F,append=F,col.names=T)

}#finalizando cada simulação

} #fechando a função
```

ANEXO 1

Função que re-escalona um vetor, de acordo com o livro de Peternelli e Mello (2011, página 85).

```
rescala<-function(vetor,novo.min=1,novo.max=6)
{
min.atual<-min(vetor)
max.atual<-max(vetor)
reescalado<-((vetor-min.atual)/(max.atual-
min.atual))*(novo.max-novo.min)+novo.min
return(reescalado)
}
```