EUCYMARA FRANÇA NUNES SANTOS

SEPARAÇÃO DE GRUPOS PRODUTIVOS EM BOVINOS LEITEIROS ATRAVÉS DE TÉCNICAS MULTIVARIADAS

RECIFE-PE

FEVEREIRO/2009

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

SEPARAÇÃO DE GRUPOS PRODUTIVOS EM BOVINOS LEITEIROS ATRAVÉS DE TÉCNICAS MULTIVARIADAS

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Mestre.

Área de Concentração: Modelagem Estatística e Computacional

Orientador: Prof. Dr. Kleber Régis Santoro

Co-orientador: Prof. Dr. Rinaldo Luiz Caraciolo Ferreira

RECIFE-PE

FEVEREIRO/2009

FICHA CATALOGRÁFICA

S237s Santos, Eucymara França Nunes
Separação de grupos produtivos em bovinos leiteiros através de técnicas multivariadas / Eucymara França Nunes Santos. -- 2009.
37 f. : il.

Orientador : Kleber Régis Santoro
Dissertação (Mestrado em Biometria) - Universidade Federal Rural de Pernambuco. Departamento Ciência da Compu - tação e Estatística.
Inclui bibliografia.

CDD 574. 018 2

1. Análise de agrupamentos
2. Análise de componentes principais
3. Análise discriminante
I Santoro, Kleber Régis
II. Título

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

SEPARAÇÃO DE GRUPOS PRODUTIVOS EM BOVINOS LEITEIROS ATRAVÉS DE TÉCNICAS MULTIVARIADAS

EUCYMARA FRANÇA NUNES SANTOS

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada por unanimidade em 27/02/2009 pela Comissão Examinadora.

Orientador:

_____

Kleber Régis Santoro

Co-orientador: Rinaldo Luiz Caraciolo Ferreira

Banca Examinadora:

_____

Eufrázio de Souza Santos – DSc (UFRPE)

_____

Severino Benone Paes Barbosa – DSc (UFRPE)

_____

Borko Stosic – DSc (UFRPE)

DEDICO

Ao meu tio Fernando Antônio Nunes Santos pelo apoio financeiro permitindo a realização desta titulação.

Agradecimentos

Agradeço, primeiramente, a DEUS, criador dos mais nobres sentimentos. Foram esses sentimentos que me deram força para alcançar essa vitória. Em seguida, aos meus pais Pedro Augusto e Maria Eucia por me ensinarem os valores da vida e por serem sempre o meu pilar de sustentação.

Foi gratificante receber o carinho e a paciência do meu noivo Igor Divino, os cuidados do meu irmão Fábio França, o apoio financeiro do meu tio Fernando Nunes e o apoio material do meu tio Francisco Nunes.

O início desta minha caminhada começou na Universidade Federal de Sergipe – UFS. De onde até hoje recebo apoio do professor Manuel Luiz Figueirôa, presença constante a cada passo da minha vida profissional, a quem agradeço os conselhos e o incentivo. Agradeço também a torcida e a ajuda dos professores: Daniel Neyra, Marcela Bernardes, Lázaro Araújo, Suzana Russo, Samuel Ribeiro e Kleber Fernandes.

E agradeço, particularmente, as pessoas que contribuíram diretamente na construção desta dissertação:

Ao professor Kleber Santoro pela magnífica orientação, dedicação e paciência.

Aos professores que me transmitiram conhecimentos: Eufrázio Santos, Gauss Cordeiro, Rinaldo Ferreira, José Aleixo, Maria Adélia, Borko Stosic e Tatijana Stosic.

Aos meus colegas de classe, especialmente aos amigos: Katiane Conceição, Lenaldo Azevedo e Magali Teresópolis.

Ao colega Luiz Henrique pela fundamental ajuda no desenvolvimento da dissertação, e ao zootecnista Gladston Santos pela colaboração e apoio.

Aos funcionários Marco Santos e Zuleide.

Para não correr o risco da injustiça, agradeço, de antemão, a todos que de alguma forma passaram pela minha vida e contribuíram para a construção de quem sou hoje.

“No futuro, o pensamento estatístico será tão necessário para a cidadania eficiente como saber ler e escrever.”

Herbert George Wells (1866 – 1946).

RESUMO

Foram utilizadas as técnicas multivariada: análise de componentes principais, análise de agrupamentos e análise discriminante com o objetivo de separar os grupos produtivos geneticamente divergentes, utilizando dados referente a produção de leite de três diferentes grupos genéticos: 1/2 HG; 3/4 HG; 7/8 HG. As variáveis utilizadas foram: grupo genético, peso do leite (kg) produzido no dia do controle, peso do leite (kg) produzido na primeira ordenha, peso do leite (kg) produzido na segunda ordenha, peso do leite (kg) produzido na terceira ordenha, idade da vaca (dias) na data do controle, idade da vaca ao parto e intervalo de partos. Os objetivos da análise de componentes principais foram: propor a utilização dos dados mais adequados e verificar as variáveis mais importantes. Proporcionando a explicação de 92,84% da variabilidade dos dados com os dados transformados e a eliminação de cinco variáveis não significativas. Foram utilizadas quatro medidas de distância e cinco métodos de agrupamentos na análise de agrupamentos objetivando indicar a melhor distância e o melhor método. Constatou-se que a distância de Mahalanobis juntamente aos métodos de ligação média, ligação simples e centróide são os mais indicados para agrupar os diferentes grupos genéticos. A análise discriminante foi utilizada para selecionar as variáveis mais importantes e estabelecer equações discriminantes que possibilite a inclusão de novos indivíduos. Foram selecionadas duas variáveis, e eliminada uma, o grupo 1/2 HG obteve mais classificações corretas e a função apresentada foi referente aos dados padronizados por possuir melhores classificações.

Palavras-chave: componentes principais, agrupamentos, discriminante, grupos genéticos.

ABSTRACT

Many varieties of techniques were used: analysis of main components, analysis of grouping and discriminant analysis with the objective of separating the productive groups genetically divergents, using data regarded to the production of milk from three different genetic groups: 1/2 HG; 3/4 HG; 7/8 HG. The used variables were: group genetic, weigh of the milk (kg) produced in the day of the control, weight of the milk (kg) produced in the first it milks, weigh of the milk (kg) produced in the second it milks, weigh of the milk (kg) produced in the third it milks, the age of the cow (days) in the date of the control, the age of the cow to the childbirth and interval of childbirths. The objectives of the analysis of main components were: proposing the use of the most appropriate data and verifing the most important variables. Providing the explanation of 92,84% of the variability of the data with the transformed data and the elimination of five no significants variables. Four distance measures and five methods of groupings were used in the analysis of groupings aiming at the indication of the best distance and the best method. It was verified that the distance of Mahalanobis taken together to the methods of medium connection, simple connection and centroid are the most suitable to contain the differents genetic groups. The discriminant analysis was used to select the most important variables and to establish discriminant equations that makes possible the new animals inclusion. Two variables were selected, and one was eliminated, the group 1/2 HG has got more correct classifications and the presented function was regarding to the standardized data for its better classifications.

Keywords: principal components, clustering, discriminant, genetic groups.

SUMÁRIO

LISTA DE FIGURAS

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

INTRODUÇÃO

A estatística multivariada é um conjunto de métodos que permite a análise simultânea de medidas múltiplas em cada indivíduo ou objeto. As medidas referem-se às variáveis nas unidades de pesquisa, amostral, experimental ou mesmo observações.

Os primeiros contribuintes da estatística multivariada no ínicio do século XX foram: Pearson (1901), Fisher (1928), Hotelling (1931), Wilks (1932) e Bartlett (1937) apud Reis (2001).

O conjunto de técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados consiste em analisar várias variáveis que são medidas simultaneamente numa mesma unidade experimental, em cada elemento amostral. Como ferramentas para o desenvolvimento dessas metodologias são empregadas técnicas como a análise de componentes principais, análise de agrupamentos, análise discriminante entre outras que permitem a redução de dados ou simplificação estrutural, ordenação e agrupamento e a investigação da dependência entre variáveis (MINGOTI, 2005).

A análise de componentes principais (Principal Components Analysis – PCA) é uma técnica de análise multivariada que permite a redução da dimensionalidade dos dados e do número de variáveis. Consistem em transformar um conjunto de variáveis originais em um pequeno número de combinações lineares, os chamados componentes principais, de dimensões equivalentes, e com propriedades importantes. Os escores (não correlacionados) dos componentes, colocados uns contra os outros, podem favorecer a análise de gráficos, onde se pode observar a formação de grupos. O objetivo é obter variáveis ou conjunto de variáveis que retenham o máximo possível de informações nelas contidas e expliquem a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre eles.

Portanto, componentes principais é uma técnica de análise intermediária, e não se constitui em um método final e conclusivo, que se presta fundamentalmente como um passo intermediário em grandes investigações científicas (DILLON, 1984; MARDIA, 1979; REYMENT, 1984).

A análise de agrupamentos também conhecida como conglomerados, classificação, tipologia numérica, taxonomia ou cluster, constitui-se em uma técnica

a ser utilizada para a descoberta de uma estrutura de grupos e de relações entre esses grupos. Tem como objetivo dividir os elementos da amostra, ou população, em grupos de tal forma que exista homogeneidade dentro dos grupos e heterogeneidade entre grupos (CRUZ & REGAZZI, 1997).

Um conceito fundamental na utilização das técnicas de análise de agrupamento é a escolha de um critério que meça a distância entre dois objetos, ou que quantifique o quanto eles são parecidos. Esta medida é chamada coeficiente de parecença, sendo dividida em duas categorias: medidas de similaridade (quanto maior o valor, maior a semelhança entre os objetos) e de dissimilaridade (ou medida de distância, quanto maior o valor, mais diferentes são os objetos). A seguir um algoritmo de agrupamento é aplicado sobre essa matriz, de modo a identificar e conectar grupos homogêneos, os quais podem ser representados graficamente por um diagrama denominado dendograma. A maioria dos algoritmos utilizados na formação dos agrupamentos podem ser classificados como métodos hierárquicos ou de partição (BUSSAB, 1990; OLIVEIRA, 2004).

As diferentes técnicas podem levar a diferentes soluções. A avaliação passou a ser utilizada para se comparar diferentes classificações advindas de um mesmo conjunto de dados. O uso de um método de validação é apropriado quando se deseja escolher o melhor método de aglomeração. Sokal e Rohlf (1962) definiram o coeficiente de correlação cofenética, calculado entre os índices de similaridade da matriz original (matriz de distância) e os índices reconstruídos, com base no dendograma (matriz cofenética ou coeficiente de fusão-distância a que os indivíduos se juntam pela primeira vez para formar grupos), como o coeficiente r de Pearson.

Outros métodos podem ser utilizados para avaliar os agrupamentos formados, tais como, a determinação do número ideal de grupos através da visualização do dendograma, em busca de grandes alterações dos níveis de similaridade para as sucessivas fusões; o grau de covergência dos vários critérios de agregação pode ser aferido através de uma tabela de contingência representando a associação de grupos nos métodos aplicados, e o nível de significância do teste de independência qui-quadrado, dos resultados de cada par de critérios de agregação (REIS, 2001).

A análise discriminante é destinada a interpretar grupos de objetos, definidos a priori pelos métodos de agrupamento. A técnica consiste em encontrar funções (combinações lineares das variáveis observadas em grupos definidos a priori) que possam explicar as diferenças entre as populações ou permitam classificar novos

objetos em umas das populações, de tal modo que seja minimizada a probabilidade de incorreta classificação a posteriori (VALENTIN, 2000).

A análise multivariada é empregada nas diversas áreas e em diferentes espécies, têm sido utilizadas pelos melhoristas genéticos, constituindo-se de uma valiosa ferramenta para os programas de melhoramento genético. Elas podem ajudar na identificação de grupos de animais produtivamente semelhantes, apesar de geneticamente divergentes, para facilitar os sistemas de manejo nos animais (nutricional, reprodutivo e produtivo), auxiliando na tomada de decisões. As mais empregadas no melhoramento genético são as medidas de dissimilaridade: distância Euclidiana e a de Mahalanobis (CRUZ & REGAZZI, 1997).

Há uma falta de conhecimento sobre as possibilidades de agrupamentos de animais geneticamente divergentes, mas produtivamente similares. Não se encontrou qualquer referência na literatura. A falta de pesquisa nesta área decorre, possivelmente, do pequeno número de animais mestiços sob controle leiteiro no país. Evidenciar esses estudos seria de grande importância prática para os produtores.

Portanto, este trabalho tem como objetivo utilizar as técnicas multivariadas como: componentes principais, agrupamentos e discriminante. Cada uma dessas técnicas serão organizadas em capítulos, a fim de se avaliar os dados e métodos mais adequados, analisar as variáveis utilizadas e proporcionar a inclusão de novos animais nos grupos estudados.

REFERÊNCIAS BIBLIOGRÁFICAS

BUSSAB, W. de. O.; MIAZAKI, E. S.; ANDRADE, D. F. de. Introdução à análise de agrupamentos. São Paulo. Associação Brasileira de estatística. 9º Simpósio Nacional de Probabilidade e Estatística, 1990. 105p.

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

DILLON, W. R. Multivariate analysis. Canadá. John Wiley e Sons, 1984.

MARDIA, K. V.; KENT, J. T.; BIBBLY, J. M. Multivariate analysis. London. Editors Z. M. Birnbaum and Lukacs Academic Press. 1997. 518p.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada. Belo Horizonte: Editora UFMG, 2005, 295p.

OLIVEIRA, J. C. V. Caracterização e perfil etnológico de rebanhos caprinos nos municípios de Ibimirim e Serra Talhada, Estado de Pernambuco. 2004. 60f. Dissertação (Mestrado em Zootecnia) – Universidade Federal Rural de Pernambuco, Recife, 2004.

REIS, E. Estatística multivariada aplicada. 2 ed. Lisboa: Silabo, 2001. 253p.

REYMENT, R. Multidimensional paleobiolgy. Pergamon Press: New York, 1984, 377p.

SOKAL, R.; ROHLF, F. The comparison of dendograms by objective methods. Taxon 11, p. 34 – 40, 1962.

VALENTIN, J. L. Ecologia Numérica: Uma introdução à análise multivariada de dados ecológicos. Rio de Janeiro. Editora Interciência, 2000. 117p.

FORMAÇÃO DE GRUPOS PRODUTIVOS EM VACAS LEITEIRAS ATRAVÉS DE COMPONENTES PRINCIPAIS

Eucymara França Nunes SANTOS[1]
Kleber Régis SANTORO[2]
Rinaldo Luiz Caraciolo FERREIRA[3]
Eufrázio de Souza SANTOS[4]
Gladston Rafael de Arruda SANTOS[5]

- RESUMO: Os objetivos deste trabalho foram propor o método mais adequado, através da análise de componentes principais, utilizando dados referentes à produção de leite de três grupos geneticamente divergentes, com o intuito de visualizar a separação destes grupos através de gráficos utilizando os escores dos componentes principais, e eliminar as variáveis menos importante sem muita perda de informação. As variáveis características para a produção de leite analisadas foram: grupo genético, peso do leite (kg) produzido no dia do controle, peso do leite (kg) produzido na primeira, segunda e terceira ordenhas, idade da vaca (dias) ao parto, idade da vaca (dias) na data do controle leiteiro e intervalo de partos (dias). As análises foram realizadas com os dados originais de coleta e com os dados transformados, devido as diferentes medidas. A análise com os dados transformados proporcionou a obtenção de três componentes com a explicação de 92,84% da variabilidade dos dados. Eliminou cinco variáveis não significativas e apresentou o melhor gráfico de separação dos grupos genéticos.
- PALAVRAS-CHAVE: componentes principais, grupos genéticos, vacas leiteiras, separação.

- *ABSTRACT: This work objetives went to propose the most appropriate method, through the principal components analysis, using data regarding production to the of milkfrom three genetically divergents groups, with the intention of visualizing the separation of these groups through graphs using the scors of the principal components and eliminaing less important variables without a lot of loss of information. The characteristic variables for the production of milk analyzed were: group genetic, weigh of the milk (kg) produced in the day of the control, weight of the milk (kg) produced in the first it milks, weigh of the milk (kg) produced in the second it milks, weigh of the milk (kg) produced in the third it milks, age of the cow (days) in the date of the control, age of the cow to the childbirth and interval of childbirths. The analysis were accomplished with the original data of collection and with the transformed data, due the different measures. The analysis with the transformed data provided the aquisition of three components with the explanation of 92,84% of the variability of the data. I also eliminated five non significant variables and presented the best separation graph of the genetic groups.*
- *KEYWORDS: principal components, genetic groups, cows milk, separation.*

1 Introdução

A estatística multivariada é um conjunto de técnicas exploratórias de sintetização (ou simplificação) da estrutura de variabilidade dos dados, consistindo em analisar várias variáveis simultaneamente.

Dentre elas está a análise de componentes principais (Principal Components Analysis – PCA), que consiste em transformar um conjunto de variáveis originais em um pequeno número de combinações lineares, os chamados componentes principais, de dimensões equivalentes, porém com propriedades importantes. O objetivo é obter variáveis ou conjunto de variáveis que retenham o máximo possível de informações e expliquem a maior parte da variabilidade total, revelando que tipo de relacionamento existe entre eles (REIS, 2001).

Este método é utilizado para identificar o fator dimensão dos dados, a redução da dimensão fornece gráficos para a análise através dos escores dos componentes, nos quais se pode observar a formação de grupos. Portanto a representação gráfica dos componentes principais são ferramentas valiosas na explanação na análise de dados (DILLON & GOLDSTEIN, 1984) e (MARDIA et al., 1997).

[1] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: eucymara@gmail.com
[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco – UFRPE - UAG, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: krsantoro@uag.ufrpe.br
[3] Departamento de Ciência Florestal, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: rinaldo@dcfl.ufrpe.br
[4] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: eufrazio@deinfo.ufrpe.br
[5] Empresa Pernambucana de Pesquisa Agropecuária – IPA, CEP:56.600-000 , Sertânia, Pernambuco, Brasil, E-mail: gladstonrafael@ipa.br

Em estudos de divergência genética, a análise de componentes principais contribui na interpretação das relações existentes entre as variáveis, propõe relativa economia de tempo e custo em experimentos futuros descartando variáveis redundantes sem perda considerável de informação, auxilia na tomada de decisões e pode levar a aceleração de um programa de melhoramento genético (MORRISON, 1976).

Os objetivos deste estudo foram verificar a possibilidade de formação de diferentes grupos de animais através da análise de componentes principais, utilizando os dados originais e padronizados para propor os mais adequados, verificar a possibilidade de redução no número de variáveis envolvidas na análise e melhorar a compreensão da dispersão dos dados.

2 Análise de Componentes Principais

A técnica de componentes principais foi originalmente desenvolvida por Karl Pearson (1901) e, posteriormente, aplicada por Harold Hotteling (1933) em diversas áreas da ciência (apud MARDIA et al., 1997).

As medidas das variáveis originais são correlacionadas entre si, indicando que algumas informações contidas em uma variável também estão em outra. Então o objetivo da análise de componentes principais é transformar a quantidade de variáveis correlacionadas em uma quantidade de variáveis não correlacionadas, ou seja, os componentes principais (Cruz & Regazzi, 1997). Os melhores resultados são obtidos quando as variáveis originais são altamente correlacionadas, positiva ou negativamente.

Seja um conjunto de variáveis observadas X_1, X_2, ... , X_j o componente principal é obtido pela combinação linear

$$Y_i = a_{i1} X_1 + a_{i2} X_2 + ... + a_{ij} X_j, \quad (1)$$

a partir da matriz de correlação (R) ou variância-covariância (S). A solução é obtida através da equação característica

$$|R - \lambda I| = 0 \quad (2)$$

sendo I a matriz identidade.

Para cada autovalor (λ) determina-se um autovetor a_i:

$$|R - \lambda I| \, a_i = 0 \quad (3)$$

Cada autovalor representa a variância de cada componente. A interpretação relativa de cada componente principal Y_i é então avaliada pela porcentagem da variância total que ele explica. A variância total é definida pela soma das variâncias de cada uma das variáveis. As variâncias individuais constituem os elementos da diagonal principal (traço da matriz), somando-os encontra-se-á a variância total.

Os autovetores são determinados pelos coeficientes das equações lineares de cada componente principal. Substituindo-se as variáveis originais nas equações dos componentes, encontram-se-ão os escores de cada componente.

A importância da variância dos componentes são organizados em ordem decrescente. Os últimos componentes são responsáveis por uma fração muito pequena da variabilidade dos dados. Segundo Johnson & Wichern (1998), 80 ou 90% da variabilidade total pode ser explicada pelas primeiras componentes, que poderão ser usadas no lugar das variáveis originais sem perder muita informação.

Examinando-se as correlações entre cada variável com o respectivo componente principal, aquela que apresentar a maior correlação com o componente de menor variância total pode ser descartada. A utilização desta técnica possibilita a redução em poucos componentes, enquanto mantém a maior quantidade de informação quanto possível. A variável que possui maior correlação com o componente principal de menor variância deve ser menos importante para explicar a variância total, praticamente insignificante (MARDIA et al., 1997).

3 Materiais e Métodos

Neste trabalho foram utilizados dados provenientes de cruzamentos entre animais Holandês (HO) e Gir (GL) de vacas leiteiras de três diferentes grupos genéticos: 1/2 HO, GL; 3/4 HO, GL e 7/8 HO,GL, com 326 animais por categoria. Os dados foram coletados semanalmente no período de dezembro/2000 a dezembro/2006 numa fazenda com agropecuária semi-intensiva de leite, na região de Ribeirão, Zona da Mata de Pernambuco.

Foram analisadas as seguintes características da produção de leite: grupo genético (GRAUSANGUE), peso do leite (Kg) produzido no dia do controle (PESOLEITE), peso do leite (Kg) produzido na primeira ordenha (PESOLEITE1), peso do leite (Kg) produzido na segunda ordenha (PESOLEITE2), peso do leite (Kg) produzido na terceira ordenha (PESOLEITE3), idade da vaca (dias) na data do controle (IDADEPESALEITE), idade da vaca (dias) ao parto (IDADEPARTO) e intervalo de partos (IEP), com 13.643, 13.643, 13.600, 12.128, 9.643, 13.643, 13.643 e 3.119 observações por variável, respectivamente.

Os dados foram analisados através da análise de componentes principais. Como as variáveis observadas possuiam diferentes unidades de medidas, então as componentes principais podem ser acompanhadas de unidade de medidas sem sentido, neste caso é coveniente fazer uma padronização das variáveis com média zero e variância um, ou seja, distribuição N(0,1) dos dados. O critério para seleção segue o sugerido por Manly (2004) e Hair Jr. et al. (2005), onde o número de componentes a ser utilizado chega à retenção de noventa por cento.

Para descarte de variáveis, Jolliffe (1973 apud Barbosa, 2006) recomenda que quando a análise de componentes principais utiliza a matriz de correlação, estabelece-se que o número de variáveis descartadas deveria ser menor ou igual ao número de componentes cuja variância (autovalor) é inferior a 0,7. É possível descartar variáveis que pouco contribuem para a discriminação dos dados.

Para análise dos componentes principais utilizou-se o procedimento PRINCOMP, do software SAS (SAS INSTITUTE, 2002).

4 Resultados e Discussão

Pode-se observar a matriz de correlação obtida através da análise de componentes principais na Tabela 1, que houve alta correlação positiva entre as variáveis peso do leite no dia do controle em relação ao peso do leite da primeira, segunda e terceira ordenhas, e entre a idade da vaca no dia do controle e a idade da vaca no dia do parto. Sendo baixa a correlação entre a variável grupo genético em relação as demais variáveis, sendo que o intervalo de partos foi a que obteve uma melhor correlação dentre elas. A variável intervalo de partos obteve baixa correlação com as características do peso do leite. McManus et al. (2002) encontraram médias/altas correlações entre a variável fertilidade real, com os pesos da vaca e alta com intervalo de partos. Para explicar a distribuição dos grupos, não será necessário um grande número de componentes, por possuir altas correlações entre as variáveis, podendo variar de acordo com a população estudada (MANLY, 2004)

Tabela 1 - Matriz de correlação entre as variáveis utilizadas na análise de componentes principais

Variável	Variável							
	GRAU SANGUE	PESO LEITE	PESO LEITE1	PESO LEITE2	PESO LEITE3	IDADE-PESA LEITE	IDADE-PARTO	IEP
GRAUSANGUE	1,0000							
PESOLEITE	0,0895	1,0000						
PESOLEITE1	0,0922	0,9657	1,0000					
PESOLEITE2	0,0781	0,9505	0.8715	1.0000				
PESOLEITE3	0,0848	0,9484	0,8779	0,8560	1,0000			
IDADE_PESALEITE	-0,0035	0,1225	0,1520	0,1040	0,0854	1,0000		
IDADE_PARTO	-0,0266	0,2261	0,2512	0,2018	0,1853	0,9765	1,0000	
IEP	0,2653	0.0757	0,0831	0,0682	0,0628	0,1311	0,1033	1,0000

Conforme pode-se verificar na Tabela 2 obteve-se um número de oito componentes principais, podendo reduzir as variáveis originais em quatro componentes, os quais retiveram 96,53% da variação dos dados, quatro componentes (50%) apresentam variância (autovalor) inferior a 0,7 (JOLLIFFE, 1973 apud BARBOSA, 2006). Assim, as quatro componentes a partir do último componente principal, são passíveis de descarte. Em estudo de redução da dimensionalidade do espaço multivariado por meio de componentes principais, Barbosa et al. (2006) sugeriram a redução de dez para quatro variáveis por apresentarem variância inferior a 0,7. A Figura 1 sugere que a partir do quinto componente, a variância dos componentes torna-se praticamente nula. A razão para isto é que as variáveis correlacionadas aos componentes principais explicam menores percentuais de variância.

Tabela 2 - Autovalores, proporção individual e acumulada da variação dos dados através da análise dos componentes principais

Componente	Autovalor	Proporção Individual	Proporção Acumulada
1	3,88161555	0,4852	0,4852
2	1,88062257	0,2351	0,7203
3	1,24090391	0,1551	0,8754
4	0,71912235	0,0899	0,9653
5	0,14451534	0,0181	0,9833
6	0,11624323	0,0145	0,9979
7	0,01697706	0,0021	1,0000
8	0,00000000	0,0000	1,0000

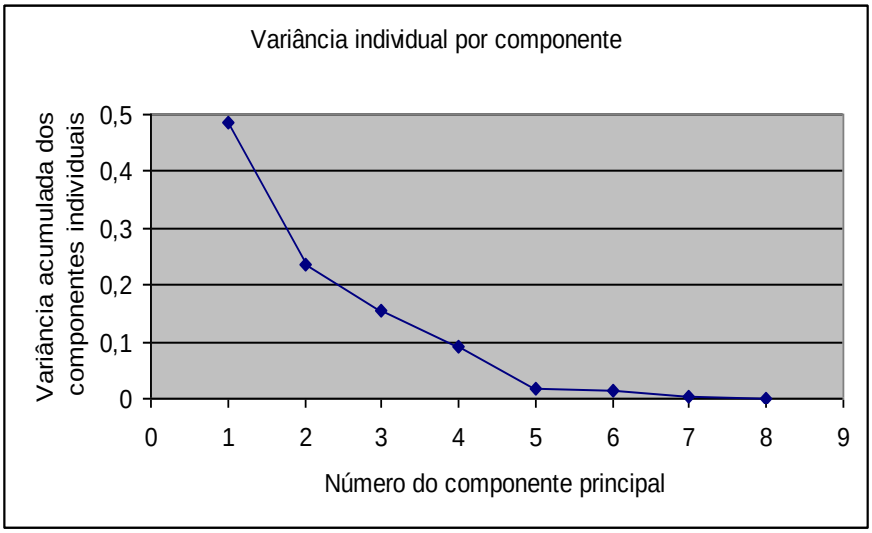


Figura 1 – Variância individual por componente.

Analizou-se os coeficientes de ponderação (autovetores), presente na Tabela 3, para um possível descarte de variáveis. A variável que possui maior correlação com o componente principal de menor variância (autovalor), deve ser menos importante para explicar a variância total. As variáveis passíveis de descarte em ordem crescente de importância foram: PESOLEITE2, PESOLEITE1, IDADE_PARTO e PESOLEITE, as demais variáveis devem ser mantidas. Barbosa et al. (2005) descartaram 17 das 33 variáveis; das características de carcaça de suínos, por apresentarem, a partir do último componente principal, maiores coeficientes.

Tabela 3 – Autovetores (coeficiente de ponderação) e suas correlações (em percentagem) para descarte de variáveis*

	Autovetor			
Variável	Prin5	Prin6	Prin7	Prin8
GRAUSANGUE	0,007607(0,289)	0,007793(0,266)	-0,018730(0,244)	0,000000(0)
PESOLEITE	0,009854(0,375)	-0,051677(-1,762)	0,021229(0,277)	**-0,854377(0)**
PESOLEITE1	-0,124353(-4,727)	**-0,788241(-26,875)**	0,024818(0,323)	0,344837(0)
PESOLEITE2	**0,757732(28,805)**	0,314919(10,737)	0,013122(0,171)	0,288205(0)
PESOLEITE3	-0,640422(-24,346)	0.524742(17,891)	0,022220(0,290)	0,260893(0)
IDADE_PESALEITE	0,002235(0,085)	0,029648(1,011)	0,700234(9,124)	0,000000(0)
IDADE_PARTO	-0,007955(-0,302)	0,021900(0,747)	**-0,712184(-9,279)**	0,000000(0)
IEP	-0,003899(-0,148)	0,007678(0,262)	-0,019565(-0,255)	0,000000(0)

*Correlações entre parênteses

O gráfico de dispersão mostrado na Figura 2, mostra a separação dos grupos genéticos produtivos. Utilizando-se os escores dos componentes principais 3 e 4, percebe-se que eles foram os que melhor representou a distribuição dos grupos genético. Oliveira (2004) observou grande heterogeneidade entre as características fenotípicas dos caprinos nos gráficos de dispersão e encontrou a melhor distribuição entre os escores nos componentes principais 1 e 3.

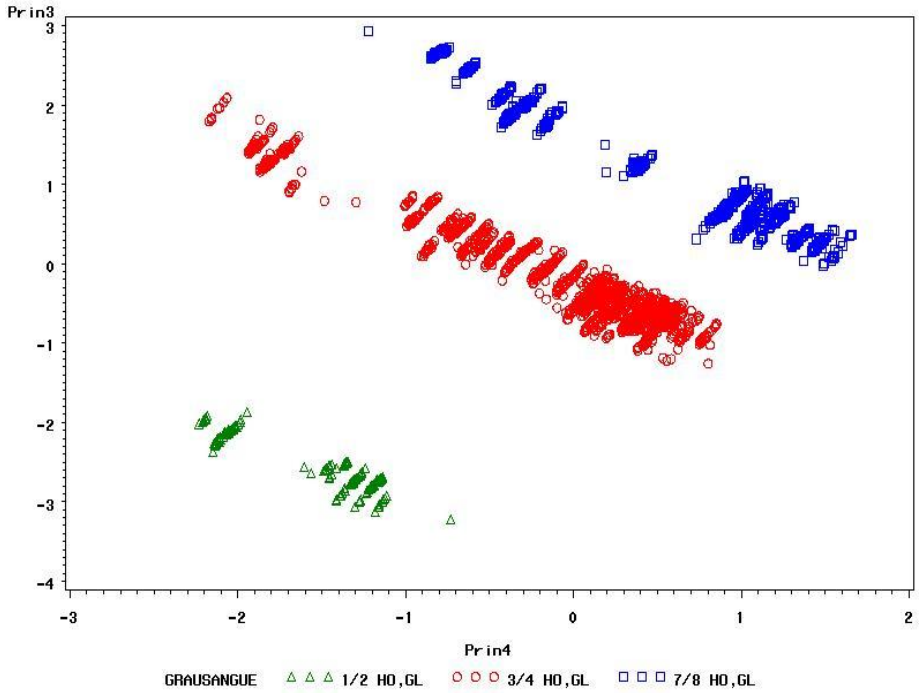


Figura 2 – Separação entre os três grupos genéticos produtivos através da dispersão entre os escores do terceiro (Prin3) e quarto (Prin4) componentes principais.

A análise de componentes principais foi refeita sem as variáveis PESOLEITE2, PESOLEITE1, IDADE_PARTO e PESOLEITE que foram descartadas. Pode-se verificar na Tabela 4 que três componentes principais puderam explicar 82,84% da variação dos dados. Não houve muita diferença da análise anterior (sem descarte de variáveis) onde três componentes retiveram 87,54%.

Tabela 4 - Autovalores, proporção individual e acumulada da variação dos dados através da análise dos componentes principais, com o descarte de variáveis

Componente	Autovalor	Proporção Individual	Proporção Acumulada
1	1,36818186	0,3420	0,3420
2	1,00627853	0,2516	0,5936
3	0,93896319	0,2347	0,8284
4	0,68657642	0,1716	1,0000

Devido as variáveis possuírem unidades de medidas diferentes padronizou-se as variáveis originais para uma distribuição Normal com média zero e variância um, procedendo à análise. Segundo Morrison (1976), em virtude dos coeficientes dos componentes principais serem influenciados pela escala das variáveis, recomenda-se utilizar variáveis padronizadas com variância igual à unidade.

Pode-se observar a matriz correlação na Tabela 5, verifica-se alta correlação positiva entre as variáveis, peso do leite no dia do controle com peso do leite na segunda e terceira ordenhas, e entre a idade da vaca no dia do controle e a idade da vaca ao parto. Obteve-se baixa correlação entre a idade da vaca no dia do controle em relação às características do peso do leite. Houve diferença no aumento da correlação da variável intervalo de partos com as variáveis relacionadas ao peso do leite, em relação à análise anterior, com os dados originais.

Tabela 5 - Matriz de correlação entre as variáveis padronizadas utilizadas na análise de componentes principais

	Variável							
Variável	GRAU SANGUE	PESO LEITE	PESO LEITE1	PESO LEITE2	PESO LEITE3	IDADE_ PESA LEITE	IDADE_ PARTO	IEP
GRAUSANGUE	1,0000							
PESOLEITE	0,2261	1,0000						

PESOLEITE1	0,0466	0,7296	1,0000					
PESOLEITE2	0,1442	0,9030	0,8117	1,0000				
PESOLEITE3	0,1522	0,9177	0,8134	0,8876	1,0000			
IDADE_ PESALEITE	-0,1453	0,0315	0,0215	0,0155	-0,0004	1,0000		
IDADE_ PARTO	-0,1636	0,1086	0,1062	0,1010	0,0848	0,9837	1,0000	
IEP	-0,1151	0,2416	0,2042	0,2286	0,2176	0,8824	0,8976	1,0000

Conforme pode-se observar na Tabela 6, houve redução das variáveis originais em três componentes, os quais retiveram 92,84% da variação dos dados, cinco componentes (62,5%) apresentaram variância (autovalor) inferior a 0,7. Assim os cinco componentes a partir do último são passíveis de descarte. A análise anterior (dados originais) explicou quase a mesma variação com quatro componentes principais, e com possibilidade de descarte de quatro componentes.

Tabela 6 - Autovalores, proporção individual e acumulada da variação dos dados através da análise dos componentes principais das variáveis padronizadas

Componente	Autovalor	Proporção	Acumulado
1	3.74037931	0.4675	0.4675
2	2.74062164	0.3426	0.8101
3	0.94619105	0.1183	0.9284
4	0.27217519	0.0340	0.9624
5	0.11721187	0.0147	0.9771
6	0.11290569	0.0141	0.9912
7	0.05877796	0.0073	0.9985
8	0.01173729	0.0015	1.0000

A Figura 3 mostra que a partir do terceiro componente a variância dos próximos componentes torna-se insignificante.

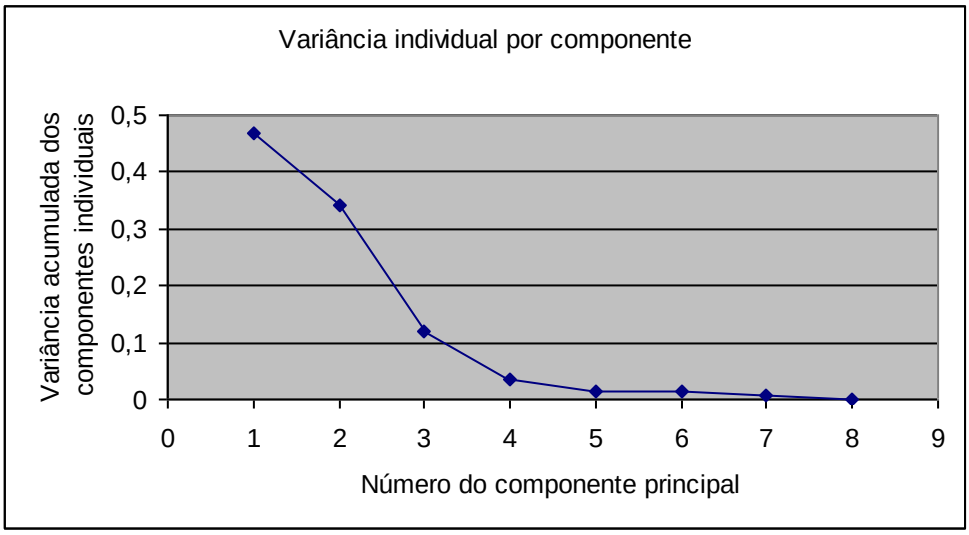


Figura 3 - Variância individual por componente (variáveis padronizadas).

As variáveis passíveis de descarte podem ser vistas na Tabela 7, aquelas que possuem maior correlação dos últimos cinco componentes, como é o caso das variáveis em ordem crescente de importância: PESOLEITE1, IEP, PESOLEITE2, PESOLEITE e IDADE_PARTO, as demais devem ser mantidas. Nesta análise foi eliminada a variável IEP, não vista na eliminação da análise anterior (dados originais). Com menos variáveis a serem analisadas, poupa-se relativa economia de tempo e custo, na tomada de novas medidas e em análises futuras.

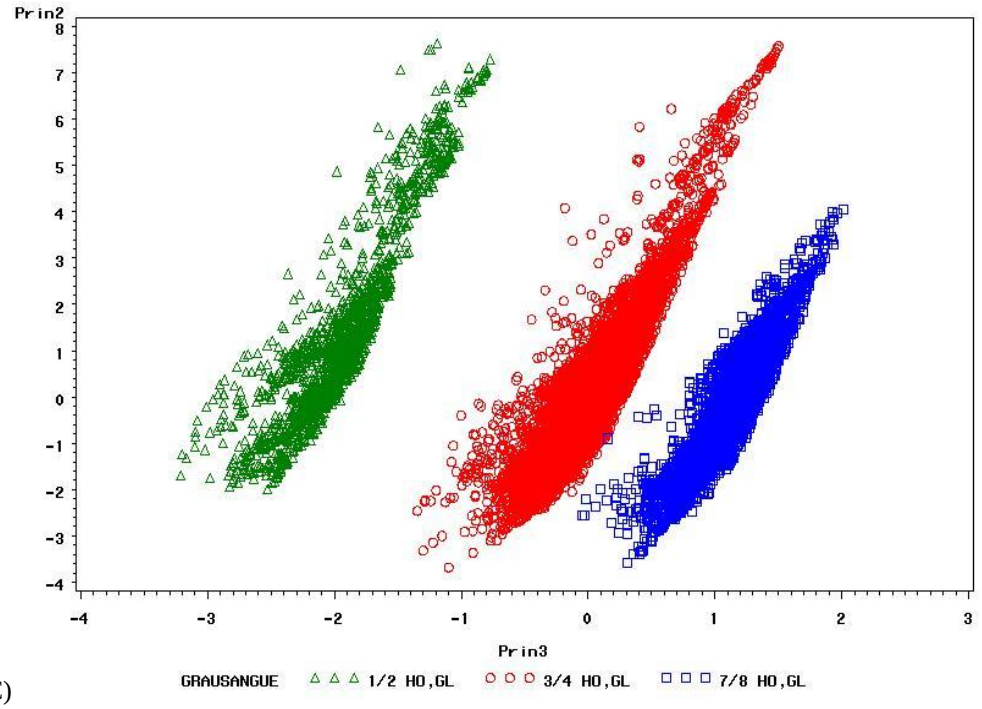
Tabela 7 – Autovetores (coeficiente de ponderação) e suas correlações (em percentagem) para eliminação dos dados (variáveis padronizadas)

	Prin4	Prin5	Prin6	Prin7	Prin8
GRAUSANGUE	0,169126(8,823)	-0,009113(-0,312)	-0,010782(-0,362)	0,042567(1,032)	-0,022178(0,240)

PESOLEITE	-0,478093(-24,942)	0,080545(2,758)	0,121790(4,092)	**-0,705077(-7,094)**	-0,008888(-0,096)
PESOLEITE1	**0,827590(43,176)**	-0,040032(-1,371)	0,043402(1,458)	-0,244348(-5,924)	0,018958(0,205)
PESOLEITE2	-0,141166(-7,365)	0,272604(9,333)	**-0,721816(-24,254)**	0,354850(8,603)	0,030430(0,330)
PESOLEITE3	-0,172256(-8,978)	-0,160736(-5,503)	0,608382(20,443)	0,559019(13,553)	0,030701(0,333)
IDADE_PESALEITE	0,041271(2,153)	0,365417(12,511)	0,122893(4,129)	0,006150(0,149)	0,704382(7,631)
IDADE_PARTO	0,050770(2,649)	0,371183(12,708)	0,119083(4,001)	0,046752(1,133)	**-0,707835(-7,669)**
IEP	-0,063542(-3,315)	**-0,787636(-26,966)**	-0,250471(-8,416)	-0,026556(-0,644)	-0,004132(-0,045)

*Valor da correlação entre parênteses

O gráfico de dispersão mostrado na Figura 2, foi obtido através dos escores dos componentes principais 2 e 3. Foi a que melhor representou a separação dos grupos genéticos produtivos. O gráfico da análise anterior (dados originais) foi composta dos escores 3 e 4.



(C)

Figura 4 – Separação entre os três grupos genéticos produtivos através da dispersão entre os escores do segundo (Prin2) e terceiro (Prin3) componentes principais (dados padronizados).

A análise de componentes principais foi refeita sem as variáveis PESOLEITE1, IEP, PESOLEITE2, PESOLEITE e IDADE_PARTO que foram descartadas. Pode-se verificar na Tabela 8 que dois componentes principais puderam explicar 73,67% da variação dos dados. Não houve muita melhora, pois na análise anterior, sem o descarte das variáveis, dois componentes retiveram 81,01%.

Tabela 8 - Autovalores, proporção individual e acumulada da variação dos dados através da análise dos componentes principais, com o descarte de variáveis (dados padronizados)

Componente	Autovalor	Proporção	Acumulado
1	1,21058361	0,4035	0,4035
2	0,99957186	0,3332	0,7367
3	0,78984453	0,2633	1,0000

Os resultados encontrados para a análise de componentes principais realizados com os dados originais e com a padronização deles foram divergentes no número de componentes retidos. A análise realizada com os dados transformados forneceu poucas componentes sem perder muitas informações, possibilitou o descarte de uma variável a mais, forneceu uma boa visualização gráfica da dispersão dos dados através dos escores dos componentes.

Conclusões

Com base nos resultados, a análise dos grupos produtivos em vacas leiteiras através da análise de componentes principais foi melhor apresentado com a padronização dos dados, do qual pode reduzir para três das oito componentes, eliminou o maior número de variáveis redundantes e obteve-se uma melhor visualização da separação dos grupos através das componentes retidas.

Agradecimentos

À UFRPE (Propesquisador 2003/2005) por proporcionar boas condições de trabalho, e ao proprietário da fazenda, pela colaboração e cessão dos dados.

SANTOS, E. F. N.; SANTORO, K. R.; FERREIRA, R. L. C.; SANTOS, E. S.; SANTOS, G. R. Formation of productive genetic groups in dairy cows through principal components.

Referências

BARBOSA, L.; LOPES. P. S.; REGAZZI, A. J.; GUIMARÃES, S. E. F.; TORRES, R. de. A. Avaliação de carcaça de suínos utilizando-se a análise dos componentes principais. Revista Brasileira de Zootecnia, v. 34, n. 6, p. 2209 – 2217, 2005 (supl.)

BARBOSA, L.; LOPES. P. S.; REGAZZI, A. J.; GUIMARÃES, S. E. F.; TORRES, R. de. A. Avaliação de características de qualidade da carne de suínos por meio de componentes principais. Revista Brasileira de Zootecnia, v. 35, n. 4, p. 1639 – 1645, 2006 (supl.)

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

DILLON, W. R.; GOLDSTEIN, M. Multivariate analysis. New York: John Wiley e Sons, 1984.

HAIR JR.; ANDERSON, R.; TATHAM, R.; BLACK, W. Análise multivariada de dados. 5 ed. Porto Alegre: Artmed, 2005. 593 p.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 4 ed. New Jersey: Pratice Hall, 1998. 816 p.

MANLY, B. F. J. Multivariate statistical methods a primer. 3 ed. London: Chapman & Hall/CRC, 2004. 208 p.

MARDIA, A. K. V.; KENT, J. T.; BIBBLY. J. M. Multivariate analysis. London: Academic Press, 1997, 518 p.

MACMANUS, C.; SAUERESSING, M. G.; FALCÃO, R. A.; SERRANO, G.; MARCELINO, K. R. A.; PALUDO, G. R. Componentes reprodutivos no rebanho de corte da Embrapa Cerrados. Revista Brasileira de Zootecnia, v. 31, n. 2, p. 648-657, 2002.

MORRISON, D. F. Multivariate statistical methods. 2 ed. New York: McGraw-Hill Company, 1976. 415 p.

OLIVEIRA, J. C. V. Caracterização e perfil etnológico de rebanhos caprinos nos municípios de Ibimirim e Serra Talhada, Estado de Pernambuco. 2004. 60 f. Dissertação (Mestrado em Zootecnia) – Universidade Federal Rural de Pernambuco, Recife, 2004.

REGAZZI, A. J. Analise multivariada. Viçosa: Universidade Federal de Viçosa, 2002. (INF-766). Notas de aula.

REIS, E. Estatística multivariada aplicada. 2 ed. Lisboa: Sílabo, 2001. 253 p.

SAS INSTITUTE. SAS Stat user's guide. Version 9. Carry: SAS Institute, 2002. CDROM.

USO DE DIFERENTES COMBINAÇÕES DE DISTÂNCIAS MULTIVARIADAS E MÉTODOS DE AGRUPAMENTO NA FORMAÇÃO DE GRUPOS PRODUTIVOS

Eucymara França Nunes SANTOS[1]
Kleber Régis SANTORO[2]
Rinaldo Luiz Caraciolo FERREIRA[3]
Eufrázio de Souza SANTOS[4]
Gladston Rafael de Arruda SANTOS[5]

- RESUMO: As distâncias multivariada: Euclidiana, Chebichev, Minkowski e Mahalanobis foram utilizadas juntamente aos métodos de agrupamento de ligação completa, ligação simples, ligação média, centróide e Ward com o objetivo de agrupar diferentes conjuntos produtivos de vacas leiteiras de três diferentes grupos genéticos: 1/2 HG; 3/4 HG; 7/8 HG. Inicialmente, os animais foram divididos em classes produtivas de quatro em quatro kilos referente ao peso do leite no dia do controle para todos os grupos genéticos. Para o cálculo das distâncias mencionadas utilizaram-se as variáveis: grupo genético, peso do leite (kg) produzido no dia do controle, peso do leite (kg) produzido na primeira, segunda e terceira ordenhas, idade da vaca (dias) ao parto, idade da vaca (dias) na data do controle leiteiro e intervalo de partos (dias). E para a construção dos dendogramas foram usadas às classes produtivas. Os resultados foram avaliados através do coeficiente cofenético, tabela de contingência, que indicará a porcentagem de grupos coincidentes, o teste qui-quadrado e o grau de associação. Constatou-se que a distância de Mahalanobis juntamente aos métodos de ligação média, ligação simples ou centróide são os mais indicados para agrupar os diferentes grupos genéticos.
- PALAVRAS-CHAVE: distâncias multivariada, métodos de agrupamento, grupos genéticos, classes produtivas, vacas leiteiras.

- *ABSTRACT: The many varieties of distances: Euclidiana, Chebichev, Minkowski and Mahalanobis were used with to the grouping methods of complete connection, simple connection, average connection, centroid and Ward with the objective of containing different productive groups of cows milkmaid from three different genetic groups: 1/2 HG; 3/4 HG; 7/8 HG. Initially, the animals were divided into productive classes of four in four kilos regarding to the weight of the milk in the day of the control for all the genetic groups. For the calculation of the mentioned distances the following variables were used: group genetic, weigh of the milk (kg) produced in the day of the control, weight of the milk (kg) produced in the first, second and third milking, age of the cow (days) to the childbirth, age of the cow (days) in the control milkmaid date and interval among childbirths (days). And for the construction of the dendograms were used to the productive classes. The results were appraised through the coefficient cofenetic, contingency table that will indicate the percentage of coincident groups, the qui-square test and the association level. It was verified that the distance of Mahalanobis, and the methods of average connection, simple connection or centroid, are the most suitable for containing the different genetic groups.*
- KEYWORDS: distance multivariate, methods of grouping, genetic groups, productive classes, cows milk.

1 Introdução

A estatística multivariada consiste em técnicas exploratórias de sintetização da estrutura da variabilidade dos dados, consistindo em analisar várias variáveis que são medidas simultaneamente numa mesma unidade experimental (Mingotti, 2005).

A análise de agrupamentos é uma das ferramentas para o desenvolvimento dessa tecnologia, consistindo em um conjunto de técnicas utilizadas na identificação de padrões de comportamento em banco de dados através da formação de grupos homogêneos. Estas técnicas têm sido utilizadas pelos melhoristas genéticos, em busca de identificar grupos produtivos e genéticos divergentes, para serem utilizados em programa de melhoramento genético, para estimar a diversidade genética de grupos de progenitores ou para proposição de sistemas de manejo (Cruz & Regazzi, 2007).

[1]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: eucymara@gmail.com
[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco – UFRPE - UAG, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: krsantoro@uag.ufrpe.br
[3] Departamento de Ciência Florestal, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: rinaldo@dcfl.ufrpe.br
[4] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: eufrazio@deinfo.ufrpe.br
[5] Empresa Pernambucana de Pesquisa Agropecuária – IPA, CEP:56.600-000 , Sertânia, Pernambuco, Brasil, E-mail: gladstonrafael@ipa.br

No Brasil, a maior parte da produção de leite é oriunda da utilização de mestiços zebuínos, especificamente Holandês x Gir, que é um dos mais importantes para a produção de leite no Estado de Pernambuco, gerando a necessidade de estudo no desempenho produtivo, de modo a contribuir para o conhecimento das potencialidades para a produção de leite e as possíveis estratégias para o melhoramento genético. O aumento da produtividade passa necessariamente pela busca de animais de valor genético superior para produção de leite, sendo utilizado em larga escala, o cruzamento das raças zebuínas (Faco, 2005).

Há uma grande necessidade de se conhecer a possibilidade de agrupamento de animais geneticamente divergentes, mas produtivamente semelhantes para apoiar e facilitar os programas de melhoramento genético, com isto objetivou-se formar agrupamentos utilizando os diferentes métodos de distância multivariada e métodos de agrupamento, avaliar a similaridade entre esses grupos e indicar a melhor distância e o melhor método a partir de grupos geneticamente divergentes.

2 Análise de Agrupamentos

A análise de agrupamentos também conhecida como conglomerados, classificação ou cluster, constitue-se em uma técnica a ser utilizada para a descoberta de uma estrutura de grupos e de relações existentes entre esses grupos. Tem como objetivo dividir os elementos da amostra, ou população, em grupos de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos.

Essa técnica permite extrair informações a respeito da independência entre as variáveis que caracterizam cada elemento. Procura-se elaborar critérios para agrupar objetos, dada uma amostra de n objetos cada um deles, medido segundo variáveis, procura-se um esquema de classificação que agrupe os objetos em k grupos.

Os grupos são formados com base na similaridade ou dissimilaridade (distâncias), reconhecer entre eles um grau de semelhança suficiente para reuní-los num mesmo conjunto, destacando os grupos de objetos similares entre si, segundo suas características ou variáveis. O grau de associação tem que ser elevado entre os membros de uma mesma categoria, e baixo, entre os elementos de categorias distintas (Valentin, 2000).

Segundo Reis (2001), as etapas de uma análise de agrupamentos são as seguintes:

1) Seleção de um indivíduo ou de uma amostra de indivíduos a serem agrupados.
2) Definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento de indivíduos.
3) Coleta dos dados que serão reunidas numa tabela com m colunas (descritores) e n linhas (objetos).
4) Escolha de um critério de similaridade ou dissimilaridade.
5) Adoção e execução de um algoritmo de agrupamento.
6) Elaboração e interpretação do dendograma.
7) Validação dos resultados encontrados.

Um conceito fundamental é a escolha de um critério que meça a distância entre dois objetos, ou que quantifique o quanto eles são parecidos, esta medida é chamada coeficiente de parecença, sendo dividida em duas categorias: medidas de similaridade (quanto maior o valor, maior a semelhança entre os objetos) e de dissimilaridade (quanto maior o valor, maior a diferença entre os objetos) (Bussab, 1990). Existem várias medidas diferentes e cada uma delas produz um determinado tipo de agrupamento. As medidas mais comuns, apropriadas para variáveis quantitativas, são as de dissimilaridade entre elementos de uma matriz de dados e as mais utilizadas são:

1) Distância Euclidiana – a distância entre dois elementos (x_i e x_j) é a raiz quadrada do somatório dos quadrados das diferenças entre valores de i e j para todas as variáveis (v = 1, 2, ..., p).

$$d(x_i, x_j) = \sqrt{\sum_{v=1}^{p} (x_{iv} - x_{jv})^2} \quad (1)$$

2) Distância Minkowski – a distância entre dois elementos (x_i e x_j), $i \neq j$ é definido por:

$$d(x_i, x_j) = \left[\sum_{h=1}^{v} w_h \left| x_{hi} - x_{hj} \right|^{\lambda} \right]^{\frac{1}{\lambda}} \quad (2)$$

onde w_h's são os pesos ponderados para as variáveis

3) Distância de Mahalanobis – a distância generalizada entre dois indivíduos x_i e x_j, $i \neq j$ é definida por:

$$d(x_i, x_j) = (x_i - x_j)' S^{-1} (x_i - x_j) \quad (3)$$

onde S é a matriz de covariância.

4) Distância de Chebichev – a distância entre dois indivíduos xi e xj é o valor máximo para todas as variáveis, das diferenças entre dois indivíduos, é definida por:

$$d(x_i - x_j) = máx\left|x_{iv} - x_{jv}\right| \quad (4)$$

Os métodos de agrupamento têm o objetivo de reunir entidades em grupos homogêneos. O método de classificação denominado hierárquico consiste em reunir indivíduos em grupos, e o processo repete-se em diferentes neveis até formar uma árvore chamada dendograma. Segundo Chatfield (1997), o tipo de algoritmo é chamado algoritmo aglomerativo porque ele opera por uma série de união, começando com n grupos de apenas um indíviduo e termina em um grupo de n indivíduos, os mais utilizados são:

1) Ligação simples – esse método tem seu procedimento iniciado com a procura dos dois objetos mais similares na matriz de dissimilaridade. Ele consiste em reconhecer os indivíduos mais próximos os quais são reunidos formando o grupo inicial, calcula-se então a distância daquele grupo em relação aos demais indivíduos e nos estágios mais avançados. A distância entre os progenitores k e um grupo formado pelos progenitores i e j é dada por:

$$d_{(ij)k} = \min (d_{ik}; d_{jk}) \quad (5)$$

2) Ligação completa – este método, após agrupar os dois indivíduos mais semelhantes de menor distância, verifica a distância máxima deste primeiro grupo para os objetos restantes, Frei (2006). As distâncias entre um grupo e um indivíduo devem ser calculadas pela expressão:

$$d_{(ij)k} = \text{máx} (d_{ik}; d_{jk}) \quad (6)$$

3) Ligação média – este método trata a distância entre dois conglomerados como a média das distâncias entre todos os pares de elementos que podem ser formados com os elementos dos dois conglomerados que estão sendo comparados. As distâncias entre o objeto que se quer incluir num grupo são:

$$d(C_1, C_2) = \sum_{i \in C_1} \sum_{j \in C_2} \left(\frac{1}{n_i n_j} \right) d(x_i, x_j) \quad (7)$$

onde C_1 e C_2 são grupos aos quais pertencem os elementos n_i e n_j, e estes são os números de elementos nos grupamentos i e j.

4) Centróide – este método define a coordenada de cada grupo como sendo a média das coordenadas de seus objetos. A distância entre os grupos é obtida através do cálculo das distâncias entre os centróides.

5) Ward – neste método um grupo será reunido a um outro se essa união proporcionar o aumento da variância intragrupo, Valentini (2000). A variância intragrupo será calculada para todas as alternativas de aglomeração, escolhendo a que proporcionará a menor variância, sendo aplicado a todos os passos da análise. A distância é definida como:

$$d\left(x_i, x_j\right) = \left(\frac{n_i n_j}{n_i + n_j} \right) \left(\bar{x}_i - \bar{x}_j\right)' \left(\bar{x}_i - \bar{x}_j\right) \quad (8)$$

onde n_i e n_j são os números de elementos nos grupamentos i e j.

A aplicação dos métodos hierárquicos permite a apresentação dos resultados sob a forma de dendograma. O dendograma é um diagrama em forma de árvore que mostra a subdivisão dos grupos formados, buscando máxima homogeneidade entre os indivíduos no grupo e máxima heterogeneidade entre os grupos (Sneath & Sokal, 1973).

3 Materiais e Métodos

Neste trabalho foram utilizados dados provenientes de cruzamento entre animais Holandês (HO) e Gir (GL) de vacas leiteiras de três diferentes grupos genéticos: 1/2 HO, GL; 3/4 HO, GL e 7/8 HO, GL, com 326 animais

por categoria. Os dados foram coletados semanalmente no período de dezembro/2000 a dezembro/2006 numa fazenda com agropecuária semi-intensiva de leite, na região de Ribeirão, Zona da Mata de Pernambuco.

Analizou-se as seguintes características de produção de leite: grupo genético (GRAUSANGUE), peso do leite (Kg) produzido no dia do controle (PESOLEITE), peso do leite (Kg) produzido na primeira ordenha (PESOLEITE1), peso do leite (Kg) produzido na segunda ordenha (PESOLEITE2), peso do leite (Kg) produzido na terceira ordenha (PESOLEITE3), idade da vaca (dias) na data do controle (IDADEPESALEITE), idade da vaca (dias) ao parto (IDADEPARTO) e intervalo de partos (IEP), com 13.643, 13.643, 13.600, 12.128, 9.643, 13.643, 13.643 e 3.119 observações por variável, respectivamente.

Com a finalidade de tornar os elementos amostrais mais convenientes, foi realizada uma análise descritiva (Tabela 1) para construir classes produtivas de animais, referentes ao peso do leite produzido no dia do controle. Para construir os intervalos de classe, utilizou-se a fórmula de Sturges:

$$h = \frac{A}{1 + 3{,}322 \log N} \tag{9}$$

onde A é a amplitude, calculada através da diferença entre o valor máximo e o valor mínimo, de leite produzido pelas vacas dos três grupos genético juntos, e N é o número total de observações.

Utilizou-se um intervalo de quatro em quatro quilos para os diferentes grupos genéticos.

Tabela 1 – Análise descritiva para a produção total do leite (kg) no dia do controle para os diferentes grupos genéticos

	Geral	1/2 HG	3/4 HG	7/8 HG
Número de observações	13.643	1.903	8.261	3.479
Valor mínimo	1,0	1,0	1,0	1,0
Valor máximo	46,40	29,80	46,40	38,80
Média	15,08	10,31	15,90	15,72
Desvio padrão	7,33	4,87	7,55	6,91
Coeficiente de variação	48,62	47,26	47,50	43,96
Número de classes*		8	12	10

*Divididas de quatro em quatro kilos

Para a análise de agrupamentos adotou-se as distâncias: Euclidiana, Minkowski, Mahalanobis e Chebichev como medidas de dissimilaridade, sobre as quais empregou-se os métodos de agrupamentos hierárquicos: ligação completa, ligação simples, ligação média, centróide e Ward. Estas distâncias e métodos serão combinados resultando em 20 combinações, que resultarão em 20 dendogramas. Os dendogramas foram representados pela classe produtiva, para a determinação do número de grupos foi utilizada uma distância de similaridade de 0,5.

Os agrupamentos serão avaliados pelo número e pela homogeneidade dos grupos. Serão avaliados também pelo coeficiente de correlação cofenética (coeficiente r de Pearson), como medida de validação para o grau de ajuste entre a matriz original e a matriz de agrupamento (cofenética) obtida após a construção do dendograma. Considera-se aceitável um coeficiente cofenético superior a 0,7, sendo que o maior valor encontrado na análise é o critério de escolha do melhor método de agrupamento.

Os resultados serão comparados através de uma tabela de contingência usada para registrar e analisar o relacionamento entre duas variáveis, que indicará o número de indivíduos alocados num mesmo cluster. O teste qui-quadrado será utilizado para verificar a existência de independência entre as características e mostrar o nível de significância para cada par de resultado através da fórmula:

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$

onde *fo* é a freqüência observada e *fe* é a freqüência esperada.

O coeficiente de associação foi utilizado como complementação do teste qui-quadrado para medir o grau de associação entre as variáveis, pode variar entre 0 e 1, sendo que quanto mais próximo de 1, maior é o nível de associação entre as variáveis. Tal coeficiente é determinado pela seguinte fórmula:

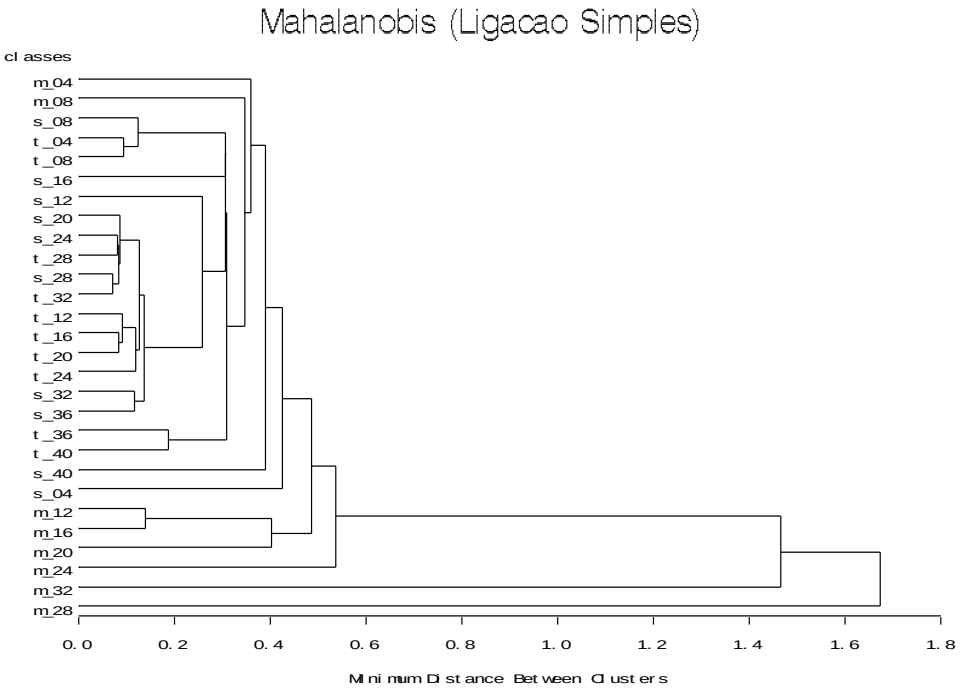
$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

onde N é o valor total da tabela de contingência e χ^2 é o valor do teste qui-quadrado

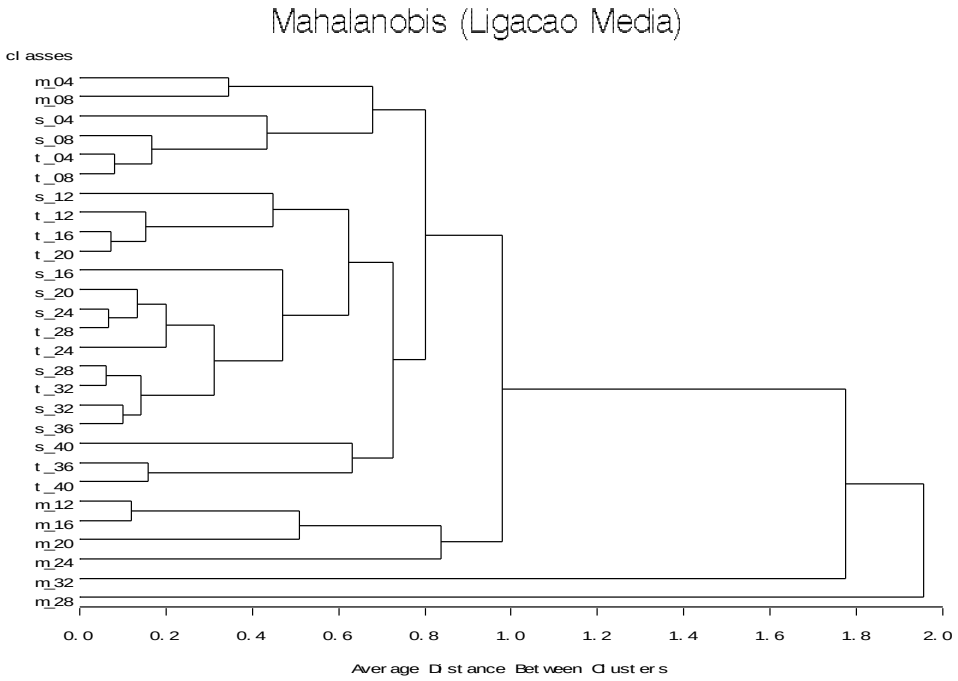
As análises foram realizadas através dos softwares R 2007 (versão 2.7.0) e SAS 8.02, conforme Johnson e Wichern (2007).

4 Resultados e Discussão

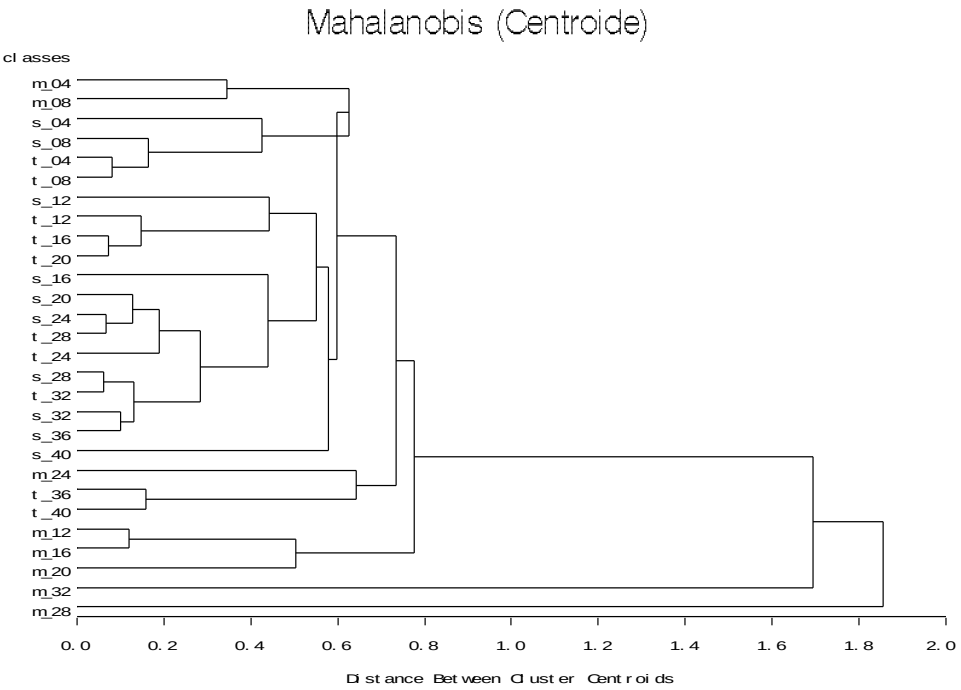
Os dendogramas (Figura 1) foram formados utilizando as distâncias: Euclidiana, Chebichev, Minkowski e Mahalanobis, cada uma delas utilizando os métodos de agrupamento: ligação completa, ligação simples, ligação média, centróide e Ward, totalizando 20 dendogramas com estruturas diferenciadas. Os grupos foram representados pela classe produtiva. Como critério para determinar a formação dos agrupamentos empregou-se uma distância de similaridade de 0,50, para determinar uma quantidade significativa de grupos, feito igualmente a todos os dendogramas para efeito de comparação. Albuquerque (2005) mostrou significativa estabilidade entre os métodos de: ligação simples, ligação completa, centróide, mediana, média das distâncias e Ward, relacionando-os com a distância de Mahalanobis, em dados de vegetação.



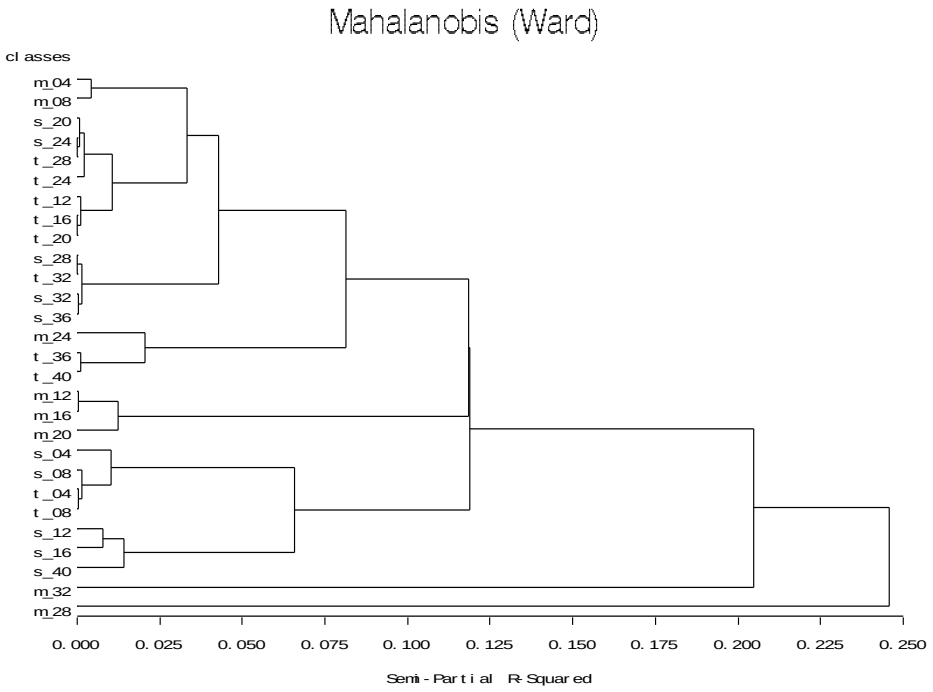
(A)


Mahalanobis (Ligacao Media)
classes
m_04
m_08
s_04
s_08
t_04
t_08
s_12
t_12
t_16
t_20
s_16
s_20
s_24
t_28
t_24
s_28
t_32
s_32
s_36
s_40
t_36
t_40
m_12
m_16
m_20
m_24
m_32
m_28
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0
Average Distance Between Clusters

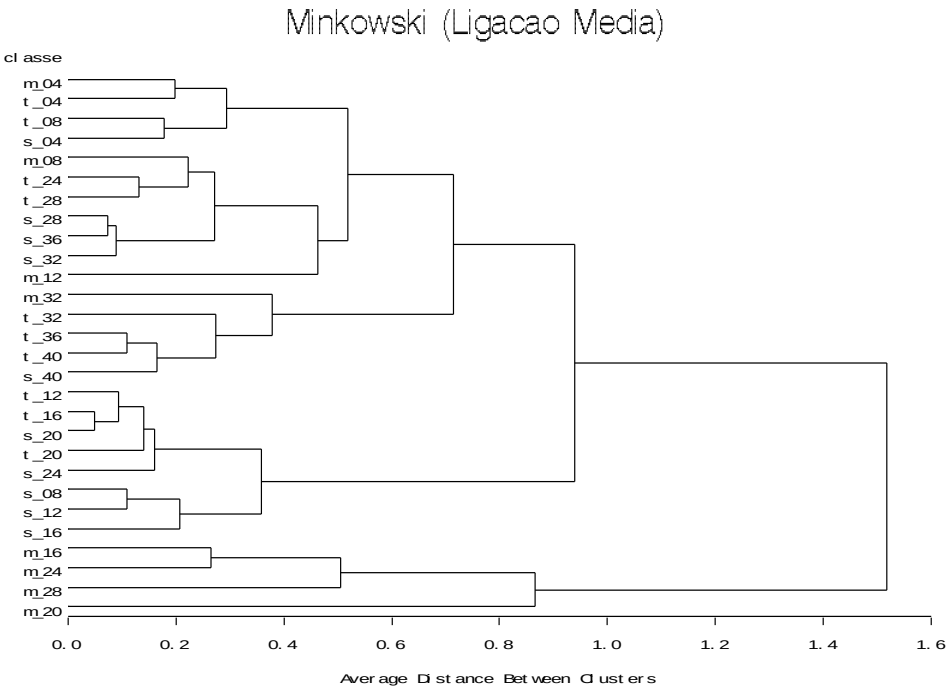
(B)


Mahalanobis (Centroide)
classes
m_04
m_08
s_04
s_08
t_04
t_08
s_12
t_12
t_16
t_20
s_16
s_20
s_24
t_28
t_24
s_28
t_32
s_32
s_36
s_40
m_24
t_36
t_40
m_12
m_16
m_20
m_32
m_28
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0
Distance Between Cluster Centroids

(C)



(D)



(E)

Figura 1 – Dendogramas obtidos através das medidas de distância e métodos de agrupamentos, (A) Mahalanobis – Ligação Simples, (B) Mahalanobis – Ligação Média, (C) Mahalanobis – Centróide, (D) Mahalanobis – Ward, (E) Minkowski – Ligação Média. Com "m" representando o 1/2 HG; "t" o 3/4 HG; "s" o 7/8 HG juntamente ao número máximo de produção total do leite (kg) no dia do controle.

Tabela 2 – Número de agrupamentos formados das distâncias e métodos de agrupamentos utilizados

	Distâncias de Dissimilaridade			
Métodos de Agrupamento	Euclidiana	Chebichev	Minkowski	Mahalanobis

Ligação completa	9	8	10	15
Ligação simples	2	2	2	4
Ligação média	6	5	7	11
Centróide	4	4	5	11
Ward	2	2	1	1

Observou-se que os grupos formados (Tabela 2) pela distância e método de agrupamento: Mahalanobis – ligação completa, Minkowski – ligação completa, Mahalanobis – ligação média e Mahalanobis – centróide apresentaram agrupamentos mais homogêneos ao unir animais com classes produtivas mais semelhantes em um mesmo grupo, sendo as mais conservadoras por agrupar animais mais parecidos produtivamente sob o ponto de vista de melhoramento genético.

A correlação cofenética (Tabela 3) mostra que o método Ward para todas as medidas de distância utilizadas não obteve um bom ajuste da matriz de dissimilaridade na forma de dendograma com valores abaixo do considerado aceitável, assim como para as distâncias – métodos: Chebichev – ligação completa, Chebichev – ligação simples e Minkowski – ligação completa. O melhor ajuste foi para a distância Mahalanobis utilizada com o método centróide.

Comparando-se os métodos de ligação simples, ligação completa e ligação média entre grupos, através da distância Euclidiana, Totti et al. (2001) concluíram que o método da ligação média foi o mais eficiente, em acessos de *Paspalum*, pois apresentou o maior valor da correlação cofenética e a melhor representação no dendograma. Utilizando o método de agrupamento de Ward, Giannotti et al. (2005) almejavam conseguir homogeneidade, de acordo com as características de crescimento em bovino de corte da raça Nelore, dentro de todos os cinco grupos, no entanto essa homogeneidade só foi confirmada em apenas três grupos.

Tabela 3 – Coeficiente de correlação cofenética das distâncias e métodos de agrupamentos utilizados

	Distâncias de Dissimilaridade			
Métodos de Agrupamento	Euclidiana	Chebichev	Minkowski	Mahalanobis
Ligação média	0,76	0,75	0,76	0,85
Ward	0,60	0,62	0,66	0,58
Centróide	0,74	0,74	0,74	0,88
Ligação completa	0,76	0,65	0,68	0,73
Ligação simples	0,72	0,68	0,71	0,83

Os resultados obtidos através da tabela de contingência (Tabela 4) identificaram que os métodos de agrupamentos: ligação completa, ligação média e centróide para a distância Euclidiana possuem 40% de semelhança, não são dependentes ao nível de 5% de significância e um baixo nível de associação. Para a distância de Chebichev os métodos de ligação média e centróide tiveram 67% de semelhança mas não são dependentes. A distância de Minkowski obteve 47% de semelhança e independência entre os métodos de ligação completa e ligação média. Houve 100% de semelhança dos métodos de ligação média e Ward com alto nível de associação e dependência a um nível de significância de 1% na distância de Mahalanobis.

Tabela 4 – Porcentagem de grupos coincidentes, qui-quadrado e grau de associação das distâncias multivariadas e métodos de agrupamentos hierárquicos (dentro das distâncias)

Euclidiana	LC[a]	LS	LM	C
	18% [b]			
LS	$1,68^{ns}$			
	0,36			
	40%	25%		
LM	$0,42^{ns}$	$0,89^{ns}$		
	0,17	0,32		
	15%	33%	40%	
C	$5,31^{*}$	$0,37^{ns}$	$0,28^{ns}$	
	0,54	0,24	0,17	
	0%	0%	0%	33%
W	$11,10^{**}$	$4,00^{*}$	$8,00^{**}$	$0,37^{ns}$
	0,71	0,71	0,71	0,24
Minkowski	LC	LS	LM	C

Chebichev	LC	LS	LM	C
	20%			
LS	$1,41^{ns}$			
	0,35			
	15%	29%		
LM	$5,94^{*}$	$0,63^{ns}$		
	0,56	0,29		
	17%	33%	67%	
C	$4,71^{*}$	$0,37^{ns}$	$1,11^{ns}$	
	0,53	0,24	0,33	
	0%	0%	0%	0%
W	$10,00^{**}$	$4,00^{*}$	$7,02^{**}$	$5,96^{*}$
	0,71	0,71	0,71	0,71
Mahalanobis	LC	LS	LM	C

	17%					32%				
LS	1,95^{ns}				LS	0,05^{ns}				
	0,37					0,05				
	47%	22%				62%	40%			
LM	0,01^{ns}	1,17^{ns}			LM	1,76^{ns}	0,0085^{ns}			
	0,02	0,34				0,25	0,024			
	13%	29%	17%			62%	40%	100%		
C	7,33**	0,63^{ns}	5,20*		C	1,76^{ns}	0,0085^{ns}	22^{ns}		
	0,57	0,29	0,55			0,25	0,024	0,71		
	0%	0%	0%	0%		0%	0%	0%	0%	
W	11,09**	3,04^{ns}	7,71**	5,88*	W	16,67**	16,67**	12,50**	12,50**	
	0,71	0,71	0,70	0,70		0,71	0,71	0,71	0,71	

[a]Métodos de agrupamento: LC=ligação completa; LS=ligação simples; LM=ligação média; C=centróide; W=Ward
[b]Em ordem: porcentagem de grupos coincidentes entre os métodos; valor e significância do teste de qui-quadrado (^{ns} (não-significativo); * e ** (significativo a 5 e 1% de probabilidade respectivamente)**); g**rau de associação

O método de ligação simples foi 100% semelhante entre as distâncias: Euclidiana e Chebichev; Euclidiana e Minkowski; Chebichev e Minkowski, dependentes ao nível de significância de 5%, também com o método do centróide com a distância Euclidiana e Chebichev sendo dependentes a 1% e o método de Ward com as distâncias Minkowski e Mahalanobis não sendo dependentes, mas com um alto nível de associação, sendo explicado por serem calculados por diferentes métodos e possuírem a mesma quantidade de grupos (Tabela 5).

Tabela 5 – Porcentagem de grupos coincidentes, qui-quadrado e grau de associação das combinações das distâncias miltivariadas e métodos de agrupamentos hierárquicos (entre distâncias)

Euclidiana /Chebichev	LC[a]	LS	LM	C	W
	47%[b]	20%	43%	17%	0%
LC	0,05^{ns}	1,41^{ns}	0,22^{ns}	4,71*	10,00**
	0,054	0,35	0,12	0,53	0,71
	18%	100%	25%	33%	0%
LS	1,69^{ns}	4,00*	0,89^{ns}	0,37^{ns}	4,00*
	0,36	0,71	0,32	0,24	0,71
	14%	29%	36%	67%	0%
LM	6,62*	0,63^{ns}	0,79^{ns}	1,11^{ns}	7,02**
	0,57	0,29	0,26	0,33	0,71
	15%	33%	40%	100%	33%
C	5,31*	0,37^{ns}	0,28^{ns}	8,00**	0,37^{ns}
	0,54	0,24	0,17	0,71	0,24
	0%	0%	0%	0%	0%
W	11,12**	4,00*	8,00**	5,96*	4,00*
	0,71	0,71	0,71	0,71	0,71
Euclidiana /Mahalanobis	LC	LS	LM	C	W
	17%	12%	10%	11%	0%
LC	9,96**	3,11^{ns}	12,39**	8,90**	16,66**
	0,54	0,39	0,61	0,56	0,70
	15%	0%	0%	0%	0%
LS	5,31*	5,96*	10,00**	8,00**	5,96*
	0,54	0,71	0,71	0,71	0,71
	20%	15%	12%	13%	0%
LM	7,10**	2,15^{ns}	9,35**	6,49*	12,90**
	0,51	0,38	0,60	0,55	0,71
	20%	15%	12%	13%	0%
C	7,10**	2,15^{ns}	9,35**	6,49*	12,90**
	0,51	0,38	0,60	0,55	0,71

	0%	0%	0%	0%	0%
W	10,00**	3,04ns	7,15**	5,00*	3,04ns
	0,71	0,71	0,71	0,71	0,71
Euclidiana /Minkowski	LC	LS	LM	C	W
	63%	17%	25%	14%	0%
LC	1,34ns	1,95ns	3,48ns	5,94*	12,13**
	0,26	0,37	0,42	0,55	0,71
	18%	100%	25%	33%	0%
LS	1,69ns	4,00*	0,89ns	0,37ns	4,00*
	0,36	0,71	0,32	0,24	0,71
	38%	22%	15%	18%	0%
LM	0,91ns	0,17ns	6,19*	4,08*	9,10**
	0,23	0,34	0,57	0,52	0,71
	14%	29%	36%	44%	0%
C	6,62*	0,63ns	0,79ns	0,09ns	7,02**
	0,57	0,29	0,26	0,10	0,71
	0%	0%	0%	0%	0%
W	10,00**	3,04ns	7,15**	5,00*	3,04ns
	0,71	0,71	0,71	0,71	0,71
Chebichev /Minkowski	LC	LS	LM	C	W
	22%	17%	13%	14%	0%
LC	5,42*	1,95ns	7,33**	5,94*	12,13**
	0,48	0,37	0,57	0,55	0,71
	20%	100%	29%	33%	0%
LS	1,41ns	4,00*	0,63ns	0,37ns	4,00*
	0,35	0,71	0,29	0,24	0,71
	13%	22%	17%	18%	0%
LM	8,02**	1,17ns	5,20*	4,08*	9,10**
	0,59	0,34	0,55	0,52	0,71
	15%	29%	40%	44%	0%
C	5,94*	0,63ns	0,40ns	0,09ns	7,02**
	0,56	0,29	0,20	0,10	0,71
	0%	0%	0%	0%	0%
W	9,09**	3,04ns	5,88*	5,00*	3,04ns
	0,71	0,71	0,70	0,71	0,71
Chebichev /Mahalanobis	LC	LS	LM	C	W
	8,7%	12%	10%	11%	0%
LC	15,05**	3,11ns	10,76**	8,90**	16,66**
	0,63	0,39	0,59	0,56	0,70
	12%	0%	0%	0%	0%
LS	12,04**	5,96*	8,98**	8,00**	5,96*
	0,71	0,71	0,71	0,71	0,71
	11%	15%	13%	13%	0%
LM	11,67**	2,15ns	8,07**	6,49*	12,90**
	0,62	0,38	0,58	0,55	0,71
	11%	15%	13%	13%	0%
C	11,67**	2,15ns	8,07**	6,49*	12,90**
	0,62	0,38	0,58	0,55	0,71
	0%	0%	0%	0%	0%
W	9,09**	3,04ns	5,88*	5,00*	3,04ns
	0,71	0,71	0,70	0,71	0,71
Minkowski /Mahalanobis	LC	LS	LM	C	W
	24%	12%	18%	10%	18%
LC	6,25*	3,11ns	7,41**	10,76*0,59	7,41**
	0,45	0,39	0,50		0,50
	29%	0%	18%	0%	0%

LS	1,27^{ns}	5,96*	4,08*	8,98**	5,00*
	0,29	0,71	0,52	0,71	0,71
	29%	15%	22%	13%	0%
LM	3,84^{ns}	2,15^{ns}	5,11*	8,07**	12,50**
	0,39	0,38	0,47	0,58	0,71
	29%	15%	22%	13%	0%
C	3,84^{ns}	2,15^{ns}	5,11*	8,07**	12,50**
	0,39	0,38	0,47	0,58	0,71
	0%	0%	0%	0%	100%
W	11,11**	3,04^{ns}	7,71**	5,88*	2,00^{ns}
	0,71	0,71	0,70	0,70	0,71

[a]Métodos de agrupamento: LC=ligação completa; LS=ligação simples; LM=ligação média; C=centróide; W=Ward
[b]Em ordem: porcentagem de grupos coincidentes entre os métodos; valor e significância do teste de qui-quadrado (^{ns} (não-significativo); * e ** (significativo a 5 e 1% de probabilidade respectivamente)); grau de associação

Pode-se verificar que cada análise realizada fornece seu resultado mais adequado. As metodologias de julgamento para a melhor combinação distância-método de agrupamento não foram concordantes entre si. Sendo que a forma conclusiva, não haveria uma melhor combinação. Na inspeção visual dos gráficos (Figura 1 e Tabela 2) observou-se que o mais adequado ao objetivo foi a distância de Minkowski e o método de ligação média. A análise do coeficiente cofenético (Tabela 2) indicou a distância de Mahalanobis e o método do centróide. A tabela qui-quadrado informa que entre os dois métodos citados anteriormente existe uma dependência de 5% de significância, 22% de coincidência e um coeficiente de 0,47 de associação.

Em um programa de melhoramento genético e por apoio ao manejo produtivo na propriedade, quando utilizada a distância Euclidiana, os métodos de agrupamento mais indicado é o de ligação completa e ligação média, quando a distância é a de Chebichev o mais indicado é o de ligação média e centróide, quando a distância é a de Minkowski indica-se o método de ligação média e quando utilizada a distância de Mahalanobis o método mais adequado são os métodos de ligação média e centróide.

Conclusões

A distância de Mahalanobis, utilizando-se os métodos de ligação média, ligação simples ou centróide, seria então mais apropriada por apresentar agrupamentos mais homogêneos, unir classes produtivas mais semelhantes, melhor representação no dendograma e maior valor do coeficiente cofenético. Fatores que melhoram a produtividade dos sistemas de produção de leite e contribuem para as possíveis estratégias no melhoramento genético.

Agradecimentos

À UFRPE (Propesquisador 2003/2005) por proporcionar boas condições de trabalho e ao proprietário da fazenda pela colaboração e cessão dos dados.

SANTOS, E. F. N.; SANTORO, K. R.; FERREIRA, R. L. C.; SANTOS, E. S.; SANTOS, G. R. Use of different combinations of multivariate distance and clustering methods in formation of productive groups in dairy cows.

Referências

ALBUQUEQUE, M. A. Estabilidade em análise de agrupamento. 2005. 64 f. Dissertação (Mestrado em Biometria) – Universidade Federal Rural de Pernambuco, Recife, 2005.

BUSSAB, W. DE O.; MIAZAKI, E. S.; ANDRADE, D. F. DE. Introdução à análise de agrupamentos, São Paulo. Associação Brasileira de estatística. 9º Simpósio Nacional de Probabilidade e Estatística, 1990. 105 p.

CHATFIELD, C.; COLLINS, A. J. Introduction to multivariate analysis. London. Chapman e Hall, 1997.

CRUZ, C. D.; REGAZZI, A. J. Modelos biométricos aplicados ao melhoramento genético. 2 ed. Viçosa: Editora UFV, 2007. 390 p.

DILLON, W. R.; GOLDSTEIN, M. Multivariate analysis. New York: John Wiley e Sons, 1984.

FACÓ, O.; LÔBO, R. N. B.; MARTINS FILHO, R.; LIMA, A. M. Idade ao primeiro parto e intervalo de partos de cinco grupos genéticos holandês x gir no Brasil. Revista Brasileira Zootecnia, v. 34, n. 6, p. 1920-1926, 2005.

FREI, F. Introdução à análise de agrupamentos. São Paulo: Editora UNESP, 2006. 111 p.

GIANNOTTI, A. D. G.; PACKER, I. U.; MERCADANTE, M. E. Z.; LIMA, C. G. Análise de agrupamento para implementação da meta-análise em estimativas de herdabilidade para características de crescimento em bovino de corte. Revista Brasileira de Zootecnia, v. 34, n. 4, p. 1165-1172, 2005.

JOHNSON, R. A.; WICHERN, D. W. Applied multivariate statistical analysis. 4 ed. New Jersey: Pratice Hall, 1998. 816 p.

MINGOTTI, S. A. Análise de dados através de métodos de estatística multivariada. Belo Horizonte: Editora UFMG, 2005, 295 p.

REIS, E. Estatística multivariada aplicada. 2 ed. Lisboa: Sílabo, 2001. 253 p.

SAS INSTIRUTE. SAS Stat user's guide. Version 9. Carry: SAS Institute, 2002. CDROM.

SNEATH, P. A.; SOKAL, R. R. Numeric taxonomy: the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973. 573 p.

TOTTI, R.; VENCOVSJY, R.; BATISTA, L. A. R. Utilização de métodos de agrupamentos hierárquicos em acessos de Paspalum (Gramínea (Poaceal)). Semina, v. 22, p. 25-35, 2001.

VALENTIN, J. L. Ecologia numérica: Uma introdução à análise multivariada de dados ecológicos. Rio de Janeiro: Editora Interciência, 2000. 117 p.

CLASSIFICAÇÃO DE INDIVÍDUOS EM GRUPOS GENÉTICOS PRODUTIVOS EM VACAS LEITEIRAS ATRAVÉS DE ANÁLISE DISCRIMINANTE

Eucymara França Nunes SANTOS[1]
Kleber Régis SANTORO[2]
Rinaldo Luiz Caraciolo FERREIRA[3]
Eufrázio de Souza SANTOS[4]
Gladston Rafael de Arruda SANTOS[5]

- RESUMO: A análise discriminante foi utilizada para selecionar as variáveis que mais contribuem para a diferença dos grupos. Utilizar os gráficos de dispersão da variável canônica para ajudar na discriminação dos grupos, verificar a correta classificação dos grupos e fornecer equações que possibilite a inclusão de novos indivíduos em três diferentes grupos genéticos (1/2 HG, 3/4 HG e 7/8 HG). As variáveis usadas foram: grupo genético, peso do leite (kg) produzido no dia do controle, peso do leite (kg) produzido na primeira, segunda e terceira ordenhas, idade da vaca (dias) ao parto, idade da vaca (dias) na data do controle leiteiro e intervalo de partos (dias). Foram utilizados dados originais e com padronização. A análise de seleção STEPDISC do SAS eliminou a variável peso do leite (kg) produzido no dia do controle. As variáveis mais importantes em todas as análises foram: idade da vaca (dias) na data do controle leiteiro e intervalo de partos (dias). O grupo genético 1/2 HG foi o que obteve a maior percentagem de classificação correta. A função discriminante apresentada foi referente aos dados padronizados por possuírem maior poder de classificação.
- PALAVRAS-CHAVE: análise discriminante, grupos genéticos, classificação

- *ABSTRACT: The discriminant analysis was used to select the variables that most contributes to the difference of the groups. The graph of dispersion of the canonical variable were used to help in the discrimination of the groups, to verify the correct classification of the groups and to supply equations that makes possible the inclusion of new animals in three different genetic groups (1/2 HG, 3/4 HG and 7/8 HG. The used variables were: group genetic, weigh of the milk (kg) produced in the day of the control, weight of the milk (kg) produced in the first it milks, weigh of the milk (kg) produced in the second it milks, weigh of the milk (kg) produced in the third it milks, age of the cow (days) in the date of the control, age of the cow to the childbirth and interval of childbirths. Original data were used with standardization. The selection analysis STEPDISC of SAS eliminated the variable weigh of the milk (kg) produced in the day of the control. The most important variables in all the analyses were: age of the cow (days) in the date of the control and interval of childbirths. The genetic group 1/2 HG was the one which has gotten the largest percentage of correct classification. The discriminant presented function was regarding to the standardized data by possessing the largest power classification.*
- *KEYWORDS: discriminant analysis, genetic groups, classification*

1 Introdução

A análise discriminante é um método estatístico utilizado para classificação de elementos de uma amostra ou população através de combinações das variáveis (funções) que melhor discriminam grupos definidos a priori, de tal forma que seja minimizada a probabilidade de classificação incorreta a posteriori. Emprega-se esta técnica para descobrir características que distinguem os membros de um grupo dos de outro, de modo que conhecendo as características de um novo indivíduo se possa prever a qual grupo ele pertence. Para tal são estimadas as funções discriminantes utilizadas para previsão de pertença de um indivíduo não agrupado (REIS, 2001).

Segundo Cruz e Regazzi (1997) a técnica estatística de análise discriminante é uma alternativa para combinar as múltiplas informações contidas na unidade experimental, de modo que seja possível identificar as características dos grupos geneticamente divergentes.

[1]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: eucymara@gmail.com
[2]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco – UFRPE - UAG, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: krsantoro@uag.ufrpe.br
[3] Departamento de Ciência Florestal, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: rinaldo@dcfl.ufrpe.br
[4] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - UFRPE, CEP: 52.171-900, Recife, Pernambuco, Brasil, E-mail: eufrazio@deinfo.ufrpe.br
[5] Empresa Pernambucana de Pesquisa Agropecuária – IPA, CEP:56.600-000 , Sertânia, Pernambuco, Brasil, E-mail: gladstonrafael@ipa.br

De acordo com Dias et al. (1997) a técnica pode ser utilizada em estudos de divergência genética, pois proporciona enriquecimento das informações extraídas dos dados experimentais. A seleção da análise mais adequada é função da precisão desejada, da facilidade de análise e da maneira como os dados foram obtidos.

Assim, o presente trabalho foi conduzido com os objetivos de identificar as características que mais contribuem na discriminação dos grupos e estabelecer funções discriminantes que possibilite a inclusão de novos indivíduos, utilizando os dados originais e com a pressuposição de normalidade.

2 Análise Discriminante

A análise discriminante surgiu para distinguir estatisticamente entre dois ou mais grupos de indivíduos, previamente definidos a partir de características conhecidas para todos os membros do grupo.

O objetivo da análise discriminante é encontrar uma combinação linear das várias variáveis independentes que minimize a probabilidade de incorreta classificação dos indivíduos. Pressupõe-se que as variáveis independentes têm distribuição conjunta normal multivariada enquanto o dependente é fixa e do tipo nominal.

Segundo Mingoti (2005), o princípio da máxima verossimilhança pode ser aplicado mesmo em situações em que a distribuição amostral envolvida não é normal. Em análise discriminante nos deparamos frequentemente com problemas de violação das pressuposições de normalidade, que podem exercer influência indevida sobre o resultado na análise. Um fator importante na análise discriminante é a pressuposição de normalidade, que quase nunca ocorre na prática. Seber (1984) diz que uma transformação apropriada pode frequentemente produzir um conjunto de dados que segue uma distribuição normal, possibilitando o uso da técnica baseada na suposição de normalidade.

Seja a função linear discriminante para a alocação de uma nova observação (X_0) nas populações π_i(i = 1,2)

$$D(x_0)=[\mu_1-\mu_2]'\Sigma^{-1}x_0 \quad (1)$$

onde μ_1 *e* μ_2 são os vetores de média para as duas populações e Σ é a matriz de covariância.

A regra de classificação é alocar x_0 em:

$$\Pi_1 \text{ se } D(x_0) - m \geq 0 \quad (2)$$

$$\Pi_2 \text{ se } D(x_0) - m < 0 \quad (3)$$

onde m é o ponto médio entre:

$$D(x_1) =[\mu_1-\mu_2]'\Sigma^{-1}x_1 \quad (4)$$

e

$$D(x_2) =[\mu_1-\mu_2]'\Sigma^{-1}x_2 \quad (5)$$

O procedimento STEPDISC do SAS (SAS, 2002) executa uma análise discriminante do tipo seleção stepwise, escolhe as variáveis que irão sair ou permanecer de acordo com o teste F e a correlação quadrada parcial quadrada (R^2). O teste F parcial checa se a variável absorverá uma quantidade significante da variação em relação àquela removida por variáveis, avalia se o valor F obtido em um dado passo satisfaz o mínimo, e se satisfizer, a variável entrará. Em relação a correlação quadrada parcial, o algoritmo seleciona, do grupo de variáveis remanescentes, aquela que dá a maior redução na variância residual (não explicada), isto é, a variável cuja correlação parcial é maior. A cada passo entra uma variável que contribui com o maior poder de discriminação e a variável que contribui menos ao poder distintivo é afastada.

A análise discriminante canônica é uma técnica que permite a redução da dimensionalidade de dados e possibilita a identificação de populações similares no espaço bi ou tridimensional. Esse procedimento procura, com base em um conjunto de características originais correlacionadas, transforma-lás em características padronizadas e não correlacionadas. A importância relativa de cada variável canônica é dada pela razão entre a variância por ela explicada e o total da variância escolhida, sendo que a importância decresce da primeira para a última, pois as últimas variáveis são responsáveis pela explicação de uma fração mínima da variância total disponível (CRUZ, 1990). Quando há, nas primeiras variáveis, a concentração de grande proporção da variância total, em torno de 80%, elas podem ser utilizadas para ilustrar graficamente as posições relativas e as orientações dos grupos.

3 Materiais e Métodos

Neste trabalho foram utilizados dados provenientes de cruzamento entre animais Holandês (HO) e Gir (GL) de vacas leiteiras de três diferentes grupos genéticos: 1/2 HO, GL; 3/4 HO, GL e 7/8 HO, GL, com 326 animais por categoria. Os dados foram coletados semanalmente no período de dezembro/2000 a dezembro/2006 numa fazenda com agropecuária semi-intensiva de leite, na região de Ribeirão, Zona da Mata de Pernambuco.

Analizou-se as seguintes características de produção de leite: grupo genético (GRAUSANGUE), peso do leite (Kg) produzido no dia do controle (PESOLEITE), peso do leite (Kg) produzido na primeira ordenha (PESOLEITE1), peso do leite (Kg) produzido na segunda ordenha (PESOLEITE2), peso do leite (Kg) produzido na terceira ordenha (PESOLEITE3), idade da vaca (dias) na data do controle (IDADE_PESALEITE), idade da vaca (dias) ao parto (IDADE_PARTO) e intervalo de partos (IEP), com 13.643, 13.643, 13.600, 12.128, 9.643, 13.643, 13.643 e 3.119 observações por variável, respectivamente.

Para avaliar a relação entre as características de produção de leite, os dados foram inicialmente submetidos ao processo de seleção das variáveis através do procedimento PROC STEPDISC, do SAS (SAS, 2002) com seleção stepwise para selecionar as variáveis que melhor discriminem os grupos.

A análise de variáveis canônicas foi utilizada com o propósito de possibilitar a visualização bidimensional de grupos similares e identificar as variáveis mais importantes na divergência entre os grupos.

Foi utilizada a análise discriminante para verificar a consistência dos grupos formados, obtendo funções que permitam classificar um indivíduo x_0, com base nas características do mesmo, buscando minimizar a probabilidade de má classificação de um indivíduo em um grupo πi, quando ele realmente pertence ao grupo π_j, $(i \neq j)$.

Um fator importante em análise discriminante é a pressuposição de normalidade, então utilizou-se o conjunto de dados com e sem padronização da distribuição Normal para efeito de comparação

As análises foram realizadas através do software SAS 8.02 – Statistical Analysis System (SAS, 2001), utilizando os procedimentos PROC STEPDISC, PROC CANDISC e PROC DISCRIM da análise discriminante.

4 Resultados e Discussão

O processo de seleção das variáveis que contribuíram para a máxima explicação da variância foi analisado por meio da seleção Stepwise do procedimento Stepdisc (Tabela 1). Foram analisados um total de sete características da produção de leite, seis delas foram selecionadas, as quais potencialmente contribuíram para explicar à variação dos grupos, observou-se a retirada da variável PESOLEITE. Araújo (2004) selecionou 27 das 49 espécies de formigas, através do mesmo processo, como representativas da comunidade local, baseando-se na sua maior contribuição individual para explicação da variância total.

Tabela 1 – Resumo da seleção STEPWISE do PROC STEPDISC (SAS, 2002)

Passo	Variável selecionada	R^2	Valor de F
1	PESOLEITE2	0,0339	54,06
2	IDADE_PARTO	0,0305	48,41
3	IDADE_PESALEITE	0,0290	45,08
4	IEP	0,0121	18,89
5	PESOLEITE1	0,0111	17,33
6	PESOLEITE3	0,0111	17,34

O teste de normalidade foi realizado através do procedimento UNIVARIATE NORMAL (SAS, 2002), que gerou através do teste de Kolmogorov-Smirnov, índices entre 0,065 e 0,239, evidenciando a distribuição normal das características da produção de leite.

A seleção das variáveis seguindo a pressuposição da normalidade dos dados (Tabela 2) não excluiu nenhuma das variáveis de acordo com os critérios da correlação quadrada parcial e do teste F.

Tabela 2 – Resumo da seleção STEPWISE do PROC STEPDISC com os dados padronizados

Passo	Variável selecionada	R^2	Valor de F
1	PESOLEITE	0,0685	501,79
2	IDADE_PARTO	0,0388	274,93
3	IDADE_PESALEITE	0,0486	348,02
4	PESOLEITE3	0,0180	124,69
5	PESOLEITE1	0,0049	33,30
6	IEP	0,0027	18,45
7	PESOLEITE2	0,0007	4,88

A análise de variáveis canônicas indicou diferenças significativas entre os grupos nos dois casos, para os dados originais ($P < 0,0001$) e para os dados padronizados ($P < 0,0001$).

Conforme pode-se verificar nas Tabelas 3 e 4, o processo de extrair as variáveis canônicas será igual ao número de variáveis originais ou o número de classes menos um, seja qual for o menor. Em ambos os casos a

análise proporcionou duas variáveis canônicas referentes ao número de três grupos menos um, podendo assim representar toda a variação no gráfico de dispersão bidimensional.

A divergência genética entre quatro linhas de matrizes de frango estudada por Carneiro (2002), por meio de variáveis canônicas, verificou-se que as duas primeiras variáveis canônicas explicaram mais de 96% da variação total nos três períodos.

Tabela 3 - Autovalores, proporção individual e acumulada da variação dos dados através da análise de variáveis canônicas dos dados originais

Variável Canônica	Autovalor	Proporção Individual	Proporção Acumulada
1	0,1153	0,8490	0,8490
2	0,0205	0,1510	1,0000

Tabela 4 - Autovalores, proporção individual e acumulada da variação dos dados através da análise de variáveis canônicas dos dados padronizados

Variável Canônica	Autovalor	Proporção Individual	Proporção Acumulada
1	0,1916	0,9435	0,9435
2	0,0115	0,0565	1,0000

As Figuras 1 e 2 ilustram os gráficos de dispersão dos grupos representados pelos escores das variáveis canônicas. Na Figura 1, o efeito conjunto das variáveis foi capaz de capturar uma leve variação entre os grupos 1/2 HG, dos demais (3/4 HG e 7/8 HG). Na Figura 2 a análise de variáveis canônicas não foi capaz de capturar variância significativa entre os grupos. Pode-se concluir que as sete características não são suficientes para discriminar indivíduos da população analisada em três classes diferentes. Uma nova análise de variáveis canônicas foi realizada sem a variável PESOLEITE pelo fato da variável ter sido eliminada por meio da seleção Stepwise, constatando que não houve diferença na explicação da variabilidade e no gráfico de dispersão. A inspeção gráfica dos escores das duas variáveis canônicas em Messeti (2004) detectou a existência de divergência genotípica entre as 12 populações de girrasol, com base nas 12 variáveis agronômicas.

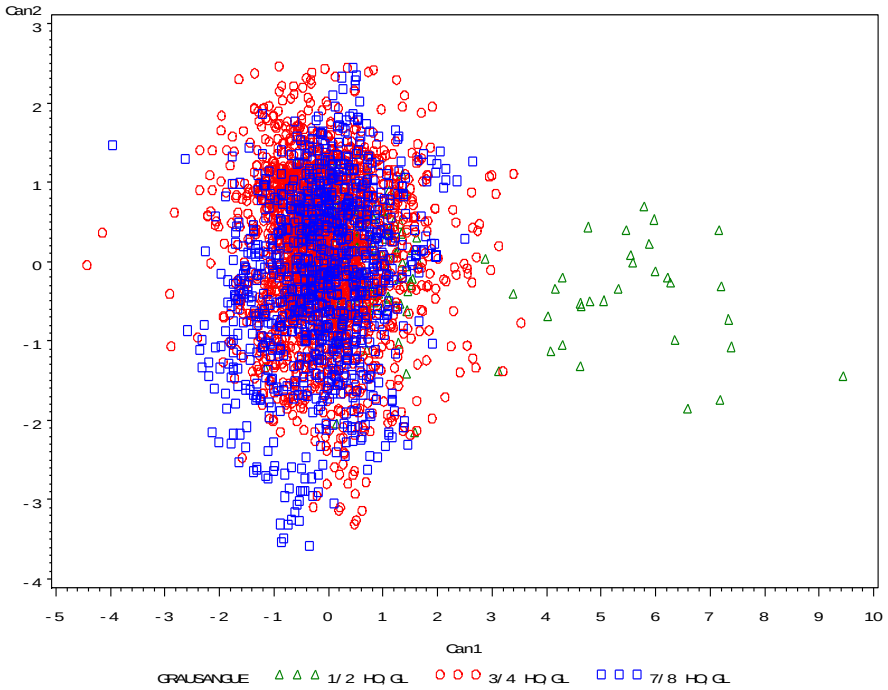


Figura 1 – Dispersão gráfica dos três grupos genéticos produtivos entre os escores das duas variáveis canônicas: Can1 x Can2 para os dados originais.

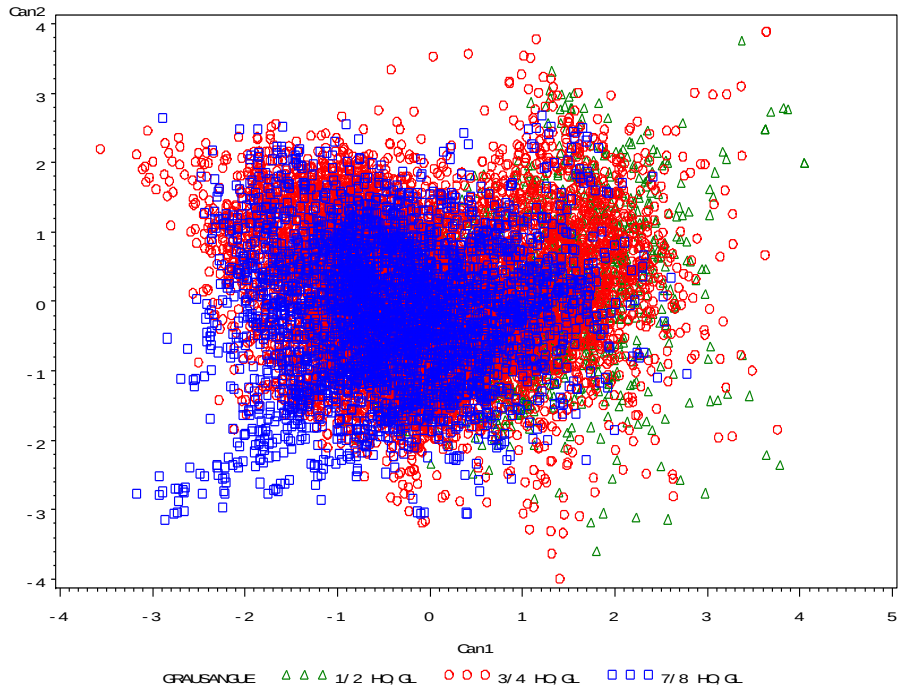


Figura 2 – Dispersão gráfica dos três grupos genéticos produtivos entre os escores das duas variáveis canônicas: Can1 x Can2 para os dados transformados.

Para identificar as variáveis mais importantes no contraste entre os grupos, a análise discriminante canônica forneceu as correlações entre as variáveis originais e as variáveis canônicas. Os coeficientes associados a primeira e a segunda variáveis canônicas, foram listados nas Tabelas 5 e 6. Observa-se que as variáveis IDADE_PARTO e IDADE_PESALEITE são responsáveis pela explicação da maior parte da variação nas duas variáveis canônicas em qualquer uma das análises.

Tabela 5 – Coeficientes padronizados das variáveis canônicas dos dados originais sem a variável PESOLEITE

Variável	Can1	Can2
PESOLEITE1	-0,79	0,13
PESOLEITE2	0,89	0,32
PESOLEITE3	0,71	-0,34
IDADE_ PESALEITE	2,09	-3,36
IDADE_PARTO	2,57	3,12
IEP	0,10	0,75

Tabela 6 – Coeficientes padronizados das variáveis canônicas dos dados padronizados com todas as variáveis

Variável	Can1	Can2
PESOLEITE	-1,81	0,08
PESOLEITE1	0,25	0,46
PESOLEITE2	0,18	-0,07
PESOLEITE3	0,58	0,35
IDADE_ PESALEITE	-2,78	-2,66
IDADE_PARTO	3,23	1,96
IEP	0,06	1,01

A análise discriminante forneceu a classificação dos três grupos genéticos, obtida com as características da produção de leite, conforme apresentado na Tabela 7 (dados originais) e Tabela 8 (dados padronizados). Os animais 3/4 HG são os mais difíceis de classificar corretamente. Pinto (2005) gerou funções e classificou potros e potras da raça Mangalarga Machador, com percentual de acerto entre 60,5 a 100%.

Tabela 7 – Classificação dos três grupos genéticos, em função das características da produção de leite para os dados originais

Grupo observado	1/2HG	3/4HG	7/8HG	TOTAL
1/2HG	96	64	32	192
%	50	33,33	16,67	100
3/4HG	292	951	672	1915
%	15,25	49,66	35,09	100
7/8HG	185	358	436	979
%	18,9	36,57	44,67	100
TOTAL	573	1373	1140	3086

Tabela 8 – Classificação dos três grupos genéticos, em função das características da produção de leite para os dados padronizados

Grupo observado	1/2HG	3/4HG	7/8HG	TOTAL
1/2HG	1468	227	208	1903
%	77,14	11,93	10,93	100
3/4HG	2257	2347	3657	8261
%	27,32	28,41	44,27	100
7/8HG	706	905	1868	3479
%	20,39	26,01	53,69	100
TOTAL	4431	3479	5733	13643

A Tabela 9 apresenta os parâmetros que permitem elaborar as equações, de acordo com os grupos para inclusão de novos animais, possibilitando o descarte antecipado daqueles que não pertencem à correta classificação do grupo, referente à análise discriminante realizada com os dados normalizados por apresentar uma maior percentagem de classificação correta.

Tabela 9 – Parâmetros da função discriminante linear de Fisher com os dados padronizados

Grupo observado	1/2HG	3/4HG	7/8HG
CONSTANTE	-1,29647	-0,20448	-0,13063
PESOLEITE	-3,48213	-1,49509	-1,09181
PESOLEITE1	0,34786	0,16297	0,01571
PESOLEITE2	0,88144	0,66418	0,63764
PESOLEITE3	1,52064	0,88770	0,65419
IDADE_ PESALEITE	-1,66125	0,57189	1,63153
IDADE_ PARTO	2,17438	-0,56625	-1,58307
IEP	0,03951	0,05528	0,00424

Conclusões

A análise discriminante proporcionou a seleção das variáveis que mais contribuem para a separação dos grupos, por apresentarem diferenças significativas, e as variáveis mais importantes na explicação da maior parte da variabilidade. O emprego do gráfico de dispersão por variáveis canônicas não foi eficiente para separar os três grupos genéticos. Apresentou-se os parâmetros das equações que possibilitaram a inclusão de novos animais com a mínima probabilidade de classificação incorreta.

Agradecimentos

À UFRPE (Propesquisador 2003/2005) por proporcionar boas condições de trabalho, e ao proprietário da fazenda, pela colaboração e cessão dos dados.

SANTOS, E. F. N.; SANTORO, K. R.; FERREIRA, R. L. C.; SANTOS, E. S.; SANTOS, G. R. Classification of individuals in productive genetic groups in dairy cows through discriminant analysis.

Referências

ARAÚJO, M. DA. S.; LUCIA, T. M. C. D.; VEIGA, C. E. DA.; NASCIMENTO, I. C. DO. Efeito da queima da palhada de cana-de-açúcar sobre comunidade de formicídeos. Associación Argentina de Ecologia, n. 14, p. 191-200, 2004.

CARNEIRO, P. L. S.; FONSECA, R.; PIRES, A. V.; TORRES FILHO, R. A.; TORRES, R. A.; PEIXOTO, J. O.; LOPES, P. S.; EUCLYDES, R. F. Estudo da divergência genética entre linhagens de matrizes de frango de corte por meio de análise multivariada. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, v. 34, n. 1, 2002.

CRUZ, C. D. Aplicação de algumas técnicas multivariadas no melhoramento de plantas. Piracicaba: ESALQ, 1990. 188 p. Tese (Doutorado em Melhoramento Genético) – Escola Superior de Agricultura "Luiz de Queiroz", 1990.

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

DIAS, L. A. S.; KAGEYAMA, P. Y.; LIRA, M. A. et al. Cultivo da palma forrageira em Pernambuco. Recife: Empresa Pernambucana de Pesquisa Agropecuária IPA, 1984. 5 p. (Documento 21).

MESSETTI, A. V. L.; PADOVANI, C. R. O uso da dispersão gráfica por variáveis canônicas com ênfase em melhoramento genético. 49ª Reunião da RBRAS, p. 373-376. Minas Gerais, 2004.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada. Belo Horizonte: Editora UFMG, 2005. 295 p.

PINTO, L. F. B.; ALMEIDA, F. Q. DE.; QUIRINO, C. R.; CABRAL, G. C.; AZEVEDO, P. C. N. DE.; SANTOS, E. M. Análise multivariada das medidas morfométricas de potros da raça mangalarga marchador: Análise discriminante. Revista Brasileira de Zootecnia, v. 34, n. 2, p. 600-612, 2005.

REIS, E. Estatística multivariada aplicada. 2 ed. Lisboa: Sílabo, 2001. 253 p.

SAS INSTITUTE. SAS Stat user's guide. Version 9. Carry: SAS Institute, 2002. CDROM.

SEBER, G. A. F. Multivariate observations. 1 ed. John Wiley, Canadá. 1984.

ANEXO I

Normas para preparação de trabalhos científicos para publicação na Revista Brasileira de Zootecnia

Escopo e política

A **Revista Brasileira de Zootecnia** (RBZ) é uma publicação mensal da Sociedade Brasileira de Zootecnia (SBZ), com o objetivo de publicar artigos originais nas áreas de Aqüicultura; Forragicultura; Melhoramento, Genética e Reprodução; Monogástricos; Produção Animal; Ruminantes; e Sistemas de Produção e Agronegócio.

No processo de publicação, os trabalhos técnico-científicos são avaliados por revisores ad hoc, indicados pelo Conselho Científico, composto por especialistas com doutorado nas diferentes áreas de interesse, e coordenados pela Comissão Editorial da RBZ. A política editorial da RBZ consiste em manter o alto padrão científico das publicações, por intermédio de colaboradores de renomada conduta ética e elevado nível técnico.

Só serão aceitos trabalhos escritos em português ou inglês e que não foram publicados nem submetidos à publicação em outro veículo. Deve-se ressaltar que isto não se aplica a resumos expandidos. Os trabalhos fracionados ou subdivididos em partes devem ser encaminhados juntos, pois serão submetidos aos mesmos revisores.

O conteúdo dos artigos publicados na Revista Brasileira de Zootecnia é de exclusiva responsabilidade de seus respectivos autores.

Encaminhamento de trabalhos

A **RBZ** publica artigos científicos originais nas áreas de Aqüicultura, Forragicultura, Melhoramento, Genética e Reprodução, Monogástricos, Produção Animal, Ruminantes, e Sistemas de Produção e Agronegócio.

O envio dos manuscritos é feito exclusivamente pela *home page* da RBZ (www.sbz.org.br), link Revista, juntamente com a carta de encaminhamento, conforme instruções no link "Envie seu manuscrito".

O pagamento da taxa de tramitação (pré-requisito para emissão do número de protocolo), no valor de R$ 30,00 (trinta reais), deverá ser realizado por meio de boleto bancário, disponível na *home page* da SBZ (www.sbz.org.br).

Uma vez aprovado o artigo, será cobrada uma taxa de publicação, que, no ano de 2009, para assinantes da RBZ, será de R$ 90,00 (noventa reais) para artigos em português e R$ 180,00 (cento e oitenta reais) para artigos em inglês com até oito páginas no formato final. Serão cobrados ainda, por página excedente, R$ 40,00 (quarenta reais) para artigos em português e R$ 80,00 (oitenta reais) para artigos em inglês. Entretanto, se entre os autores houver algum não assinante (exceto co-autores que não militam na área zootécnica, desde que não seja o primeiro autor), serão cobrados valores diferenciados (consultar link "Instruções aos autores" na *home page* da RBZ).

Forma e preparação de trabalhos

Os trabalhos já publicados ou sob consideração em qualquer outra publicação não serão aceitos. Ressalta-se que esta norma não é válida para resumos expandidos.

Só serão aceitos trabalhos escritos em português ou inglês.

O texto deve ser elaborado segundo as normas da RBZ e orientações disponíveis no link "Instruções aos autores".

Formatação de texto

O texto deve ser digitado em fonte Times New Roman 12, espaço duplo (exceto Resumo, Abstract e Tabelas, que devem ser elaborados em espaço 1,5), margens superior, inferior, esquerda e direita de 2,5; 2,5; 3,5; e 2,5 cm, respectivamente.

Pode conter até 25 páginas, numeradas seqüencialmente em algarismos arábicos.

As páginas devem apresentar linhas numeradas (a numeração é feita da seguinte forma: MENU ARQUIVO/CONFIGURAR PÁGINA/LAYOUT/NÚMEROS DE LINHA.../ NUMERAR LINHAS), com paginação contínua e centralizada no rodapé.

Estrutura do artigo

O artigo deve ser dividido em seções com cabeçalho centralizado, em negrito, na seguinte ordem: Resumo, Abstract, Introdução, Material e Métodos, Resultados e Discussão, Conclusões, Agradecimento e Literatura Citada.

Não serão aceitos cabeçalhos de terceira ordem.

Os parágrafos devem iniciar a 1,0 cm da margem esquerda.

Título

Deve ser preciso e informativo. Quinze palavras são o ideal e 25, o máximo. Digitá-lo em negrito e centralizado, segundo o exemplo: Valor nutritivo da cana-de-açúcar para bovinos em crescimento. Indicar sempre a entidade financiadora da pesquisa, como primeira chamada de rodapé numerada.

Autores

Deve-se listar até seis autores. A primeira letra de cada nome/sobrenome deve ser maiúscula (Ex.: Anacleto José Benevenutto). Não listá-los apenas com as iniciais e o último sobrenome (Ex.: A.J. Benevenutto).

Outras pessoas que auxiliaram na condução do experimento e/ou preparação/ avaliação do trabalho devem ser mencionadas em Agradecimento.

Resumo

Deve conter no máximo 1.800 caracteres com espaço. As informações do resumo devem ser precisas e informativas. Resumos extensos serão devolvidos para adequação às normas.

Deve sumarizar objetivos, material e métodos, resultados e conclusões. Não deve conter introdução. Referências nunca devem ser citadas no resumo.

O texto deve ser justificado e digitado em parágrafo único e espaço 1,5, começando por RESUMO, iniciado a 1,0 cm da margem esquerda.

Abstract

Deve aparecer obrigatoriamente na segunda página e ser redigido em inglês científico, evitando-se sua tradução por meio de aplicativos comerciais.

O texto deve ser justificado e digitado em espaço 1,5, começando por ABSTRACT, em parágrafo único, iniciado a 1,0 cm da margem esquerda.

Palavras-chave e Key Words

Apresentar até seis (6) palavras-chave e Key Words imediatamente após o RESUMO e ABSTRACT, respectivamente, em ordem alfabética. Devem ser elaboradas de modo que o trabalho seja rapidamente resgatado nas pesquisas bibliográficas. Não podem ser retiradas do título do artigo. Digitá-las em letras minúsculas, com alinhamento justificado e separado por vírgulas. Não devem conter ponto final.

Introdução

Deve conter no máximo 2.500 caracteres com espaço.

Deve-se evitar a citação de várias referências para o mesmo assunto.

Trabalhos com introdução extensa serão devolvidos para adequação às normas.

Material e Métodos

Descrição clara e com referência específica original para todos os procedimentos biológicos, analíticos e estatísticos. Todas as modificações de procedimentos devem ser explicadas.

Resultados e Discussão

Os resultados devem ser combinados com discussão. Dados suficientes, todos com algum índice de variação incluso, devem ser apresentados para permitir ao leitor a interpretação dos resultados do experimento. A discussão deve interpretar clara e concisamente os resultados e integrar resultados de literatura com os da pesquisa para proporcionar ao leitor uma base ampla na qual possa aceitar ou rejeitar as hipóteses testadas.

Evitar parágrafos soltos e citações pouco relacionadas ao assunto.

Conclusões

Devem ser redigidas em parágrafo único e conter no máximo 1.000 caracteres com espaço.

Não devem ser repetição de resultados. Devem ser dirigidas aos leitores que não são necessariamente profissionais ligados à ciência animal. Devem explicar claramente, sem

abreviações, acrônimos ou citações, o que os resultados da pesquisa concluem para a ciência animal.

Abreviaturas, símbolos e unidades

Abreviaturas, símbolos e unidades devem ser listados conforme indicado na *home page* da RBZ, link Revista>Instruções aos autores.

Deve-se evitar o uso de abreviações não consagradas e de acrônimos, como por exemplo: "o T3 foi maior que o T4, que não diferiu do T5 e do T6". Este tipo de redação é muito cômoda para o autor, mas é de difícil compreensão para o leitor.

Tabelas e Figuras

É imprescindível que todas as Tabelas sejam digitadas segundo menu do Word "Inserir Tabela", em células distintas (não serão aceitas tabelas com valores separados pelo recurso ENTER ou coladas como figura). Tabelas e figuras enviadas fora de normas serão devolvidas para adequação.

Devem ser numeradas seqüencialmente em algarismos arábicos e apresentadas logo após a chamada no texto.

O título das tabelas e figuras deve ser curto e informativo, devendo-se adotar as abreviaturas divulgadas oficialmente pela RBZ.

A legenda das figuras (chave das convenções adotadas) deve ser incluída no corpo da figura. Nos gráficos, as designações das variáveis dos eixos X e Y devem ter iniciais maiúsculas e unidades entre parênteses.

Figuras não-originais devem conter, após o título, a fonte de onde foram extraídas, que deve ser referenciada.

As unidades, a fonte (Times New Roman) e o corpo das letras em todas as figuras devem ser padronizados.

Os pontos das curvas devem ser representados por marcadores contrastantes, como círculo, quadrado, triângulo ou losango (cheios ou vazios).

As curvas devem ser identificadas na própria figura, evitando o excesso de informações que comprometa o entendimento do gráfico.

As figuras devem ser gravadas no programa Word, Excel ou Corel Draw (extensão CDR), para possibilitar a edição e possíveis correções.

Usar linhas com, no mínimo, 3/4 ponto de espessura.

No caso de gráfico de barras, usar diferentes efeitos de preenchimento (linhas horizontais, verticais, diagonais, pontinhos etc). Evite os padrões de cinza porque eles dificultam a visualização quando impressos.

As figuras deverão ser exclusivamente monocromáticas.

Não usar negrito nas figuras.

Os números decimais apresentados no interior das tabelas e figuras devem conter vírgula, e não ponto.

Citações no texto

As citações de autores no texto são em letras minúsculas, seguidas do ano de publicação. Quando houver dois autores, usar & (e comercial) e, no caso de três ou mais autores, citar apenas o sobrenome do primeiro, seguido de et al.

Comunicação pessoal (ABNT-NBR 10520).

Não fazem parte da lista de referências, sendo colocadas apenas em nota de rodapé. Coloca-se o sobrenome do autor seguido da expressão "comunicação pessoal", a data da comunicação, o nome, estado e país da instituição à qual o autor é vinculado.

Literatura Citada

Baseia-se na Associação Brasileira de Normas Técnicas _ ABNT (NBR 6023).

Devem ser redigidas em página separada e ordenadas alfabeticamente pelo(s) sobrenome(s) do(s) autor(es).

Digitá-las em espaço simples, alinhamento justificado e recuo até a terceira letra a partir da segunda linha da referência. Para formatá-las, siga as seguintes instruções: no menu Formatar, escolha a opção Parágrafo... RECUo especial, opção DESLOCAMENTO... 0,6 cm.

Em obras com dois e três autores, mencionam-se os autores separados por ponto-e-vírgula e, naquelas com mais de três autores, os três primeiros vêm seguidos de et al. As iniciais dos autores não podem conter espaços. O termo et al. não deve ser italizado nem precedido de vírgula.

O recurso tipográfico utilizado para destacar o elemento título será negrito e, para os nomes científicos, itálico.

Indica(m)-se o(s) autor(es) com entrada pelo último sobrenome seguido do(s) prenome(s) abreviado (s), exceto para nomes de origem espanhola, em que entram os dois últimos sobrenomes.

No caso de homônimos de cidades, acrescenta-se o nome do estado (ex.: Viçosa, MG; Viçosa, AL; Viçosa, RJ).

Obras de responsabilidade de uma entidade coletiva

ASSOCIATION OF OFFICIAL ANALYTICAL CHEMISTRY - AOAC. **Official methods of analysis**. 16.ed. Arlington: AOAC International, 1995. 1025p.

UNIVERSIDADE FEDERAL DE VIÇOSA - UFV. **Sistema de análises estatísticas e genéticas - SAEG**. Versão 8.0. Viçosa, MG, 2000. 142p.

Livros e capítulos de livro

LINDHAL, I.L. Nutrición y alimentación de las cabras. In: CHURCH, D.C. (Ed.) **Fisiologia digestiva y nutrición de los ruminantes**. 3.ed. Zaragoza: Acríbia, 1974. p.425-434.

NEWMANN, A.L.; SNAPP, R.R. **Beef cattle**. 7.ed. New York: John Wiley, 1997. 883p.

Teses e dissertações

Castro, F.B. **Avaliação do processo de digestão do bagaço de cana-de-açúcar auto-hidrolisado em bovinos**. Piracicaba: Escola Superior de Agricultura Luiz de Queiroz, 1989. 123p. Dissertação (Mestrado em Zootecnia) - Escola Superior de Agricultura Luiz de Queiroz, 1989.

Boletins e relatórios

BOWMAN,V.A. **Palatability of animal, vegetable and blended fats by equine**. (S.L.): Virgínia Polytechnic Institute and State University, 1979. p.133-141 (Research division report, 175).

Artigos

Restle, j.; Vaz, r.z.; Alves Filho, d.c. et al. Desempenho de vacas Charolês e Nelore desterneiradas aos três ou sete meses. **Revista Brasileira de Zootecnia**, v.30, n.2, p.499-507, 2001.

Congressos, reuniões, seminários etc

Citar o mínimo de trabalhos publicados em forma de resumo, procurando sempre referenciar os artigos publicados na íntegra em periódicos indexados.

CASACCIA, J.L.; PIRES, C.C.; RESTLE, J. Confinamento de bovinos inteiros ou castrados de diferentes grupos genéticos. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 30., 1993, Rio de Janeiro. **Anais...** Rio de Janeiro: Sociedade Brasileira de Zootecnia, 1993. p.468.

EUCLIDES, V.P.B.; MACEDO, M.C.M.; OLIVEIRA, M.P. Avaliação de cultivares de *Panicum maximum* em pastejo. In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 36., 1999, Porto Alegre. **Anais...** São Paulo: Sociedade Brasileira de Zootecnia/Gmosis, [1999] (CD-ROM).

Artigo e/ou matéria em meios eletrônicos

NGUYEN, T.H.N.; NGUYEN, V.H.; NGUYEN, T.N. et al. [2003]. Effect of drenching with cooking oil on performance of local yellow cattle fed rice straw and cassava foliage. **Livestock Research for Rural Development**, v.15, n.7, 2003. Disponível em: <http:// www.cipav.org.co/lrrd/lrrd15/7/nhan157.htm> Acesso em: 28/07/2005.

REBOLLAR, P.G.; BLAS, C. [2002]. **Digestión de la soja integral en rumiantes**. Disponível em: <http://www. ussoymeal.org/ruminant_s.pdf.> Acesso em: 12/10/02.

SILVA, R.N.; OLIVEIRA, R. [1996]. Os limites pedagógicos do paradigma da qualidade total na educação. In: CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UFPe, 4., 1996, Recife. **Anais eletrônicos...** Recife: Universidade Federal do Pernanbuco, 1996. Disponível em: <http://www.propesq.ufpe.br/anais/anais.htm> Acesso em: 21/01/97.