ANA CAROLINA MOTA CAMPANA

VERIFICAÇÃO DOS EFEITOS DAS VARIÂNCIAS E DAS RELAÇÕES DE VARIÁVEIS LIGADAS À PECUÁRIA DE LEITE NO AGRUPAMENTO DOS PRODUTORES

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

VIÇOSA
MINAS GERAIS – BRASIL
2009

ANA CAROLINA MOTA CAMPANA

VERIFICAÇÃO DOS EFEITOS DAS VARIÂNCIAS E DAS RELAÇÕES DE VARIÁVEIS LIGADAS À PECUÁRIA DE LEITE NO AGRUPAMENTO DOS PRODUTORES

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

APROVADA: 16 de fevereiro de 2009.

_____	_____
Prof. Carlos Henrique Osório Silva	Prof. Luiz Alexandre Peternelli
(Co-orientador)	(Co-orientador)
_____	_____
Prof. Nerilson Terra Santos	Prof. José Maurício de Souza Campos

_____

Prof. José Ivo Ribeiro Júnior

(Orientador)

A Deus e às pessoas que mais AMO:
meus pais, Sonia e José Américo,
meus avós, Maria e Américo,
meu noivo, Moysés,
meu irmão, Júnior
e minha tia Tereza.
OFEREÇO.

AGRADECIMENTOS

A Deus pela grandiosidade da vida, por nunca me deixar desistir e por sempre me dar motivos para agradecer.

Ao meu noivo, Moysés, pela paciência, compreensão, companheirismo e por muito ter me incentivado a cumprir mais esta etapa de minha vida. Obrigada principalmente por me fazer muito mais feliz!

Aos meus pais José Américo e Sonia, pelo apoio de sempre, por terem me oferecido uma boa educação e por tudo que eles fizeram para que eu chegasse até aqui.

A minha tia Tereza por me receber sempre tão carinhosamente, muito obrigada pelo seu amor e carinho. Você é uma segunda mãe pra mim.

À Universidade Federal de Viçosa, particularmente ao Departamento de Informática, Setor de Estatística, pela formação acadêmica e oportunidade de crescimento intelectual e profissional.

À CAPES pelo apoio financeiro sem o qual esta dissertação não teria sido possível, ou a teria tornado mais difícil.

Ao professor José Ivo Ribeiro Júniro, pela orientação séria e comprometida, pela amizade e confiança depositada em mim.

Aos professores que participaram da banca examinadora, Carlos Henrique Osório Silva, Luiz Alexandre Peternelli, Nerilson Terra Santos e José Maurício de Souza Campos, por terem aceitado o convite e por suas contribuições oportunas, que certamente enriqueceram o trabalho.

Ao secretário da Pós-Graduação Altino, por toda amizade, incentivo e por estar sempre sorrindo e disposto a ajudar.

Aos meus colegas e amigos da UFV, principalmente à Dani, pela grande amizade construída, os desabafos, as risadas... Enfim, por todos os momentos de descontração, companheirismo e amizade.

À minha amiga Talita, por entender minha ausência e por estar sempre presente na minha vida.

À Dona Maria, minha "avó emprestada", por me acolher tão bem em sua casa, por cuidar de mim e por todo o carinho.

Por fim, a todas as pessoas que de uma forma ou de outra têm contribuído para que eu me torne uma pessoa melhor.

BIOGRAFIA

ANA CAROLINA MOTA CAMPANA, filha de José Américo Campana e Sonia Maria Mota Campana, nasceu em Vila Velha, Espírito Santo, no dia 2 de agosto de 1983.

Em março de 2002, iniciou o curso de Estatística na Universidade Federal do Espírito Santo, concluindo-o em dezembro de 2006.

Em março de 2007, ingressou no Programa de Pós-Graduação, em nível de mestrado, em Estatística Aplicada e Biometria, da Universidade Federal de Viçosa, submetendo-se à defesa da dissertação no dia 16 de fevereiro de 2009.

SUMÁRIO

Página

RESUMO

CAMPANA, Ana Carolina Mota, M.Sc., Universidade Federal de Viçosa, fevereiro de 2009. **Verificação dos efeitos das variâncias e das relações de variáveis ligadas à pecuária de leite no agrupamento dos produtores.** Orientador: José Ivo Ribeiro Júnior. Co-Orientadores: Carlos Henrique Osório Silva e Luiz Alexandre Peternelli.

Com o aumento substancial na quantidade de dados armazenados, surge a necessidade da utilização de métodos que permitam analisar simultaneamente várias variáveis medidas em cada elemento amostral, e ainda com a possibilidade de reduzir a dimensionalidade desse conjunto sem perda significativa de informação. Entre eles, pode-se citar o método dos componentes principais, cuja obtenção pode envolver a matriz de covariâncias (S) ou a de correlações (R) das variáveis de interesse. Como a utilização dessas matrizes pode fornecer diferentes componentes, objetivou-se investigar, por meio da simulação de dados, os efeitos das escalas das características sobre a qualidade e a viabilidade da classificação dos elementos amostrais, buscando assim, indicar estratégias de análise mais adequadas em diferentes casos. Além do estudo de simulação, foi realizado outro com variáveis zootécnicas e econômicas referentes a 255 produtores de leite de três regiões do estado de Minas Gerais, com o objetivo de verificar qual a melhor estrutura de dados em classificar de forma mais apropriada os produtores mais viáveis economicamente. Em ambos os estudos, foi efetuada uma transformação nos valores das variáveis baseada nos respectivos coeficientes de variação, cuja matriz de covariâncias foi denominada de S^*. Observou-se que a utilização da matriz S privilegiou as variáveis econômicas de maiores variâncias, enquanto a matriz R considerou as variáveis mais correlacionadas entre si como as mais

importantes. A obtenção dos CPs com base na matriz S^* minimizou os problemas das escalas inerentes aos usos das matrizes *S* e *R*. A primeira, por considerá-la totalmente e, a segunda, por desconsiderá-la. Desta forma, considerou-se a matriz S^* como a mais indicada no presente estudo de caso, uma vez que priorizou como mais importantes, as variáveis econômicas mais relacionadas às variáveis zootécnicas.

ABSTRACT

CAMPANA, Ana Carolina Mota, M.Sc., Universidade Federal de Viçosa, February, 2009. **Verification of the effects of variances and of the relationships among variables related to milk production in the grouping of dairy farmers.** Adviser: José Ivo Ribeiro Júnior. Co-Advisers: Carlos Henrique Osório Silva and Luiz Alexandre Peternelli.

Nowadays research often collect information on many variables from a great number of experimental units, hence produce and store large amount of data, which in turn requires methods that can handle such situations. Statistical methods such as the principal component analysis (PCA), that can reduce the dimensionality of the analysis without significant information loss, are of great interest. PCA can use either the covariance (S) or the correlation (R) matrix among variables, but the analysis may result in different Principal Components (PC) resulting from R or S. In order to indicate the best strategies for different scenarios, we conducted a simulation study to investigate the effects of variable scaling over the viability and quality of the results from PCA analysis used to cluster experimental units. In addition to this first simulation study, we also conducted a second one using animal science and economical variables from 255 dairy producers from three locations of Minas Gerais State. The goal was to verify the most appropriate data structure for cluster analysis, such that it best classifies the most economically viable producers. In both studies we used a transformation of variables based on its coefficient of variation, which resulted in a new covariance matrix named S*. Results showed that the use of matrix S favored economical variables with larger variances, while use of R matrix resulted as the most important variables the ones with larger

correlations among them. Calculations of PC using matrix S* minimized these scaling problems when S and R matrices are used. Analysis using S is entirely affected by the variable scale while using R is not affected by the scale at all. We concluded that the S* matrix was the most appropriate for the present case study because it considered the most important economical variables to be the ones most related to the animal science variables.

1. INTRODUÇÃO GERAL

Durante os últimos anos tem-se verificado um crescimento substancial na produção e no armazenamento de dados que, em grande escala, são inviáveis de serem analisados através de métodos manuais e tradicionais (SHAPIRO, 1991). Sabe-se que essas grandes quantidades de dados equivalem a um maior potencial de informação e isso faz com que, em muitos estudos práticos, os pesquisadores fiquem “tentados” a incluir toda a informação disponível. Porém esta opção conduz em muitas situações, não a um melhor resultado, mas sim à introdução de confusão ou ruído, isso porque muitas das informações contidas neles não estão caracterizadas explícita e corretamente.

Diante deste cenário, surge a necessidade de explorá-los para extrair informações e conhecimento utilizados nas soluções dos diversos problemas. Para tanto, diferentes metodologias podem ser utilizadas. Na estatística, os métodos de análise multivariada são aplicados quando várias variáveis são medidas simultaneamente em cada elemento amostral. A necessidade de analisá-las simultaneamente implica em procurar métodos que permitam reduzir a dimensionalidade sem perda significativa da informação contida nos dados.

Entre as técnicas multivariadas utilizadas para reduzir a dimensionalidade dos dados, 0destaca-se a dos componentes principais (CP), introduzida por Karl Pearson em 1901 e fundamentada no artigo de Hotelling de 1933. Seu principal objetivo é de explicar a estrutura de variância e covariância de um vetor composto por p variáveis (MINGOTI, 2007).

A obtenção dos CPs envolve a decomposição da matriz de covariâncias (S). Porém, em geral os pesquisadores optam por trabalhar com as variáveis em escala

padronizada, isto é, transformadas com médias e variâncias iguais a zero e um, respectivamente. Neste caso, a matriz *S* das variáveis padronizadas equivale à matriz de correlações (*R*) das variáveis originais de interesse. Entretanto, o uso ou não das transformações nos dados pode fornecer diferentes componentes, o que pode alterar a classificação final dos elementos amostrais, uma vez que eles não são invariantes à mudança de escala (JOHNSON e WICHERN, 2002; REIS, 1997). Este problema foi também verificado nos estudos de Dawkins (1989) e Naik e Kattree (1996), que utilizaram diferentes estratégias para a obtenção dos CPs, a fim de classificar 56 países quanto aos recordes em competições olímpicas. Ao final, eles obtiveram diferentes classificações.

Assim, essa pesquisa tem por objetivo investigar, por meio da simulação de dados, os efeitos das correlações e das variâncias das variáveis sobre a qualidade e a viabilidade da classificação dos elementos amostrais. A partir disso, então, indicar uma estratégia de análise mais adequada e eficiente nos casos em que há uma discrepância na variância e na escala das variáveis estudadas.

Além do estudo baseado em simulação de dados, foi realizado outro com variáveis zootécnicas e econômicas do banco de dados da Central de Processamento de Dados do Educampo/Sebrae-MG (CPDE), referentes a 255 produtores de leite de três regiões do estado de Minas Gerais, objetivando-se verificar qual estrutura de dados classifica de forma mais apropriada os produtores do ponto de vista econômico.

Este trabalho foi organizado da seguinte forma:

- Em sequência a esta introdução, apresenta-se uma revisão de literatura onde é feita uma explanação sobre simulação de dados, componentes principais e análise de agrupamento. Além disso, apresentam-se alguns trabalhos onde estas técnicas foram utilizadas, dando ênfase a trabalhos relacionados à pecuária leiteira.
- No Capítulo 1 foi feito o estudo das diferentes estruturas de dados sobre a estimação dos componentes principais, usando dados simulados.
- No Capítulo 2, foi feito um estudo de caso buscando-se obter a estratégia de análise mais adequada para discriminar produtores de leite de três regiões do estado de Minas Gerais.

2. REVISÃO DE LITERATURA

2.1. Simulação de Dados

Simulação computacional consiste em empregar técnicas matemáticas em computadores com o propósito de imitar um processo ou operação do mundo real. Desta forma, para ser realizada uma simulação, é necessário construir um modelo que corresponda à situação real que se deseja simular.

A capacidade de simular uma amostra de dados baseada em um modelo estatístico específico e de analisar os dados simulados pode facilitar a compreensão de um problema complexo, permitindo assim estudar uma ampla gama de situações controladas pelo pesquisador. O que consequentemente facilita a comparação e aperfeiçoamento de diversas metodologias (BAKER, 1995).

Assim, a simulação computacional tem sido de grande utilidade em diversas áreas do conhecimento como nos trabalhos relacionados à agricultura (PETERNELLI et al., 2006) e genética (GURGEL, 2007; CUNHA et al., 2006; SALGADO et al., 2008). Em geral, utiliza-se em áreas que por algum motivo seja custoso a obtenção de uma amostra.

2.2. Componentes Principais

A técnica denominada análise de componentes principais (ACP), foi introduzida por Karl Pearson em 1901 e fundamentada no artigo de Hotelling de 1933. Seu principal objetivo é de explicar a estrutura de variância e covariância de um vetor, composto de *p* variáveis aleatórias, por meio da obtenção dos componentes principais (MINGOTI, 2007).

Cada componente principal (CP) é uma combinação linear das variáveis originais, independentes entre si e estimados com o propósito de reter, em ordem de estimação, o máximo de informação em termos da variação total contida nos dados. Se forem *p* variáveis originais, será possível obter *p* CPs. No entanto, deseja-se uma redução do número de variáveis estudadas em poucos CPs, ou seja, deseja-se substituir a informação contida nas *p* variáveis originais pela informação contida em *k* ($k \leq 3$) CPs não correlacionados e com a menor perda de informação possível (BARBOSA et al., 2005).

Dada uma matriz *X* representando *n* elementos amostrais em *p* variáveis ($X_1, X_2, ..., X_p$), o propósito principal é determinar as novas variáveis ($CP_1, CP_2, ..., CP_p$), tal que CP_j ($j = 1, 2, ..., p$) seja dada pela combinação linear das *p* variáveis *Xs*.

Os CPs podem ser baseados nas variáveis originais *Xs* ou nas variáveis padronizadas *Zs*, onde: $Z_j = \dfrac{X_j - \overline{X}_j}{s_j}$.

Assim, os *p* componentes $CP_1, CP_2, ..., CP_p$ baseados nas variáveis originais *Xs* são estimados por:

$$\hat{C}P_j = \hat{\mathbf{a}}^*_{1j} X_1 + \hat{\mathbf{a}}^*_{2j} X_2 + \cdots + \hat{\mathbf{a}}^*_{pj} X_p \text{, em que:}$$

$\hat{\mathbf{a}}^*_j$ = estimativa do autovetor normalizado do CP_j.

Já os *p* componentes $CP_1, CP_2, ..., CP_p$ baseados nas variáveis padronizadas *Z*s são estimados por:

$$\hat{C}P_j = \hat{\mathbf{a}}^*_{1j} Z_1 + \hat{\mathbf{a}}^*_{2j} Z_2 + \cdots + \hat{\mathbf{a}}^*_{pj} Z_p$$

Para *p* variáveis *X*s, serão estimados *p* autovalores ($\hat{\lambda}_j$) por meio do determinante da expressão baseada na matriz amostral ($p \times p$) das variâncias e covariâncias (*S*) das variáveis *X*s:

$S - \hat{\lambda}_j I = 0$, em que:

$$S = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{12} & s_2^2 & \cdots & s_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ s_{1p} & s_{2p} & \cdots & s_p^2 \end{bmatrix};$$

I = matriz identidade de ordem p;

s_j^2 = estimativa da variância da variável X_j;

$s_{jj'}$ = estimativa da covariância entre as variáveis X_j e $X_{j'}$.

Alternativamente, os p autovalores podem ser estimados por meio do determinante da expressão baseada na matriz amostral ($p \times p$) das correlações (R) entre as variáveis Xs, que é a mesma matriz ($p \times p$) das variâncias e covariâncias entre as variáveis padronizadas Zs:

$R - \hat{\lambda}_j I = 0$, em que:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix};$$

$r_{jj'}$ = estimativa da correlação entre as variáveis X_j e $X_{j'}$.

Apesar das estimativas dos autovalores serem diferentes quando baseadas em S e R, por estarem captando informações diferentes para ambos os casos, têm-se (JOHNSON e WICHERN, 2002; FERREIRA, 2008):

$\hat{V}(\hat{CP}_j) = \hat{\lambda}_j$ (autovalor de ordem j da matriz R ou S);

onde,

$\hat{V}(\hat{CP}_1) \geq \hat{V}(\hat{CP}_2) \geq \cdots \geq \hat{V}(\hat{CP}_p)$ e

$Cov(\hat{CP}_1, \hat{CP}_2) = Cov(\hat{CP}_1, \hat{CP}_p) = \ldots = Cov(\hat{CP}_{p-1}, \hat{CP}_p) = 0$.

Portanto, o CP_1 contém mais informações sobre os dados do que o CP_2, que não contém informações do CP_1, e assim sucessivamente.

Para cada autovalor estimado $\hat{\lambda}_j$ (j=1, 2, ..., p) corresponde um autovetor estimado, normalizado ou não. Assim, para o autovalor estimado $\hat{\lambda}_j$, estima-se o autovetor não normalizado $\hat{\mathbf{a}}_j$, a partir de uma das soluções dos sistemas de equações dados a seguir:

$[S - \hat{\lambda}_j I]\ \hat{\mathbf{a}}_j = \phi$;

$[R - \hat{\lambda}_j I]\ \hat{\mathbf{a}}_j = \phi$, em que:

$\hat{\mathbf{a}}_j$ = estimativa do autovetor não normalizado do CP_j;

ϕ = vetor nulo de dimensão $px1$.

A estimativa do autovetor normalizado $\hat{\mathbf{a}}_j^*$ é dada por:

$$\hat{\mathbf{a}}_j^* = \begin{bmatrix} \hat{a}_{1j}^* \\ \hat{a}_{2j}^* \\ \vdots \\ \hat{a}_{pj}^* \end{bmatrix} = \frac{1}{\sqrt{\hat{a}_{1j}^2 + \hat{a}_{2j}^2 + \cdots + \hat{a}_{pj}^2}} \begin{bmatrix} \hat{a}_{1j} \\ \hat{a}_{2j} \\ \vdots \\ \hat{a}_{pj} \end{bmatrix}$$, em que:

$\hat{a}^{*2}_{1j} + \hat{a}^{*2}_{2j} + \cdots + \hat{a}^{*2}_{pj} = 1$.

A importância relativa de um CP_j, é avaliada pela porcentagem da variância total que ele explica. Em geral, para interpretar os dados com sucesso, basta escolher os primeiros k componentes que envolvam pelo menos 70% da variância total. Isto é, basta escolher $CP_1, CP_2, ..., CP_k$, tal que:

$$\frac{\sum_{j=1}^{k} \hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j} \times 100\% \geq 70\%$$, em que $k < p$.

Cada CP escolhido deve ser intrepretado, uma vez que constituem as novas variáveis respostas que serão utilizadas nas análises subsequentes do estudo. A interpretação dos CPs é baseada nos valores dos coeficientes da combinação linear das variáveis em estudo (originais ou padronizadas), isto é, com base, nas variáveis que mais contriuem para o CP.

A importância ou influência que cada variável exerce sobre o CP, é dada pela correlação entre cada variável X_j e o componente CP_j, que está sendo interpretado:

$r_{X_j \hat{C}P_j} = \dfrac{\hat{a}_j^* \sqrt{\hat{\lambda}_j}}{s_j}$, para CP_j baseado em S;

$r_{X_j \hat{C}P_j} = \hat{a}_j^* \sqrt{\hat{\lambda}_j}$, para CP_j baseado em R.

A importância relativa dos CPs decresce do primeiro para o último. Desta forma, conclui-se que os últimos componentes serão responsáveis pela explicação de uma pequena parte da variação total dos dados. Assim, a variável X_j que apresentar maior correlação, baseada em S, em valor absoluto, com o componente de menor autovalor, terá menor importância em explicar a variabilidade em relação aos n elementos amostrais (MARDIA et al.,1997; JOHNSON e WICHERN, 2002).

Existem diversos critérios práticos para determinar quantos componentes devem ser utilizados para a análise. Dentre estes critérios, segundo Reis (2007), os mais utilizados são: o *scree-plot* (CATTELL, 1966), que é um gráfico dos autovalores estimados ($\hat{\lambda}_j$), em função da ordem dos CPs, representando graficamente a porcentagem de variância explicada por cada componente. Quando esta porcentagem se reduz e a curva passa a ser quase paralela ao eixo das abscissas, pode-se excluir os componentes correspondentes. Outro critério muito utilizado é incluir os k primeiros CPs suficientes para explicar pelo menos 70% da variação total (MELÉM JÚNIOR et al., 2008; CRUZ, 1990). Além desses, tem-se também o critério de Kaiser (1958), que inclui apenas os CPs, cujos autovalores são superiores ou iguais à média dos autovalores.

Os escores relativos a cada elemento amostral de cada componente são estimados com base nos valores dos coeficientes de ponderação associados aos valores padronizados dos p variáveis estudadas. A dispersão destes escores em eixos cartesianos é que vai indicar quais são os mais divergentes.

As estimativas dos escores relativos aos n elementos amostrais obtidos em relação aos p componentes CP_1, CP_2, ..., CP_p baseados na matriz S são dadas por:

$$\hat{C}P_{1i} = \hat{\mathbf{a}}_{11}^* x_{1i} + \hat{\mathbf{a}}_{21}^* x_{2i} + \cdots + \hat{\mathbf{a}}_{p1}^* x_{pi}$$
$$\hat{C}P_{2i} = \hat{\mathbf{a}}_{12}^* x_{1i} + \hat{\mathbf{a}}_{22}^* x_{2i} + \cdots + \hat{\mathbf{a}}_{p2}^* x_{pi}$$
$$\vdots$$
$$\hat{C}P_{ji} = \hat{\mathbf{a}}_{1j}^* x_{1i} + \hat{\mathbf{a}}_{2j}^* x_{2i} + \cdots + \hat{\mathbf{a}}_{pj}^* x_{pi}$$

onde

x_{ji} = valor da variável X_j ($j = 1, 2, ..., p$) do elemento amostral i (i = 1, 2, ..., n).

As estimativas dos escores relativos aos n elementos amostrais obtidas em relação aos p componentes CP_1, CP_2, ..., CP_p baseados na matriz R são dadas por:

$$\hat{C}P_{1i} = \hat{\mathbf{a}}^*_{11} z_{1i} + \hat{\mathbf{a}}^*_{21} z_{2i} + \cdots + \hat{\mathbf{a}}^*_{p1} z_{pi}$$
$$\hat{C}P_{2i} = \hat{\mathbf{a}}^*_{12} z_{1i} + \hat{\mathbf{a}}^*_{22} z_{2i} + \cdots + \hat{\mathbf{a}}^*_{p2} z_{pi}$$
$$\vdots$$
$$\hat{C}P_{ji} = \hat{\mathbf{a}}^*_{1j} z_{1i} + \hat{\mathbf{a}}^*_{2j} z_{2i} + \cdots + \hat{\mathbf{a}}^*_{pj} z_{pi}$$

onde

z_{ji} = valor padronizado da variável X_j (j = 1, 2, ..., p) do elemento amostral i (i = 1, 2, ..., n).

2.3. Análise de Agrupamento

A análise de agrupamento, também conhecida como análise de conglomerados, classificação ou *cluster*, tem como objetivo dividir os n elementos da amostra ou da população, em k $(k \leq n)$ grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às p variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação à estas mesmas características (MINGOTI, 2007).

O processo de agrupamento envolve basicamente duas etapas. A primeira relaciona-se com a estimação de uma medida de dissimilaridade ou de similaridade entre os elementos amostrais ou entre as variáveis medidas (CRUZ e REGAZZI, 2004). A medida de dissimilaridade entre dois elementos amostrais cresce à medida que a diferença entre eles aumenta e, a de similaridade, aumenta à medida que essa diferença diminui (BARROSO e ARTES, 2003). Dentre as principais medidas de dissimilaridade para variáveis quantitativas, podem-se citar as distâncias Euclidiana, Euclidiana Média e Mahalanobis (MAHALANOBIS, 1936).

Cole (1998) e Han e Kamber (2001) destacam que a mais utilizada é a distância Euclidiana, que entre dois elementos i e i' é definida por:

$$d(X_i, X_{i'}) = [(X_i - X_{i'})'(X_i - X_{i'})]^{1/2} = \left[\sum_{j=1}^{p}(X_{ij} - X_{i'j})^2\right]^{1/2}$$

ou seja, os elementos amostrais são comparados em cada variável pertencente ao vetor de observações. Entretanto, um dos problemas apresentados pela distância Euclidiana é o fato dela ser influenciada pela escala das medições e pelo número de variáveis estudadas e, também, de não levar em conta o grau de correlação entre as mesmas. Para contornar o problema da escala, tem sido recomendável a padronização dos dados e, para contornar a influência do número de variáveis, utiliza-se a distância Euclidiana Média. Já em estudos onde se dispõe de repetições e, consequentemente, da estimação das variâncias e covariâncias residuais entre as variáveis disponíveis, indica-se fazer o uso da distância de Mahalanobis (FERREIRA, 2008).

As metodologias para a formação dos grupos (conglomerados) são frequentemente classificadas em hierárquicas e não hierárquicas, sendo que as primeiras podem ser divididas em aglomerativas e divisivas (RENCHER, 2002).

As técnicas hierárquicas, na maioria das vezes, são utilizadas em análise exploratória de dados com o intuito de identificar possíveis agrupamentos e o valor provável do número de grupos (g). Já para o uso das técnicas não hierárquicas, existe a necessidade que o número de grupos já esteja pré-especificado pelo pesquisador.

Nos métodos hierárquicos aglomerativos inicia-se com n conglomerados, ou seja, cada elemento do conjunto de dados é considerado como sendo um conglomerado isolado. Em cada passo do algoritmo, os elementos vão sendo agrupados de acordo com uma medida de dissimilaridade ou de similaridade, formando novos conglomerados até o momento no qual todos os elementos formarão um único grupo. Dentre os métodos hierárquicos, podem-se citar: ligação média entre grupos (UPGMA), o método de Ward (Ward, 1963), ligação simples (método do vizinho mais próximo) e ligação completa (método do vizinho mais distante).

O método UPGMA trata a distância entre dois grupos como a média da distância entre todos os pares dos elementos pertencentes a cada grupo. Desta forma, se o grupo 1 (g_1) tem n_1 elementos e o grupo 2 (g_2) tem n_2, a distância entre eles será definida por (MINGOTI, 2007):

$$d(g_1,g_2) = \sum_{j\in g_1}\sum_{k\in g_2}\left(\frac{1}{n_1 n_2}\right)d(j,k),$$

em que:

d(j,k) = distância entre os elementos *j* e *k* pertencentes aos grupos g_1 e g_2 respectivamente.

Já o de método de Ward, também chamado de "variância mínima", basea-se na "mudança de variação" causada pela inclusão de um elemento num determinado grupo (HAIR et al., 2005). A alocação de um elemento a um grupo é feita maximizando a homogeneidade dentro dos grupos ou minimizando o total das somas de quadrados dentro de grupos, também conhecida como soma de quadrados dos erros (ESS), que é calculada por:

$$ESS = \sum_{i=1}^{n} (y_i - \bar{y})'(y_i - \bar{y}),$$

onde y_i é o vetor multivariado dos *k* escores associados ao *i*-ésimo elemento e $\bar{y}$ é a média de todos os escores.

No método de ligação simples o dendrograma é estabelecido pelos elementos com maior similaridade. Já o método da ligação completa trata-se de uma antítese ao método da ligação simples. No método da ligação completa a similaridade entre dois grupos é dada pelos elementos de cada grupo que menos se parecem. Este método, geralmente, leva a grupos compactos e discretos, tendo os seus valores de similaridade relativamente pequenos (CRUZ e CARNEIRO, 2003).

2.4. Aplicações da Análise de Componentes Principais

Scremin e Bastos (2000) fizeram uso da técnica dos CPs para reduzir a dimensionalidade dos dados com objetivo de identificar as variáveis mais representativas da variabilidade dos dados e um posterior agrupamento de propriedades rurais do Estado de Santa Catarina, utilizando variáveis contábeis. No estudo 27 variáveis originais foram reduzidas em seis CPs e estabelecidos, por uma rede neural artificial, quatro grupos com características bem definidas. A caracterização em grupos homogêneos auxiliou na definição das medidas que conduziram ao sucesso de empreendimentos agrícolas. Na seleção dos CPs foi utilizado o método de Kaiser, onde foi observado um acúmulo da porcentagem de variância explicada superior a 70%.

Aleixo et al. (2007) objetivaram captar a variedade de situações tecnológicas para identificar grupos de produtores, o mais semelhantes possível. No estudo foram considerados 72 produtores selecionados conforme 29 variáveis relacionadas a fatores produtivos. Avaliaram-se as variáveis de melhor representatividade dentro de cada fator e suas comunalidades dentro do conjunto de fatores analisados. Para a avaliação desses resultados, foi utilizada a análise fatorial baseada em CPs. Posteriormente, aplicou-se o método de análise de agrupamentos.

Em outro estudo, Pla (1986) objetivou conhecer a situação do setor leiteiro por meio de informações de uma série de variáveis que influenciam na produção total e na produtividade por propriedade e por vaca. Para a seleção dos CPs foram utilizados os métodos de Kaiser e do diagrama de autovalores, retendo para análise os três primeiros componentes com um acúmulo de 67% da variância total explicada.

Em Melià e Sesé (1999) encontra-se um estudo sobre propriedades psicométricas e estrutura fatorial de um questionário orientando a medida do clima organizacional para a segurança trabalhista. Utilizando o método de Kaiser foram retidos os três primeiros componentes, os quais explicaram 58,9% da variância total dos dados. Na interpretação dos fatores, foram considerados os componentes que possuíam cargas fatoriais iguais ou superiores a 0,4. Assim, foram identificados os fatores da empresa ligados a "*estrutura e segurança*", "*política de segurança*" e "*ações de intervenção em segurança*".

REFERÊNCIAS BIBLIOGRÁFICAS

ALEIXO, S.S.; SOUZA, J.G.; FERRAUDO, A.S. Técnicas de análise multivariada na determinação de grupos homogêneos de produtores de leite. **R. Bras. Zootec**. v. 36, p. 2168-2175, 2007.

BAKER, R.J. **Selection indices in plant breeding**. Boca Raton, Florida: CRC, 1995. 218 p.

BARBOSA, L.; LOPES, P.S.; REGAZZI, A.J.; GUIMARÃES, S.E.F.; TORRES, R.A. de. Seleção de variáveis de desempenho de suínos por meio da análise de componentes principais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 57, n.6, p. 805-810, 2005.

BARROSO, L.P.; ARTES, R. Análise Multivariada. In: 10º SEAGRO – Simpósio de Estatística Aplicada à Experimentação Agronômica, 48º RBRAS – Reunião Anual da Região Brasileira da Sociedade Internacional de Biometria, 2003, Lavras. **Minicurso**...Lavras: UFLA, 2003. 156p.

CATTELL, R.B. The screen test for number of factors. **Multivarite Behavioral Research**, v.1, p.140-161, 1966.

COLE, R.M. **Clustering with Genetic Algorithms**. 1998. M. Sc., Department of Computer Science, University of Western Australia, Australia.

CRUZ, C.D. **Aplicação de algumas técnicas multivariadas no melhoramento de plantas. Piracicaba. Genética e Melhoramento Vegetal**. 1990. 188 p. Tese de Doutorado em Agronomia. Escola Superior de Agricultura "Luiz de Queiroz" da Universidade de São Paulo. Piracicaba. São Paulo. 1990.

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

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** 3. ed., Viçosa: Universidade Federal de Viçosa, 2004. 480p.

DAWKINS, B. Multivariate analysis of national track records. **The American Statistician**, v.43, p.110-115, 1989.

CUNHA, E.E.; EUCLYDES, R.F., TORRES, A., CARNEIRO, P.L.S. Simulação de dados para avaliação genética de rebanhos de gado de corte. **Arq. Bras. Med. Vet. Zootec**., v.58,p.381-387, 2006.

FERREIRA, D.F. **Estatística Multivariada**. 1. ed. Lavras: Editora UFLA, 2008. 662p.

GURGEL, F.L. Simulação computacional no melhoramento genético de plantas. **Revista Científica do UNIDESC**, v. 1,p.1-17, 2007.

HAIR, J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W. **Análise multivariada de dados.** Porto Alegre, Bookman, 2005, 600p.

HAN, J.; KAMBER, M. **Cluster Analysis**. In: Morgan Kaufmann Publishers (eds.), *Data Mining: Concepts and Techniques*, 1 ed., chapter 8, New York, USA, Academic Press, 2001.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, p. 417-441, 1933.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 5. ed. New Jersey: Prentice Hall, 2002. 767p.

KAISER, H.F. The varimax criterion for analytic in factor analysis. **Psychometrika**, v. 23, p. 187-200, 1958.

MAHALANOBIS, P.C. On the generalized distance in statistics. **Proceedings of the National Institute of Sciences of India,** New Delhi, v.2, p.49-55, 1936.

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

MELÉM JÚNIOR, N.J.; FONSECA, I.C.B.; BRITO, O.R.; DECAËNS, T.; CARNEIRO, M.M.; MATOS, M.F.A.; GUEDES, M.C.; QUEIROZ, J.A.L.; BARROSO, K.O. Análise de componentes principais para avaliação de resultados analíticos da fertilidade de solos do Amapá. **Semina: Ciências Agrárias**, Londrina, v. 29, n.3, p. 499-506, jul./set. 2008.

MELIÀ, J.L.; SESÉ, A. **La medida del clima seguridad y salud laboral**. Anals de Psicologia, v.15, n.2, p.269-289. Sevicio de Publicaciones de la Universidad de Murcia (España, ISSN: 0212-9728), 1999.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte, Editora UFMG, 2007, 295p.

NAIK, D.N.; KATTREE, R. Revisiting Olympic track records: some practical considerations in the principal component analysis. **The American Statistician**, v.50, n.2, p.140-144, 1996.

PETERNELLI, L.A.; SILVA, G.F.; LEITE, H.G. Uma proposta para a geração de amostras aleatórias nos problemas de simulação em modelos de planejamento. **Revista Árvore**, v. 30, p. 749-758, 2006.

PLA, Laura E. **Analisis multivariado: metodo de componente principales**. Venezuela. Secretaria General de la Organización de los Estados Americanos: Whashington, D.C., 1986.

REIS, E. **Estatística Multivariada Aplicada.** Lisboa: Edições Silabo, 1997. 343p.

RENCHER, A.C. Methods of multivariate analysis. New York: John Wiley, 2002.

SALGADO, C.C.; NASCIMENTO, M.; CAMPANA, A.C.M.; CRUZ, C.D.; FERREIRA, A., BARREIRA, C.F. **Construção de mapas de ligação com dados incompletos de marcas moleculares.** In: 54º Congresso Brasileiro de Genética, Salvador, Bahia, 2008.

SCREMIN, M.A.A.; BASTOS, R.C. **Caracterização de grupos homogêneos de propriedades rurais do estado de Santa Catarina utilizando redes neurais artificiais.** In: Congresso e Mostra de Agroinformática, Ponta Grossa, Paraná, 2000.

SHAPIRO, P.G. Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. **AI Magazine**, v. 11, n. 5, Jan. 1991, Special issue, p.68-70.

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

CAPÍTULO 1

EFEITO DAS ESCALAS DAS VARIÁVEIS SOBRE AS ESTIMATIVAS DOS COMPONENTES PRINCIPAIS

RESUMO

Na era do conhecimento, a informação tornou-se uma das ferramentas mais importantes para a humanidade. Em todos os segmentos, uma gestão construída sobre uma base de dados bem formatada indica boa parte do caminho para o sucesso do negócio. A utilização de dados para monitorar o nível tecnológico e aumentar a capacidade gerencial torna-se cada vez mais relevante para alcançar os resultados pretendidos. Com isso, o que se verifica é um considerável aumento na quantidade de dados armazenados, o que faz surgir a necessidade de utilizar métodos que permitam analisar simultaneamente essas variáveis medidas em diferentes elementos amostrais, por meio da redução da massa de dados e sem perda significativa de informação. Uma alternativa é analisá-los pelo método estatístico dos componentes principais (CPs), cujas estimativas podem envolver as matrizes de covariâncias (S) ou de correlações (R) das variáveis de interesse. Como a utilização dessas matrizes pode fornecer diferentes componentes, objetivou-se investigar por meio da simulação de dados, o efeito das escalas de diferentes variáveis, correlacionadas ou não, sobre a composição das matrizes S e R e, consequentemente, sobre a qualidade e viabilidade da classificação dos elementos amostrais, por meio dos CPs mais apropriados. Desse modo, indicar estratégias de análises mais adequadas aos diferentes objetivos propostos. A matriz R é baseada nos

valores padronizados, com média igual a zero e variância igual a um, para todas as variáveis estudadas, independentemente da posição e da variação dos valores originais. Essa transformação retira o efeito das escalas das variáveis, mas cria o inconveniente de impor a mesma variação para todas elas. Por outro lado, a matriz *S* incorpora todo o efeito das escalas, o que poderá mascarar a variabilidade relativa captada pelos primeiros CPs. Desse modo, a matriz *S* estimada a partir das variáveis transformadas com médias iguais a zero e variâncias ponderadas pelos seus respectivos coeficientes de variação contornou as limitações impostas pelas análises anteriores.

1. INTRODUÇÃO

Quando se analisam dados oriundos de muitas variáveis medidas em diferentes elementos amostrais, com o auxílio de métodos estatísticos multivariados, a possibilidade de reduzir a dimensão do conjunto de dados sem grandes perdas de informações, desempenha papel crucial (CADIMA, 2001). Aproximações para espaços bi ou tri-dimensionais, além de tornarem possíveis visualizações gráficas aproximadas das relações entre as várias variáveis estudadas, possibilitam interpretações técnicas mais simplificadas e globalizadas.

Na análise de componentes principais (ACP), a redução da dimensão se baseia na substituição das p variáveis originais, que podem apresentar relações entre si, em k ($k<p$) novas variáveis não correlacionadas entre si (CPs), que são combinações lineares das variáveis originais estimadas com o propósito de reter o máximo de informação contida nas mesmas, em termos de variação total (CRUZ et al., 2004; BARBOSA et al., 2005; STEARNS et al., 2005).

As estimativas dos CPs se baseiam na decomposição da matriz de covariâncias amostral (S) das variáveis de interesse e, caso seja feita alguma transformação nos dados, na matriz de covariâncias associada às variáveis transformadas (MINGOTI, 2007). Uma transformação frequentemente utilizada tem sido a padronização pelas respectivas médias e desvios-padrão, resultando em variáveis com médias e variâncias iguais a zero e um, respectivamente. Neste caso, estimar os CPs pela decomposição da matriz de covariâncias (S) das variáveis padronizadas equivale a extrair os CPs da matriz de correlações (R) das variáveis originais. Uma vez estimados os CPs, os seus valores numéricos, denominados de escores, podem ser calculados para cada elemento amostral, o que permite a realização de uma análise estatística univariada para cada CP.

Porém, os componentes estimados a partir da matriz *S* são diferentes daqueles oriundos da matriz *R* (JOHNSON e WICHERN, 2002). Além disso, não há orientações claras sobre a escolha apropriada das matrizes a serem utilizadas nos diversos casos. Esta diferença foi ilustrada no trabalho de Naik e Kathree (1996), que questionaram a classificação final encontrada por Dawkins (1989). Pelo fato de existirem discrepâncias entre as variâncias das variáveis observadas, Dawkins (1989) utilizou a matriz de correlações (*R*) e ordenou os indivíduos em relação ao desempenho global. Por outro lado, Naik e Kathree (1996) usaram uma transformação nos dados originais para a estabilização da variância, diferente daquela realizada por Dawkins (1989) e encontraram uma nova classificação. Esse questionamento mostrou que a análise por CPs pode levar a diferentes resultados e, consequentemente, alguns deles podem estar associados a interpretações e conclusões erradas sobre o estudo realizado.

Devido ao fato das matrizes *S* e *R* influenciarem diferentemente as estimativas dos CPs, este trabalho teve como objetivo investigar por meio de um estudo de simulação de dados, os efeitos das variâncias, das correlações e da escala de medida das variáveis sobre a qualidade e a viabilidade da classificação dos elementos amostrais baseada nos escores dos CPs.

2. MATERIAL E MÉTODOS

Foram simulados, utilizando o software *R* versão 2.7.1 (R Development Core Team, 2007), diferentes conjuntos de dados compostos por duas variáveis (X_1 e X_2) com 200 observações sob diferentes situações de relações entre as mesmas, a partir da distribuição normal com os parâmetros descritos na Tabela 1, para os casos 1 (C1), 2 (C2), 3 (C3), 4 (C4) e 5 (C5).

Tabela 1. Parâmetros de média, variância e coeficiente de variação de duas variáveis (X_1 e X_2) simuladas sob cinco diferentes casos

Caso	Parâmetro	X_1	X_2
C1	**Média**	100	100
	Variância	100	100
	CV (%)	10	10
C2	**Média**	300	100
	Variância	900	100
	CV (%)	10	10
C3	**Média**	10	100
	Variância	36	100
	CV (%)	60	10
C4	**Média**	100	100
	Variância	36	100
	CV (%)	6	10
C5	**Média**	10	100
	Variância	100	100
	CV (%)	100	10

Em cada caso, foram simulados valores para as variáveis X_1 e X_2 considerando três diferentes situações:

a) X_1 e X_2 não são correlacionadas ($\rho_{12} = 0$);

b) X_1 e X_2 são moderadamente correlacionadas ($\rho_{12} = 0{,}5$);

c) X_1 e X_2 são totalmente correlacionadas ($\rho_{12} = 1$).

Formou-se assim um universo de 15 conjuntos de dados para análise dos CPs, utilizando-se os dados originais (matriz *S*), padronizados com média e variância iguais a zero e um, respectivamente (matriz *R*) e transformados levando em consideração o coeficiente de variação (CV) dos mesmos. A transformação utilizada foi definida como segue:

$z_{ij}^{*} = z_{ij} \times CV_j$, para $i = 1, 2, \ldots, 200$ e $j = 1, 2$.

em que:

z_{ij} = valor da *i*-ésima observação da variável X_j padronizada, com média zero e variância um;

z_{ij}^{*} = valor da *i*-ésima observação da variável X_j transformada, com média zero e variância ponderada pelo respectivo CV_j ;

CV_j = coeficiente de variação da variável X_j.

Esta transformação foi utilizada devido ao fato do CV ser uma medida mais apropriada da variabilidade relativa dos dados quando os mesmos estão sob diferentes escalas de medida. Desta forma, as variáveis passam a estar numa mesma escala de medida, com médias iguais a zero, porém com variâncias diferentes de um e ponderadas pelos seus respectivos CVs.

No caso C1, as variáveis X_1 e X_2 foram simuladas com a mesma média e desvio-padrão e consequentemente, com o mesmo CV. Este caso serviu de testemunha para o efeito da escala e possibilitou estudar, de forma pura, o efeito da correlação sobre as estimativas dos CPs. No caso C2, buscou-se verificar o efeito da escala causadora de média alta associada à variância alta em X_1, mas com a mesma variação relativa em X_2. Em C3, a variável X_1 de menor média e variância foi aquela que teve maior variação relativa. Já nos casos C4 e C5, a relação da média com a variância e o CV apresentou-se inversamente proporcional (Tabela 1).

Dada uma matriz ***X*** representando 200 elementos amostrais medidos por duas variáveis (X_1 e X_2), o propósito principal foi estimar as novas variáveis (CP_1, CP_2), tal que CP_j (j = 1, 2) é a combinação linear das duas variáveis *Xs*.

Os dois CPs estimados a partir da matriz S das variáveis *Xs* originais foram obtidos por:

$$\hat{C}P_1 = \hat{\mathbf{a}}_{11}^{*} X_1 + \hat{\mathbf{a}}_{12}^{*} X_2$$

$$\hat{C}P_2 = \hat{\mathbf{a}}_{21}^{*} X_1 + \hat{\mathbf{a}}_{22}^{*} X_2$$

em que:

$\hat{a}_j^{*} = [\hat{a}_{1j}^{*} \quad \hat{a}_{2j}^{*}]$ = estimativa do autovetor normalizado do CP_j, para j =1, 2.

Como as variáveis estudadas apresentaram diferentes escalas de medida foi realizada a padronização para que ficassem com médias e variâncias iguais. A padronização utilizada foi: $Z_{ij} = \dfrac{X_{ij} - \overline{X}_j}{s_j}$, i = 1,2,...,200 e j = 1,2.

Onde Z_j é a variável X_j padronizada, com média zero e variância um; $\overline{X}_j$ e s_j são a média e o desvio-padrão amostral da j-ésima variável, respectivamente.

Neste caso, a obtenção dos CPs a partir da matriz *S* dos dados padronizados é equivalente a obtê-los utilizando a matriz de correlações (*R*) dos dados originais.

Assim, os dois componentes (CP_1 e CP_2) baseados nas variáveis padronizadas *Zs* foram estimados por:

$$\hat{C}P_1 = \hat{a}_{11}^{*} Z_1 + \hat{a}_{12}^{*} Z_2$$

$$\hat{C}P_2 = \hat{a}_{21}^{*} Z_1 + \hat{a}_{22}^{*} Z_2$$

Portanto, para os dois CPs, foram estimados dois autovalores ($\hat{\lambda}_j$) por meio do determinante da expressão baseada na matriz de variâncias e covariâncias amostral (*S*) das variáveis originais *Xs*:

$S - \hat{\lambda}_j I = 0$, em que:

$$S = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix};$$

I = matriz identidade de ordem 2;

s_1^2 = variância amostral da variável X_1;

s_2^2 = variância amostral da variável X_2.

s_{12} = covariância amostral entre as variáveis X_1 e X_2.

Por outro lado, os dois autovalores também foram estimados por meio do determinante da expressão baseada na matriz de correlações amostrais das variáveis originais *Xs* (*R*), que é a mesma matriz de variâncias e covariâncias das variáveis padronizadas *Zs*:

$R - \hat{\lambda}_j I = 0$, em que:

$$R = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix};$$

r_{12} = coeficiente de correlação amostral entre as variáveis X_1 e X_2.

As estimativas dos autovalores baseadas em *S* e *R* são diferentes entre si, porém para ambos os casos têm-se:

$\hat{V}(\hat{CP}_j) = \hat{\lambda}_j$ (autovalor de ordem *j* da matriz *R* ou *S*);

em que,

$\hat{V}(\hat{CP}_1) \geq \hat{V}(\hat{CP}_2)$ e $Cov(\hat{CP}_1, \hat{CP}_2) = 0$.

Portanto, o CP_1 contém mais informações sobre os dados do que o CP_2, que não contém informações do CP_1.

Para cada autovalor estimado $\hat{\lambda}_j$ (j = 1, 2) corresponde um autovetor estimado normalizado ou não. A estimativa do autovetor não normalizado $\hat{a}_j$, a partir de uma das soluções dos sistemas de equações é dada a seguir:

$$[S - \hat{\lambda}_j I]\hat{a}_j = \phi;$$

$$[R - \hat{\lambda}_j I]\hat{a}_j = \phi,$$

em que:

$\hat{a}_j = [\hat{a}_{1j} \quad \hat{a}_{2j}]$ = estimativa do autovetor $\hat{a}_j$ não normalizado do CP_j;

ϕ = vetor nulo de dimensão 2x1.

A estimativa do autovetor normalizado $\hat{a}_j^*$ foi dada por:

$$\hat{a}_j^* = \begin{bmatrix} \hat{a}_{1j}^* \\ \hat{a}_{2j}^* \end{bmatrix} = \frac{1}{\sqrt{\hat{a}_{1j}^2 + \hat{a}_{2j}^2}} \begin{bmatrix} \hat{a}_{1j} \\ \hat{a}_{2j} \end{bmatrix}, \text{ em que } \hat{a}_{1j}^{*2} + \hat{a}_{2j}^{*2} = 1.$$

A importância relativa do CP_j foi dada pelo percentual da variância total que ele explica. Em geral, para interpretar os dados com sucesso, basta escolher os primeiros k componentes que envolvam pelo menos 70% da variância total (FERREIRA, 2008). Isto é, basta escolher os CP_j's, tal que:

$\frac{\sum_{j=1}^{k}\hat{\lambda}_j}{\sum_{j=1}^{p}\hat{\lambda}_j}\times 100 \geq 70\%$, onde p é o número de variáveis estudadas e $k \leq p$.

A importância ou influência que cada variável exerceu sobre o CP_j, foi dada pela correlação entre cada variável X_j e o componente CP_j, que está sendo interpretado, como segue:

$r_{X_j \hat{C}P_j} = \frac{\hat{a}_j^* \sqrt{\hat{\lambda}_j}}{s_j}$, para CP_j baseado em S;

$r_{X_j \hat{C}P_j} = \hat{a}_j^* \sqrt{\hat{\lambda}_j}$, para CP_j baseado em R.

Como a importância relativa dos CPs decresce do primeiro para o último, a variável X_j que apresentar maior correlação, em valor absoluto, com o segundo componente, ou seja, o de menor autovalor, terá menor importância em explicar a variabilidade em relação aos 200 elementos amostrais (MARDIA et al., 1997; JOLLIFFE, 1972;1973).

As estimativas dos escores relativos às 200 observações obtidas em relação aos dois componentes baseados nas matrizes S e R são dadas por:

$\hat{C}P_{ji} = \hat{\mathbf{a}}_{j1}^* X_{1i} + \hat{\mathbf{a}}_{j2}^* X_{2i}$ e $\hat{C}P_{ji} = \hat{\mathbf{a}}_{j1}^* Z_{1i} + \hat{\mathbf{a}}_{j2}^* Z_{2i}$, para $i = 1,\ldots,200$.

$\hat{a}_j^* =$ estimativa do autovetor normalizado do CP_j.

Os resultados obtidos pelas duas abordagens estudadas (matrizes S e R), além da transformação proposta, foram comparados, com o objetivo de definir as melhores estratégias a serem utilizadas nos diferentes casos apresentados. Além disso, foi calculado, para cada caso estudado, um coeficiente de coincidência entre os resultados fornecidos pelas matrizes R, S e S^*.

Para o cálculo do coeficiente de coincidência, considerou-se os 50 primeiros elementos com maiores escores no CP_1, e em seguida foi calculado o percentual de classificações concordantes entre as três matrizes estudadas.

Todas as análises estatísticas foram realizadas no software Minitab® 14.

3. RESULTADOS E DISCUSSÃO

Na Tabela 2 são apresentadas algumas estatísticas descritivas das variáveis simuladas (X_1 e X_2), cujas estimativas estão bastante próximas dos parâmetros especificados (Tabela 1).

Tabela 2. Estimativas das médias, variâncias, coeficientes de correlação de duas variáveis (X_1 e X_2) simuladas nos diferentes casos estudados (C1, C2, C3, C4 e C5)

Caso	Estatística	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
		X_1	X_2	X_1	X_2	X_1	X_2
C1	Média	99,04	100,35	99,77	100,09	100,17	100,17
	Variância	88,90	78,44	85,21	96,83	98,73	98,73
	CV (%)	9,52	8,83	9,25	9,83	9,92	9,92
C2	Média	297,13	100,35	299,32	100,09	300,51	100,17
	Variância	800,14	78,44	766,90	96,83	888,60	98,73
	CV (%)	9,52	8,83	9,25	9,83	9,92	9,92
C3	Média	9,43	100,35	9,86	100,09	10,10	100,17
	Variância	32,01	78,44	30,68	96,83	35,54	98,73
	CV (%)	60,02	8,83	56,15	9,83	59,02	9,92
C4	Média	99,43	100,35	99,86	100,09	100,10	100,17
	Variância	32,01	78,44	30,68	96,83	35,54	98,73
	CV (%)	5,69	8,83	5,55	9,83	5,96	9,92
C5	Média	9,04	100,35	9,77	100,09	10,17	100,17
	Variância	88,90	78,44	85,21	96,83	98,73	98,73
	CV (%)	104,28	8,83	94,44	9,83	97,71	9,92

3.1. Componentes Principais baseados nas Variáveis Originais

Na Tabela 3 são apresentadas as estimativas dos autovalores dos CPs obtidos a partir dos dados originais e, portanto, da decomposição da matriz de variâncias e covariâncias (*S*), para as duas variáveis X_1 e X_2 avaliadas nos diferentes casos estudados.

Tabela 3. Estimativas dos autovalores obtidos pela matriz *S* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

	$r_{12} = 0{,}01$			$r_{12} = 0{,}51$			$r_{12} = 1$		
Caso	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$
C1	88,96	78,39	53,20	137,82	44,22	75,70	197,47	0	100
C2	800,14	78,44	91,10	794,71	69,02	92,00	987,35	0	100
C3	78,45	32,00	71,00	107,00	20,50	83,90	134,28	0	100
C4	78,45	32,00	71,00	107,00	20,50	83,90	134,28	0	100
C5	88,95	78,39	53,20	137,82	44,20	75,70	197,47	0	100

$IR_{\lambda_1}(\%)$ = importância relativa do CP_1.

Para $r_{12} \cong 0$, as estimativas dos autovalores ($\hat{\lambda}_1$ e $\hat{\lambda}_2$) foram aproximadamente iguais à maior e menor variâncias das variáveis X_1 e X_2, respectivamente (Tabelas 2 e 3). Isso mostra que o grau de explicação de cada componente foi diretamente proporcional à quantidade da variância de cada variável, independentemente se ela esteve ou não relacionada à escala.

Essa situação pode ser observada no caso C3, onde a maior variância de X_2 (78,44) não foi relacionada à maior variabilidade relativa (CV=8,83%). Portanto, simplesmente uma mudança na escala de X_1 trará alterações significativas nas estimativas dos autovalores obtidos em uma outra estratégia de análise de CPs.

Como pode-se observar para $r_{12} \cong 0$ (Tabelas 2 e 3), a soma das estimativas dos autovalores ($\hat{\lambda}_1 + \hat{\lambda}_2$) foi igual à soma das estimativas das variâncias ($s_1^2 + s_2^2$), sendo a estimativa do primeiro autovalor obtida por:

$$\hat{\lambda}_1 = s_j^2 + r_{jj'} s_j^2, \text{ em que :}$$

$$s_j^2 \geq s_{j'}^2, \text{ para } j, j' = 1, 2;$$

$$\hat{\lambda}_2 = s_j^2 + s_{j'}^2 - \hat{\lambda}_1.$$

Percebe-se que nos casos em que se utilizaram os dados originais (não transformados), as variáveis com maior variância foram as de maior importância no primeiro CP, aquele que explicou a maior parte da variação total. Essa influência poderá ser drástica, quando houver uma discrepância muito acentuada entre as variâncias. Isto pode ser observado em todos os casos de $r_{12} \cong 0$, dado que a estimativa de λ_1 foi muito parecida com a da maior variância (Tabelas 3 e 4).

Portanto, a existência de variâncias discrepantes entre as variáveis, devido à escala de medida ou à própria variação, proporcionou maior importância relativa ao primeiro autovalor, o que será um problema, quando esta importância for devida à escala e não à variabilidade propriamente dita. Este fato pode ser observado nos casos C3 e C5, onde a variável X_2, de maior importância nos CPs por apresentar a maior variância, foi a de menor variação relativa, ou seja, de menor CV (Tabelas 2 e 3).

O aumento da correlação, em módulo, aumentou a importância do primeiro CP, até 100%, para $r_{12} = 1$. No entanto, quando a diferença entre as variâncias foi grande (C2), esse aumento foi desprezível (Tabela 3).

Portanto, quando se utilizou a matriz *S*, a magnitude dos autovalores dos CPs esteve diretamente relacionada à magnitude das variâncias e covariâncias entre as variáveis estudadas. Portanto, as estimativas dos autovetores normalizados (Tabela 4), que ponderam os valores das variáveis X_1 e X_2 nos respectivos CPs, também foram proporcionais às estimativas dos respectivos autovalores.

Tabela 4. Estimativas dos autovetores normalizados, associados ao primeiro autovalor, obtidas pela matriz *S* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
Caso	$\hat{a}_{11}^*$	$\hat{a}_{12}^*$	$\hat{a}_{11}^*$	$\hat{a}_{12}^*$	$\hat{a}_{11}^*$	$\hat{a}_{12}^*$
C1	0,998	0,069	0,662	0,750	0,707	0,707
C2	1,000	0,003	0,981	0,196	0,949	0,316
C3	0,009	1,000	0,343	0,939	0,514	0,857
C4	0,009	1,000	0,343	0,939	0,514	0,857
C5	0,998	0,069	0,662	0,750	0,707	0,707

Do mesmo modo, o aumento da correlação, em módulo, aumentou a importância da variável com menor variância. No entanto, quando a diferença entre as variâncias foi grande (C2), esse aumento foi menor (Tabela 4).

3.2 Componentes Principais baseados nas Variáveis Padronizadas

No caso de se utilizar as variáveis padronizadas, a matriz *S* dos dados transformados será equivalente à matriz de correlações amostral (*R*) dos dados originais.

Na Tabela 5 são apresentadas as estimativas dos autovalores dos CPs obtidos a partir da matriz de variâncias e covariâncias dos dados padronizados (matriz *R*), para as duas variáveis X_1 e X_2 nos diferentes casos estudados.

Tabela 5. Estimativas dos autovalores obtidas pela matriz *R* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

	$r_{12} = 0{,}01$			$r_{12} = 0{,}51$			$r_{12} = 1$		
Caso	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$IR_{\lambda_1}(\%)$
C1	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C2	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C3	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C4	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100
C5	1,01	0,99	50,40	1,51	0,49	75,60	2	0	100

$IR_{\lambda_1}(\%)$ = importância relativa do CP_1.

Observou-se que a soma das estimativas dos autovalores ($\hat{\lambda}_1 + \hat{\lambda}_2$) foi igual ao traço da matriz *R* [traço(*R*)=2], sendo que as estimativas dos autovalores foram obtidas por:

$$\hat{\lambda}_1 = 1 + r_{12} \text{ e } \hat{\lambda}_2 = 2 - \hat{\lambda}_1.$$

Além disso, todos os cinco casos apresentaram o mesmo resultado para todos os três coeficientes de correlação. Portanto, a média e a variância de cada variável original não interferiram nas estimativas dos CPs (Tabela 5).

Observou-se que a técnica dos CPs utilizando os dados padronizados, resolveu o problema da discrepância na escala de medida das variáveis. Porém, todas passaram a ser igualmente importantes (Tabela 6). E se o objetivo do estudo for o de encontrar as

variáveis mais importantes em discriminar os elementos amostrais, essa estratégia não ajudará.

Tabela 6. Estimativas dos autovetores normalizados associados ao primeiro autovalor obtidas pela matriz *R* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
Caso	$\hat{a}_{11}^{*}$	$\hat{a}_{12}^{*}$	$\hat{a}_{11}^{*}$	$\hat{a}_{12}^{*}$	$\hat{a}_{11}^{*}$	$\hat{a}_{12}^{*}$
C1	0,707	0,707	0,707	0,707	0,707	0,707
C2	0,707	0,707	0,707	0,707	0,707	0,707
C3	0,707	0,707	0,707	0,707	0,707	0,707
C4	0,707	0,707	0,707	0,707	0,707	0,707
C5	0,707	0,707	0,707	0,707	0,707	0,707

Além disso, é importante ressaltar que os coeficientes dos CPs obtidos pela matriz *S* não foram numericamente iguais aos da matriz *R*. Em geral, o percentual de variação explicado pela matriz *S* foi mais concentrado no primeiro CP. Na *R* a concentração foi menor. Portanto, de acordo com a última matriz, houve melhor distribuição da variabilidade e desta forma, foi necessário um maior número de componentes para explicar a mesma quantidade da variância total.

De modo geral, a matriz *S* deu maior importância às variáveis com variâncias mais altas, importâncias iguais às variáveis de mesma variância, independentemente da escala de medida dos dados, o que não foi bom, quando o primeiro componente foi dominado pela variável de maior variância devido à escala. Já a matriz *R* deu importâncias iguais, de acordo com a correlação entre as variáveis, independentemente da escala, o que também foi um problema, quando se desejou encontrar as variáveis mais importantes em discriminar os elementos amostrais.

Logo, se a matriz *R* dá importâncias iguais a todas as variáveis e a *S* privilegia as de maiores variâncias absolutas, então foi necessário estabelecer uma transformação nos dados de forma que elas estivessem numa mesma escala, mas com variâncias proporcionais às variabilidades relativas que elas explicam.

3.3. Componentes Principais baseados nas Variáveis Transformadas

Para contornar o problema das discrepâncias das variâncias devidas às diferenças de escalas, as variáveis foram transformadas de forma que estivessem com mesma média, porém com suas variâncias ponderadas pelos respectivos coeficientes de variação (CV). Com os dados transformados (Z^*), procedeu-se a análise utilizando-se a matriz de variâncias e covariâncias (S^*), sendo as estimativas dos autovalores apresentadas na Tabela 7.

Tabela 7. Estimativas dos autovalores obtidas pela matriz S^* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

	$r_{12} = 0{,}01$			$r_{12} = 0{,}51$			$r_{12} = 1$		
Caso	$\hat{\lambda}_1$	$\hat{\lambda}_2$	IR_{λ_1} (%)	$\hat{\lambda}_1$	$\hat{\lambda}_2$	IR_{λ_1} (%)	$\hat{\lambda}_1$	$\hat{\lambda}_2$	IR_{λ_1} (%)
C1	0,008	0,006	56,90	0,013	0,004	76,10	0,019	0	100
C2	0,008	0,006	56,90	0,013	0,004	76,10	0,019	0	100
C3	0,320	0,006	98,10	0,271	0,007	97,50	0,354	0	100
C4	0,006	0,003	68,00	0,010	0,002	85,20	0,013	0	100
C5	0,967	0,006	99,40	0,763	0,007	99,10	0,952	0	100

Observou-se no caso C2, para $r_{12} = 0{,}01$**,** que foram necessários dois componentes para explicarem a variação dos dados, uma vez que o primeiro CP explicou somente 56,90% desta variação (Tabela 7). Este resultado foi diferente do observado quando se utilizou os dados originais, onde somente o primeiro componente foi necessário para explicar mais de 91% da variação total. Nesse caso, a variável mais importante foi a X_1, cuja variância foi muito superior à de X_2. Porém, esta variância foi devida somente à escala e não a uma variação real, dado que ambas apresentaram CVs semelhantes (Tabela 2). Portanto, o grau de explicação do CP_1 foi mais próximo daquele obtido pela aplicação da matriz de correlações aos dados originais, que considerou a mesma variabilidade das duas variáveis padronizadas. Isso implicou que as transformações das variáveis em função das suas variabilidades relativas medidas pelos CVs, foram eficientes em diminuírem a importância da variável que apresentou alta variância em função da escala de medição.

Nos casos C3 e C5, o CP_1 explicou quase que totalmente a variação total, sendo este componente dominado pelas variáveis X_1 no C3 e X_2 no C5, que possuíram os maiores coeficientes de variação (Tabelas 2 e 7). Este resultado foi bastante interessante quando comparado aos casos C3 e C5 da análise baseada na matriz *S*, onde as variáveis X_2 e X_1 foram consideradas as mais importantes em função das suas maiores variâncias, respectivamente.

Os resultados apresentados para o caso C4 foram similares aos obtidos utilizando-se os dados originais. Este resultado era esperado, pois ambas as variáveis estavam numa mesma escala de medida e, portanto, a maior variância associada a X_2 não foi proveniente da diferença entre as mesmas (Tabela 7).

Quando $r_{12} = 1$, ou seja, quando as variáveis foram totalmente correlacionadas, foi necessário apenas um componente para explicar toda a variação contida nos dados, sendo que o coeficiente de maior grandeza nestes componentes esteve associado à variável de maior variação relativa (CV), que foi considerada a mais importante na discriminação dos elementos amostrais (Tabelas 7 e 8).

Tabela 8. Estimativas dos autovetores normalizados associados ao primeiro autovalor obtidas pela matriz S^* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

	$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
Caso	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$	$\hat{a}^*_{11}$	$\hat{a}^*_{12}$
C1	1,000	0,031	0,617	0,787	0,707	0,707
C2	1,000	0,031	0,617	0,787	0,707	0,707
C3	1,000	0,001	0,995	0,097	0,986	0,166
C4	0,011	1,000	0,317	0,949	0,515	0,857
C5	1,000	0,001	0,998	0,057	0,995	0,101

Para todos os casos estudados, as variáveis de maior variabilidade relativa (CV), foram aquelas que apresentaram os maiores coeficientes no primeiro componente, aquele que explicou a maior parte da variabilidade contida nos dados.

Os autovetores normalizados estimados a partir da matriz *S* baseada nos dados originais (Tabela 4) foram semelhantes àqueles obtidos a partir da mesma matriz baseada nos dados transformados (Tabela 8). Como pode-se observar, um coeficiente é alto e outro, baixo, em valores absolutos. A diferença é que o coeficiente alto pondera as

variáveis de maiores variância e CV, respectivamente. Outra diferença foi o grau de explicação de cada autovetor, medido pela magnitude do respectivo autovalor. Portanto, a transformação dos dados não proporcionou estimativas desses coeficientes mais equilibradas, como aquelas oriundas da matriz *R* (Tabela 6). Desse modo, pode-se concluir que qualquer diferença de variabilidade, por menor que seja, entre as variáveis, promoverá uma grande desproporcionalidade entre os valores absolutos dos coeficientes normalizados dos autovetores.

Apesar da grande diferença provocada pelas matrizes *R*, *S* e S^* sobre as estimativas dos autovalores e autovetores normalizados, a coincidência entre as classificações dos 50 elementos amostrais com maiores escores do CP_1 foi bastante alta, para $r_{12} \geq 0{,}5$ (Tabela 9). Isso reforçou a importância da correlação na qualidade de uma análise multivariada.

Tabela 9. Coeficiente de coincidência entre as classificações dos 50 elementos amostrais com os maiores escores do CP_1 fornecidos pelas matrizes *R*, *S* e S^* nos diferentes casos estudados (C1, C2, C3, C4 e C5)

		$r_{12} = 0{,}01$		$r_{12} = 0{,}51$		$r_{12} = 1$	
		R	***S****	***R***	***S****	***R***	***S****
C1	***S***	0,62	0,98	0,98	0,72	1	1
	R	-	0, 62	-	0,96	-	1
C2	***S***	0,62	0,98	0,74	0,72	1	1
	R	-	0,62	-	0,96	-	1
C3	***S***	0,62	0,24	0,82	0,54	1	1
	R	-	0,62	-	0,76	-	1
C4	***S***	0,62	1	0,80	0,98	1	1
	R	-	0,62	-	0,82	-	1
C5	***S***	0	0,98	0,98	0,72	1	1
	R	-	0,02	-	0,76	-	1

4. CONCLUSÕES

- A análise de componentes principais baseada na matriz de variâncias e covariâncias dos dados originais leva em consideração a variância e a covariância independentemente da média e do coeficiente de variação. Desta forma, as estimativas dos componentes podem ser prejudicadas quando a variabilidade dos dados for inerente à escala.

- Quando se utiliza a matriz de correlações entre as variáveis originais, as variâncias das variáveis não são importantes para as estimativas dos componentes e, portanto, para a classificação dos elementos amostrais. Desta forma, apesar de contornar o problema da escala dos dados, ela faz com que todas as variáveis tenham a mesma importância, o que pode não ser útil quando se pretende identificar aquelas variáveis com maior grau de discriminação.

- A análise a partir das variáveis transformadas utilizando o CV como ponderação, proporciona a formação dos componentes com base nas variáveis de maior variabilidade relativa. Esta transformação contorna as limitações das análises baseadas nas matrizes de variâncias e covariâncias e de correlações das variáveis originais, dado que as mesmas passam a estar numa mesma escala, porém com variâncias diferentes.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, L.; LOPES, P.S.; REGAZZI, A.J.; GUIMARÃES, S.E.F.; TORRES, R. A. de. Seleção de variáveis de desempenho de suínos por meio da análise de componentes principais. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte, v. 57, n.6, p. 805-810, 2005.

CADIMA, J.F.C.L. Redução de Dimensionalidade através duma Análise em Componentes Principais: Um Critério para o Número de C.P. a Reter. **Revista de Estatística – INE.** v.3, p.9, 2001. Disponível em http://www.ine.pt/prodserv/estudos/debito.asp?x_estudoid=208. Acesso Agosto, 2008.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** 3. ed., Viçosa: Universidade Federal de Viçosa, 2004. 480p.

DAWKINS, B. Multivariate analysis of national track records. **The American Statistician**, v.43, p.110-115, 1989.

FERREIRA, D. F. **Estatística Multivariada**. 1. ed. Lavras: Editora UFLA, 2008. 662p.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 5. ed. New Jersey: Prentice Hall, 2002. 767p.

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

JOLLIFFE, I.T. Discarding variables in a principal component analysis. II. Real data. **Applied Statistics**, Londres, v. 22, p. 21- 31, 1973.

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

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2007. 297p.

NAIK, D.N.; KATTREE, R. Revisiting Olympic track records: some practical considerations in the principal component analysis. **The American Statistician**, v.50, n.2, p.140-144, 1996.

R DEVELOPMENT CORE TEAM (2007). **R: A language and environment for statistical computing. R Fundation for Statistical Computing**. Vienna, Austria. ISBN 3-900051-07-0. Disponível em: http://r-project .org

STEARNS, T.M.; BEEVER, J.E.; SOUTHEY, B.R.; ELLIS, M.; MCKEITH, F.K.; RODRIGUEZ-ZAS, S.L. Evaluation of approaches to detect quantitative traid loci for growth, carcass, and meat quality on swine chromosomes 2, 6, 13, and 18. II. Multivariate and principal component analysis. **American Society of Animal Science**, Savoy, v. 83, p. 2471-2481, 2005.

CAPÍTULO 2

ANÁLISE DE COMPONENTES PRINCIPAIS PARA O AGRUPAMENTO DE PRODUTORES DE LEITE POR MEIO VARIÁVEIS ECONÔMICAS

RESUMO

O objetivo deste trabalho foi de verificar qual a melhor estrutura de dados em classificar de forma mais apropriada os produtores com base em variáveis econômicas. Para tanto, utilizaram-se técnicas de estatística multivariada aos dados referentes às variáveis econômicas e zootécnicas de 255 produtores de leite de três regiões do estado de Minas Gerais participantes do projeto de assistência técnica e gerencial Educampo. A coleta dos dados foi feita mensalmente de maio de 2006 a abril de 2007. As transformações nos dados proporcionaram diferentes estimativas dos componentes principais. Percebeu-se que a utilização da matriz de variâncias e covariâncias (*S*) privilegiou as variáveis econômicas de maiores variâncias na obtenção dos CPs, já a matriz de correlações (*R*) considerou as variáveis mais correlacionadas como as mais importantes na formação dos CPs. A obtenção dos CPs com base na matriz de variâncias e covariâncias das variáveis transformadas pelos respectivos coeficientes de variação (S^*), minimizou os problemas de escala inerentes ao uso da matriz *S* e sem considerar a mesma variabilidade para todas as variáveis (matriz *R*). As variáveis econômicas com os maiores coeficientes de variação foram as mais relacionadas com as variáveis zootécnicas. Isto fez com que a estrutura de dados utilizada na obtenção dos CPs via matriz S^* fosse considerada a mais indicada na discriminação dos produtores.

1. INTRODUÇÃO

Grandes transformações têm marcado a produção de leite brasileira nos últimos anos. Tais mudanças estão principalmente associadas aos impactos advindos da estabilização monetária, desregulamentação do mercado (fim do controle estatal sobre os preços), da abertura econômica e da mudança nos padrões de consumo da população, que exigem dos produtores recorrentes adaptações no sentido de se modernizarem, buscando adequar-se à nova conjuntura e melhorar a competitividade (GOMES, 2000).

Essas tranformações geram a necessidade de adoção de técnicas gerenciais mais sofisticadas, que possam contribuir para a eficiência administrativa e produtiva dos empresários rurais, fazendo com que os mesmos busquem obter não apenas informações sobre produção e tecnologia, mas conceitos administrativos nas diversas áreas da empresa rural.

Para a empresa rural alcançar níveis satisfatórios de confiabilidade e resultado, é necessário possuir um sistema de gerenciamento eficaz, fundamentado em informações de qualidade, que promova a profissionalização de sua administração. E, na era do conhecimento, a informação tornou-se a ferramenta mais importante para a humanidade.

Diante deste cenário, surge a necessidade de armazenar e explorar informações provenientes desta atividade, para a utilização na solução dos diversos problemas. Para tanto, diferentes metodologias surgem como ferramentas de análise aplicadas a várias variáveis medidas simultaneamente em cada elemento amostral. Dentre estas, destacam-se os métodos estatísticos de análise multivariada.

Durante os últimos anos tem-se verificado uma infinidade de trabalhos ligados à pecuária de leite que fazem uso de técnicas multivariadas. Estes trabalhos visam principalmente caracterizar e agrupar os produtores de forma que seja possível realizar

ações regionalizadas, viabilizando intervenções técnicas diferenciadas, o que permite a consolidação de condições de sustentabilidade a partir das reais necessidades de incorporação tecnológica dos produtores. Dentre estes trabalhos, podem-se citar o trabalho de Aleixo et al. (2007), onde se objetivaram identificar grupos de produtores, o mais semelhantes possíveis, por meio do conjunto de variáveis e características selecionadas. Neto et al. (2005) objetivaram identificar e caracterizar sistemas de produção de leite por meio de diferentes métodos de análise multivariada. Fernandes et al. (2004) utilizaram a análise de agrupamento, seguida de análise discriminante, com o objetivo de reunir os municípios da Região Sul em áreas de produção de leite com o mesmo padrão de similaridade.

Dentre as técnicas de análise multivariada que objetivam a redução da dimensionalidade dos dados, a análise de componentes principais destaca-se, por proporcionar uma simplificação considerável nos cálculos estatísticos e na interpretação dos resultados (CRUZ e CARNEIRO, 2003), possibilitando assim, a identificação das variáveis que mais contribuem na discriminação dos elementos amostrais.

A obtenção dos componentes principais envolve a decomposição da matriz de variâncias e covariâncias das variáveis estudadas. Entretanto, estes componentes não são invariantes em relação às transformações nas escalas (REIS, 1997). Desse modo, percebe-se que a análise de componentes principais pode levar a diferentes resultados e, consequentemente, alguns deles podem estar associados a interpretações e conclusões errôneas do estudo realizado.

Desta forma, este trabalho tem por objetivo, a partir de 18 variáveis zootécnicas e econômicas, referentes a 255 produtores de leite de três regiões do estado de Minas Gerais, verificar qual é a melhor estrutura de dados, dentre as três consideradas no estudo, em classificar de forma mais apropriada os produtores mais viáveis economicamente.

2. MATERIAL E MÉTODOS

Para a realização do presente estudo de caso, utilizaram-se as informações cedidas pela Central de Processamento de Dados do Educampo/Sebrae-MG, cujos dados foram referentes às variáveis econômicas e zootécnicas de 255 produtores de leite de três regiões do estado de Minas Gerais (Triângulo/Alto do Paranaíba, Central e Vale do Mucuri), participantes do projeto de assistência técnica e gerencial. A coleta dos dados foi feita mensalmente de maio de 2006 a abril de 2007.

Foram consideradas as seguintes variáveis econômicas:

- Y_1: renda bruta da atividade leiteira (R$/ano)= venda de leite + venda de animais;
- Y_2: preço médio do leite (R$/litro) = valor médio unitário do leite recebido, incluindo frete;
- Y_3: custo operacional efetivo do leite (R$/litro) = COE x (($Y_1$ / venda do leite) x 100) / produção anual de leite. Em que, COE = gastos com mão-de-obra contratada, concentrados, manutenção de forrageiras não-anuais, mineralização, sanidade, energia e combustíveis, material de ordenha, inseminação artificial, frete de leite, impostos e taxas, reparos em benfeitorias e máquinas e outras despesas de custeio) x ((Y_1 / venda do leite) x 100) / produção anual de leite;
- Y_4: custo total do leite (R$/litro) = (COE + mão-de-obra familiar + depreciação de máquinas, benfeitorias, forrageiras não-anuais e animais de serviços + remuneração do capital médio investido em animais, benfeitorias, máquinas, forrageiras não anuais) x ((Y_1 / venda do leite) x 100) / produção anual de leite;
- Y_5: gasto com mão-de-obra/ renda bruta do leite (%) = gasto com mão-de-obra para atividade leiteira em relação ao valor da produção de leite;

- Y_6: gasto com concentrado na atividade/ renda bruta do leite (%) = gasto com concentrado para o rebanho em relação ao valor da produção de leite;
- Y_7: margem bruta/ área (R$/ha) = (COE – Y_1)/área;
- Y_8: taxa de remuneração do capital sem terra (% a.a.) = ((Y_1 – (COE + mão-de-obra familiar + depreciação de máquinas, benfeitorias, forrageiras não-anuais e animais de serviços))/(capital médio investido em animais, benfeitorias, máquinas, forrageiras não-anuais);
- Y_9: taxa de remuneração do capital com terra (% a.a.) = ((Y_1 – (COE + mão-de-obra familiar + depreciação de máquinas, benfeitorias, forrageiras não-anuais e animais de serviços))/(capital médio investido em animais, benfeitorias, máquinas, forrageiras não-anuais e terra).

As variáveis zootécnicas utilizadas foram:

- X_1: produção média de leite (litros/dia);
- X_2: área usada para pecuária (ha);
- X_3: mão-de-obra permanente para a pecuária (homem/ano);
- X_4: vacas em lactação/ total de vacas (%);
- X_5: vacas em lactação/ rebanho (%);
- X_6: vacas em lactação/ área para pecuária (cabeças);
- X_7: produção/ vaca em lactação (litros/dia);
- X_8: produção/ mão-de-obra permanente (litros/dh);
- X_9: produção/ área para pecuária (litros/ha/ano).

Foram realizadas análises descritivas para todas as variáveis econômicas estudadas. Além disso, estimou-se os coeficientes de correlações amostrais, que foram testados a partir do teste t ao nível de 5% de probabilidade.

Neste estudo as variáveis econômicas foram denominadas como dependentes e estudadas por meio da análise de componentes principais (CP), cujas estimativas foram baseadas nas matrizes de variâncias e covariâncias dos dados originais (S), padronizados (R) e transformados de acordo com o coeficiente de variação (S^*). Desse modo, buscou-se verificar qual a melhor estrutura de dados em classificar de forma mais apropriada os produtores mais viáveis economicamente.

Os CPs foram determinados como combinações lineares das variáveis econômicas (*Ys*), sendo o CP_j relativo às 255 observações estimado por:

$$\hat{C}P_{ij} = \hat{\mathbf{a}}^*_{1j} Y_{1i} + \hat{\mathbf{a}}^*_{2j} Y_{2i} + \hat{\mathbf{a}}^*_{3j} Y_{3i} + \cdots + \hat{\mathbf{a}}^*_{9j} Y_{9i},$$

em que:

$\hat{\mathbf{a}}^*_j$ = estimativa do autovetor normalizado do CP_j.

A padronização e a transformação dos dados foram obtidas respectivamente por:

$$z_{ij} = \frac{y_{ij} - \overline{Y}_j}{s_j};$$

$z^*_{ij} = z_{ij} \times CV_j$, em que:

y_{ij} = valor da *i*-ésima observação da variável econômica Y_j;

z_{ij} = valor padronizado na *i*-ésima observação da variável econômica Y_j;

z^*_{ij} = valor transformado na *i*-ésima observação da variável econômica Y_j;

CV_j = coeficiente de variação da variável econômica Y_j;

$\overline{Y}_j$ = média amostral da variável econômica Y_j;

s_j = desvio-padrão amostral da variável econômica Y_j.

Como critérios para a determinação do número de CPs utilizaram-se o *scree-plot* (CATTELL, 1966) e os *k* primeiros CPs que fossem suficientes para explicar 70% ou mais da variação total dos dados.

Para cada CP calculou-se os escores de cada elemento amostral (produtor), por meio das matrizes *S*, *R* e S^*. Posteriormente, os produtores foram agrupados a partir dos escores dos *k* CPs utilizados para a interpretação dos dados. O agrupamento foi realizado por meio dos métodos hierárquicos de ligação média (UPGMA) e Ward, utilizando a distância Euclidiana como medida de dissimilaridade.

O método de ligação UPGMA trata a distância entre dois grupos como a média da distância entre todos os pares de elementos pertencentes a cada grupo. Desta forma, se o grupo 1 (g_1) tem n_1 elementos e o grupo 2 (g_2) tem n_2, a distância entre eles será definida por (MINGOTI, 2007):

$$d(g_1, g_2) = \sum_{j \in g_1} \sum_{k \in g_2} \left(\frac{1}{n_1 n_2} \right) d(j,k),$$

em que:

d(j,k) = distância entre os produtores *j* e *k* pertencentes aos grupos g_1 e g_2 respectivamente.

Já o de método de Ward, também chamado de "variância mínima", baseia-se na "mudança de variação" causada pela inclusão de um elemento num determinado grupo (HAIR et al., 2005). A alocação de um elemento a um grupo é feita maximizando a homogeneidade dentro dos grupos ou minimizando o total das somas de quadrados dentro de grupos, também conhecida como soma de quadrados dos erros (ESS), que é calculada por:

$$ESS = \sum_{i=1}^{n} (CP_i - C\overline{P})'(CP_i - C\overline{P}),$$

onde CP_i é o vetor multivariado dos *k* escores associados ao *i*-ésimo produtor e $C\overline{P}$ é a média de todos os escores.

A distância Euclidiana foi definida como a raiz quadrada da soma dos quadrados das diferenças entre os diferentes escores dos produtores, isto é, a distância Euclidiana entre dois produtores *i* e *i'* baseada nos escores do CP_j foi definida por:

$$d(CP_{ij} - CP_{i'j}) = [(CP_{ij} - CP_{i'j})'(CP_{ij} - CP_{i'j})]^{1/2} = \left[\sum_{j=1}^{k} (CP_{ij} - CP_{i'j})^2\right]^{1/2},$$

em que:

CP_{ij}, = escore do *i*-ésimo produtor no componente *j*.

Para a determinação do número final de grupos (*g*), após a construção do dendrograma em função das três matrizes definidas, adotou-se como critério, um nível de similaridade de aproximadamente 80% entre os produtores de um mesmo grupo.

De posse dos grupos formados, foram realizadas análises de variância (ANOVA) para cada uma das três matrizes e métodos de agrupamento, de acordo com o seguinte modelo estatístico para cada uma das nove variáveis econômicas *Ys* (MONTGOMERY, 1997):

$y_{ij} = \mu + g_i + e_{ij}$, em que:

μ = média geral;

y_{ij} = valor observado da variável *Y* no grupo *i* e no produtor *j*;

g_i = efeito do grupo *i*;

e_{ij} = efeito do erro experimental associado ao valor observado y_{ij}.

As ANOVAS indiciduais das nove características econômicas, com base nos grupos formados de acordo com os *k* CPs escolhidos, permitiram obter os quadrados médios entre grupos (QMG) e residuais (QMRes), além das razões entre eles medidas pelo teste *F*. Desse modo, espera-se que quanto maior for o valor calculado da estatística *F* para uma dada variável econômica, ou seja, quanto maior o QMG e menor o QMRes, melhor será a formação dos grupos proporcionada pela respectiva matriz. Nesse caso, maior será a homogeneidade dentro dos grupos e a heterogeneidade entre eles.

Além das ANOVAS, foram realizadas análises de regressão lineares múltiplas para cada uma das nove variáveis econômicas em função das nove variáveis zootécnicas, de acordo com o seguinte modelo completo (MONTGOMERY; PECK, 1992):

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_9 x_{9i} + e_i$, em que:

$y_i =$ valor da *i*-ésima observação da variável econômica *Y*;

$\beta_0 =$ intercepto;

$\beta_j =$ coeficiente angular associado à *j*-ésima variável técnica X_j (*j* =1, 2, 3,..., 9);

$e_i =$ erro aleatório associado à *i*-ésima observação.

Para a seleção da melhor equação de regressão ajustada para cada variável econômica (*Y*), foi utilizado o método *stepwise*, proposto por Efroymson (1960). Neste método, as variáveis são adicionadas, uma por vez, na equação de regressão até que a mesma se torne satisfatória. A cada passo, a variável com a maior correlação parcial com a variável dependente é selecionada com base no p-valor de um teste F parcial. A análise termina assim que alguma variável apresente valor de F não significativo. Uma variável selecionada a pertencer ao modelo pode, em um estágio posterior, ser retirada devido sua relação com as outras variáveis posteriormente inseridas na regressão. Para a entrada e saída das variáveis técnicas do modelo de regressão considerou-se um nível de significância de 0,15.

Comparou-se os modelos finais de regressão com os resultados obtidos em cada análise de CPs (matrizes *S, R* e S^*), a fim de verificar se a variável definida como a mais importante em discriminar os elementos foi aquela que apresentou maior coeficiente de determinação (R^2) e maior número de variáveis regressoras com maiores coeficientes, em módulo.

Todas as análises estatísticas foram processadas por meio dos softwares Microsoft Excel XP 2002 e Minitab 14®.

3. RESULTADOS E DISCUSSÃO

3.1. Componentes Principais

Na Tabela 1 são apresentadas algumas estatísticas descritivas das variáveis zootécnicas (*X*) e econômicas (*Y*) estudadas.

Com respeito às variáveis econômicas, aquelas utilizadas na análise por CPs, percebeu-se grande diferença entre as suas escalas, principalmente entre as variáveis Y_7 e Y_2 que possuíram o maior e o menor desvios-padrão, dados respectivamente, por 1.025,75 e 0,06. Além disso, observou-se que um maior desvio-padrão não implicou necessariamente em maior variabilidade relativa, uma vez que ele esteve relacionado à escala de medida da variável. Isso pode ser observado pela comparação das variáveis Y_8 e Y_9. A primeira, apesar de ter possuído maior desvio-padrão, possuiu menor variação relativa. Desta forma, as estimativas dos componentes dependerão da transformação utilizada, uma vez que os CPs não são invariantes em relação à transformação nas escalas das variáveis (REIS, 1997).

Tabela 1. Médias, desvios-padrão e coeficientes de variação (CV) das variáveis zootécnicas (*X*) e econômicas (*Y*) medidas em 255 produtores de leite do estado de MG

	Variável	Média	Desvio-padrão	CV (%)
$\mathbf{X_1}$	Produção média de leite (litros/dia)	888,95	986,71	111,00
$\mathbf{X_2}$	Área usada para pecuária (ha)	106,62	78,79	73,89
$\mathbf{X_3}$	Mão-de-obra permanente (homem/ano)	1152,06	983,01	85,33
$\mathbf{X_4}$	Vacas em lactação/ total de vacas (%)	74,63	8,30	11,12
$\mathbf{X_5}$	Vacas em lactação/ rebanho (%)	37,16	8,56	23,04
$\mathbf{X_6}$	Vacas em lactação/ área para pecuária (cab./ha)	0,74	0,45	60,09
$\mathbf{X_7}$	Produção/ vacas em lactação (litros/dia)	13,10	4,22	32,21
$\mathbf{X_8}$	Produção/ mão-de-obra permanente (litros/dh)	273,34	119,33	43,65
$\mathbf{X_9}$	Produção/ área para pecuária (litros/ha/ano)	3829,72	3285,09	85,78
$\mathbf{Y_1}$	Renda bruta da atividade leiteira (mil R$/ano)	327,57	362,32	110,61
$\mathbf{Y_2}$	Preço médio do leite (R$/litro)	0,87	0,06	7,09
$\mathbf{Y_3}$	Custo operacional efetivo do leite (R$/litro)	0,61	0,15	24,09
$\mathbf{Y_4}$	Custo total do leite (R$/litro)	0,76	0,16	20,98
$\mathbf{Y_5}$	Gasto com mão-de-obra/ renda bruta do leite (%)	11,40	7,29	63,96
$\mathbf{Y_6}$	Gasto com concentrado/ renda bruta do leite (%)	32,67	8,26	25,28
$\mathbf{Y_7}$	Margem bruta/ área (R$/ha)	915,80	1.025,75	112,00
$\mathbf{Y_8}$	Taxa de remuneração do capital sem terra (% a.a.)	16,99	13,83	81,40
$\mathbf{Y_9}$	Taxa de remuneração do capital com terra (% a.a.)	8,20	8,98	109,45

Na Tabela 2 é apresentada a matriz de correlações entre as nove variáveis econômicas estudadas. Notou-se a existência de correlações significativas (p<0,05) entre quase todas as variáveis, o que justificou o uso da análise por CPs.

Tabela 2. Estimativas das correlações entre as nove variáveis econômicas medidas em 255 produtores de leite do estado de MG

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8
Y_2	0,43*	-	-	-	-	-	-	-
Y_3	0,24*	0,27*	-	-	-	-	-	-
Y_4	0,08	0,18*	0,90*	-	-	-	-	-
Y_5	-0,11*	-0,31*	0,39*	0,41*	-	-	-	-
Y_6	0,12*	-0,01	0,57*	0,42*	0,08	-	-	-
Y_7	0,18*	0,25*	-0,37*	-0,45*	-0,48*	-0,22*	-	-
Y_8	0,13*	0,24*	-0,59*	-0,72*	-0,52*	-0,33*	0,63*	-
Y_9	0,13*	0,25*	-0,39*	-0,51*	-0,45*	-0,23*	0,60*	0,85*

* significativo pelo teste *t* (p<0,05).

3.1.1. Matriz de variâncias e covariâncias (S)

De acordo com a matriz de variâncias e covariâncias (Tabela 3), verificou-se que existiram grandes diferenças entre as variâncias e covariâncias das variáveis econômicas estudadas.

Tabela 3. Matriz de variâncias e covariâncias entre as nove variáveis econômicas medidas em 255 produtores de leite de MG

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
Y_1	131.274,375	-	-	-	-	-	-	-	-
Y_2	9,648	0,004	-	-	-	-	-	-	-
Y_3	12,765	0,002	0,021	-	-	-	-	-	-
Y_4	4,401	0,002	0,021	0,025	-	-	-	-	-
Y_5	-287,374	-0,140	0,415	0,473	53,154	-	-	-	-
Y_6	365,880	0,006	0,685	0,550	5,020	68,211	-	-	-
Y_7	65781,637	15,768	-55,265	-72,765	-3.577,228	-1.822,331	1.052.233,026	-	-
Y_8	635,229	0,205	-1,191	-1,572	-51,916	-37,701	8.924,952	191,223	-
Y_9	433,735	0,136	-0,514	-0,721	-29,598	-17,380	5.532,583	104,881	80,557

Foram estimados os autovalores e os autovetores normalizados associados à matriz *S*, assim como os percentuais das variâncias explicadas por cada CP (Tabela 4).

Tabela 4. Estimativas dos autovalores, autovetores normalizados e variância explicada pelos CPs obtidas pela decomposição da matriz *S*

Variável	Componentes Principais								
	CP_1	**CP_2**	**CP_3**	**CP_4**	**CP_5**	**CP_6**	**CP_7**	**CP_8**	**CP_9**
Y_1	-0,071	-0,997	0,000	0,004	-0,001	0,000	0,000	0,000	0,000
Y_2	0,000	0,000	-0,001	0,000	-0,002	0,001	-0,252	-0,874	-0,416
Y_3	0,000	0,000	0,006	-0,005	0,005	0,007	-0,645	-0,169	0,745
Y_4	0,000	0,000	0,007	-0,002	0,002	0,010	-0,721	0,457	-0,521
Y_5	0,003	0,000	0,191	0,266	0,938	-0,113	0,003	-0,002	-0,003
Y_6	0,002	-0,004	0,231	-0,944	0,207	-0,117	0,007	-0,001	-0,003
Y_7	-0,997	0,071	0,011	0,001	0,001	0,000	0,000	0,000	0,000
Y_8	-0,008	0,000	-0,827	-0,106	0,134	-0,536	-0,012	0,003	0,000
Y_9	-0,005	0,000	-0,476	-0,166	0,243	0,828	0,007	-0,001	-0,001
Autovalor	1,057,028	126,601	160	60	35	17	0	0	0
Variância	0,893	0,107	0	0	0	0	0	0	0
Variância Acumulada	0,893	1	1	1	1	1	1	1	1

De acordo com o critério utilizado, apenas o primeiro componente explicou quase que 90% da variação total e, consequentemente, suficiente para explicar as nove variáveis econômicas (Tabela 4 e Figura 1). A redução do número de variáveis já era esperada, pelo fato de se ter variáveis correlacionadas e com variâncias muito discrepantes entre si.

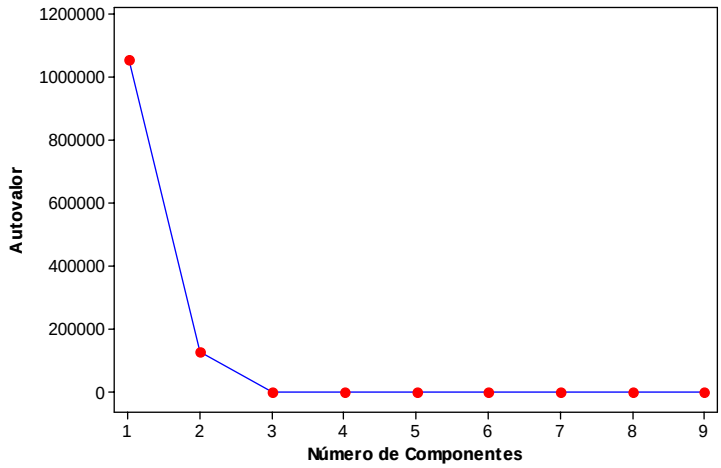

Figura 1. *Scree-plot* dos nove CPs obtidos pela matriz *S*.

Observou-se que o primeiro CP relacionou-se negativamente com margem bruta/área (Y_7) e o segundo com renda bruta da atividade leiteira (Y_1), exatamente aquelas variáveis que apresentaram maiores variâncias (Tabela 1).

Segundo Mardia et al. (1997), as variáveis altamente correlacionadas com os CPs de menores variâncias (autovalores) representam variações praticamente insignificantes, sendo estas passíveis de descarte em ensaios futuros. Desta forma, as variáveis aqui consideradas como pouco significativas para análise foram: preço médio do leite (Y_2), custo operacional efetivo do leite (Y_3) e custo total do leite (Y_4).

3.1.2. Matriz de correlações (R)

Com o objetivo de contornar o problema causado pela grande diferença entre as escalas, as variáveis foram padronizadas com média zero e com variância um. A decomposição da matriz *S* das variáveis padronizadas foi equivalente em decompor a matriz de correlações (*R*) das variáveis originais.

Na Tabela 5 são apresentadas as estimativas dos autovalores e autovetores normalizados dos CPs obtidos a partir da matriz de variâncias e covariâncias dos dados padronizados (matriz *R*), para as nove variáveis econômicas estudadas. Observou-se que a padronização dos dados fez com que houvesse melhor distribuição entre a importância das variáveis nos componentes, assim como a necessidade de um maior número deles em explicar a mesma quantidade de variação total dos dados. Isso se deveu principalmente ao fato de que a maioria das variáveis foram moderadamente ($r \cong 0{,}5$) correlacionadas entre si (Tabela 2).

Tabela 5. Estimativas dos autovalores, autovetores normalizados e variâncias explicadas pelos CPs obtidas pela decomposição da matriz *R*

	Componentes Principais								
Variável	**CP_1**	**CP_2**	**CP_3**	**CP_4**	**CP_5**	**CP_6**	**CP_7**	**CP_8**	**CP_9**
Y_1	-0,029	-0,511	0,268	0,670	-0,419	-0,020	-0,185	0,032	0,079
Y_2	-0,078	-0,577	0,326	-0,435	-0,006	0,205	0,534	0,196	0,006
Y_3	0,391	-0,382	-0,043	-0,042	0,282	-0,032	-0,206	-0,369	-0,663
Y_4	0,420	-0,260	0,069	-0,252	0,231	-0,135	-0,396	-0,068	0,674
Y_5	0,321	0,205	0,356	0,478	0,560	-0,004	0,415	0,086	0,080
Y_6	0,248	-0,253	-0,813	0,213	-0,008	0,118	0,344	0,116	0,151
Y_7	-0,364	-0,206	-0,109	0,022	0,223	-0,862	0,139	0,008	-0,001
Y_8	-0,456	-0,104	-0,051	0,114	0,240	0,298	0,101	-0,738	0,254
Y_9	-0,399	-0,183	-0,114	0,096	0,520	0,303	-0,403	0,504	-0,078
Autovalor	3,995	1,956	0,839	0,684	0,617	0,460	0,289	0,099	0,062
Variância	0,444	0,217	0,093	0,076	0,069	0,051	0,032	0,011	0,007
Variância Acumulada	0,444	0,661	0,754	0,830	0,899	0,950	0,982	0,993	1,000

A partir do *scree-plot* (Figura 2), observou-se a necessidade de três componentes para explicar 75,4% da variação total, dois a mais do que aqueles estimados a partir da matriz *S*.

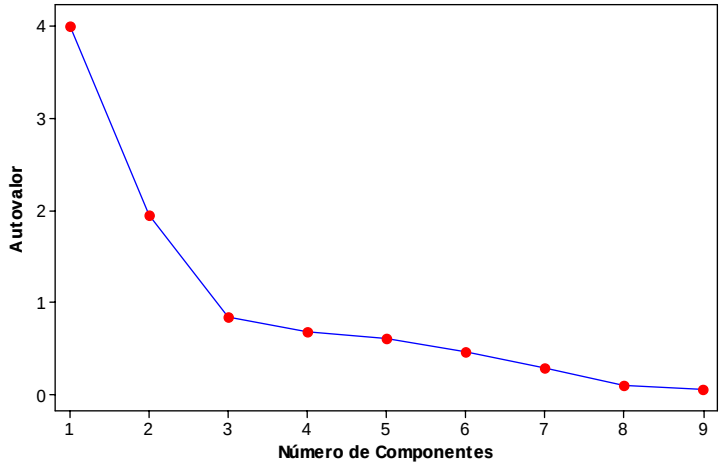

Figura 2. *Scree-plot* dos nove CPs obtidos pela matriz *R*.

O primeiro CP, apesar dos coeficientes serem menores que 0,50, representou um contraste entre as variáveis relativas ao custo operacional (Y_3, Y_4, Y_5 e Y_6) e aquelas de remuneração (Y_7, Y_8 e Y_9). O segundo CP relacionou-se negativamente com as variáveis renda bruta da atividade leiteira (Y_1), preço médio do leite (Y_2) e custo operacional efetivo do leite (Y_3), sendo as duas primeiras de maior peso (Tabela 5). Já o terceiro CP foi dominado pela variável gasto com concentrado na atividade/renda bruta do leite (Y_6).

Observou-se que a análise baseada na matriz *R*, incluiu nos primeiros CPs, as variáveis Y_2, Y_3 e Y_4 que foram consideradas passíveis de descarte pela análise via matriz *S* (Tabelas 4 e 5).

3.1.3. Matriz de variâncias e covariâncias (S^*)

Com o objetivo de contornar o problema causado pela grande diferença entre as escalas e de levar em consideração a variação relativa das variáveis, utilizou-se uma nova transformação, onde as variáveis após serem padronizadas com média zero e variância um, tiveram suas variâncias ponderadas pelos respectivos coeficientes de variação (CVs).

As médias de todas as nove variáveis econômicas foram iguais a zero, sendo a matriz de variâncias e covariâncias (S^*) das variáveis transformadas apresentada na Tabela 6.

Tabela 6. Matriz de variâncias e covariâncias entre as nove variáveis econômicas transformadas medidas em 255 produtores de leite de MG (S^*)

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
Y_1	1,223	-	-	-	-	-	-	-	-
Y_2	0,034	0,005	-	-	-	-	-	-	-
Y_3	0,064	0,005	0,058	-	-	-	-	-	-
Y_4	0,018	0,003	0,045	0,044	-	-	-	-	-
Y_5	-0,077	-0,014	0,060	0,055	0,409	-	-	-	-
Y_6	0,034	0,000	0,034	0,022	0,013	0,064	-	-	-
Y_7	0,219	0,020	-0,099	-0,105	-0,343	-0,061	1,255	-	-
Y_8	0,114	0,014	-0,115	-0,122	-0,268	-0,068	0,574	0,663	-
Y_9	0,161	0,019	-0,103	-0,116	-0,317	-0,065	0,737	0,753	1,198

As estimativas dos autovalores, autovetores normalizados e das variâncias explicadas pelos CPs obtidos a partir dos dados ponderados pelo CV (matriz S^*), para as nove variáveis econômicas estudadas, são apresentadas na Tabela 7.

Tabela 7. Estimativas dos autovalores, autovetores normalizados e variância explicada pelos CPs obtidas pela decomposição da matriz de correlações amostral dos dados ponderados pelo CV (S^*)

	Componentes Principais								
Variável	**CP_1**	**CP_2**	**CP_3**	**CP_4**	**CP_5**	**CP_6**	**CP_7**	**CP_8**	**CP_9**
Y_1	0,197	-0,972	-0,081	-0,010	-0,059	0,063	0,042	0,018	0,011
Y_2	0,014	-0,023	-0,001	-0,019	0,027	-0,035	-0,213	-0,11	-0,969
Y_3	-0,067	-0,092	0,008	0,091	0,325	-0,367	-0,547	-0,628	0,213
Y_4	-0,075	-0,054	0,024	0,060	0,285	-0,152	-0,553	0,758	0,048
Y_5	-0,234	-0,038	-0,073	0,939	-0,221	-0,054	0,058	0,019	-0,040
Y_6	-0,040	-0,053	0,009	-0,029	0,285	-0,782	0,528	0,121	-0,092
Y_7	0,585	0,047	0,778	0,219	0,050	-0,011	0,004	0,001	0,003
Y_8	0,441	0,118	-0,280	-0,074	-0,674	-0,436	-0,240	0,054	0,049
Y_9	0,598	0,151	-0,551	0,228	0,472	0,179	0,092	-0,004	-0,012
Autovalor	2,6814	1,1804	0,5303	0,2836	0,1502	0,0586	0,0284	0,0033	0,0023
Variância	0,545	0,240	0,108	0,058	0,031	0,012	0,006	0,001	0,000
Variância Acumulada	0,545	0,785	0,893	0,951	0,981	0,993	0,999	1,000	1,000

Observou-se que a ponderação dos dados fez com que as variáveis de maiores variabilidades relativas apresentassem maiores importâncias dentro dos primeiros componentes que mais explicaram a variação total dos dados. Além disso, assim como na matriz *R*, houve uma melhor distribuição entre as variâncias explicadas pelos CPs. Porém, os primeiros componentes explicaram maior quantidade de variância que na análise anterior (matriz *R*).

Com base nos critérios adotados, apenas dois componentes foram necessários para explicar 78,5% da variação dos dados (Tabela 7 e Figura 3).

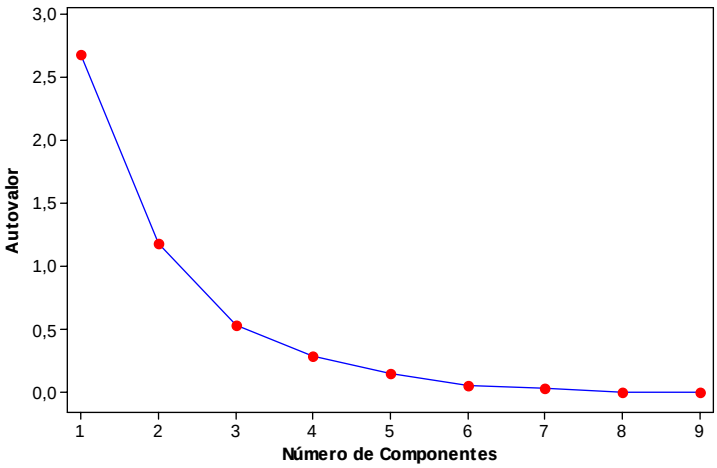

Figura 3. *Scree-plot* dos nove CPs obtidos pela matriz S^*.

Observou-se que as variáveis margem bruta/ área (Y_7), e taxa de remuneração do capital com terra (Y_9) foram as que mais contribuíram para a explicação do primeiro componente. Enquanto que o segundo foi dominado pela variável renda bruta da atividade leiteira (Y_1). Como pode-se observar na Tabela 1, essas três variáveis foram as que apresentaram os maiores CVs.

Quanto às variáveis passíveis de descarte, os resultados foram concordantes com aqueles apresentados pela análise baseada na matriz *R*, sendo as variáveis Y_2, Y_3 e Y_4 as menos importantes em explicarem a variação total deste conjunto de dados (Tabela 7).

As diferenças entre as estimativas dos autovalores e dos autovetores normalizados provocadas pelas três matrizes utilizadas (*S*, *R* e S^*), mostraram claramente as diferentes interpretações que podem ser feitas no estudo da viabilidade econômica desses produtores.

De acordo com a matriz *S*, apenas a variável Y_7 deve ser analisada. As outras oito podem ser desprezadas. No entanto, como Y_7 apresenta uma variância exageradamente superior à das demais, fica a dúvida se ela representa realmente alta variabilidade entre os produtores ou se tal magnitude foi devida à grandeza de sua medida. De acordo com os resultados da Tabela 1, os dois argumentos são verdadeiros. Porém, o inconveniente foi que Y_7 mascarou as influências das variáveis Y_1 e Y_9, com altos CVs (Tabela 1) e das duplas de Y_3 e Y_4 e de Y_8 e Y_9, altamente correlacionadas (Tabela 2), sobre as estimativas dos CPs. Consequentemente, a interpretação das nove variáveis econômicas pelo primeiro CP baseado na matriz *S* (Tabela 4) ficou comprometida.

Já a matriz *R* indicou a necessidade de três CPs para a interpretação econômica dos produtores. Além disso, eles apresentaram interpretações mais complexas. No entanto, houve a necessidade de utilizar as variáveis Y_3, Y_4, Y_5, Y_6, Y_7, Y_8 e Y_9 (Tabela 5). Delas, apenas as variáveis Y_3 e Y_4 e as variáveis Y_8 e Y_9 foram altamente correlacionadas (Tabela 2) e apenas as variáveis Y_7, Y_8 e Y_9 apresentaram altos CVs (Tabela 1). Portanto, a interpretação das variáveis Y_5 e Y_6 parece ser desnecessária. E como a matriz S^* indicou a necessidade de selecionar dois componentes interpretáveis pelas variáveis Y_1, Y_7 e Y_9, de maiores CVs (Tabela 1), mostrou que não foi necessário interpretar, além das variáveis Y_5 e Y_6, as variáveis Y_3 e Y_4 de baixo CVs (Tabela 1) e a variável Y_9 altamente correlacionada com Y_8 (Tabela 2).

Desse modo, concluiu-se que a transformação dos dados de acordo com o CV, além de não sofrer o forte efeito das diferenças de variâncias ao ponto de mascarar os efeitos das covariâncias, corrigiu ao mesmo tempo, essas duas imperfeições.

3.2. Análises de Agrupamento e de Variância

De posse dos escores relativos aos componentes escolhidos, isto é, dos dois primeiros CPs para as matrizes *S* e S^* e dos três primeiros para a matriz *R*, foram realizadas análises de agrupamento por meio dos métodos de ligação UPGMA e de Ward para cada uma das matrizes utilizadas.

De acordo com o critério utilizado, foram criados cinco grupos para cada método e cada matriz estudada, que foram comparados por meio da análise de variância. Apresenta-se na Tabela 8 o número de elementos amostrais (produtores) pertencentes a

cada grupo formado, em todos os agrupamentos realizados. Percebeu-se a existência de grandes diferenças entre o número de produtores nos grupos formados pelas diferentes matrizes e métodos de agrupamento. A heterogeneidade entre e a homogeneidade dos elementos amostrais dentro de cada grupo, foram medidas com base na estatística *F* (Tabela 9).

Tabela 8. Número de elementos amostrais pertencentes aos grupos formados com base nos escores escolhidos das matrizes *S, R* e S^* utilizando-se os métodos de agrupamento UPGMA e Ward

	S		***R***		$\boldsymbol{S^*}$	
	UPGMA	*Ward*	*UPGMA*	*Ward*	*UPGMA*	*Ward*
Grupo 1	238	105	240	44	226	72
Grupo 2	7	81	8	42	21	77
Grupo 3	5	43	2	91	5	63
Grupo 4	4	18	4	38	1	24
Grupo 5	1	8	1	40	2	19

Pode-se observar, a partir da Tabela 9, que quando se utilizou a matriz *S* para a obtenção dos CPs, as variáveis Y_2 e Y_7 foram as que proporcionaram maior poder discriminatório, isto é, os cinco grupos formados através do método de agrupamento UPGMA, foram mais homogêneos dentro e heterogêneos entre eles, uma vez que os valores de *F* obtidos para estas variáveis foram maiores do que aqueles encontrados a partir das matrizes *R* e S^*. Já as variáveis Y_3, Y_4, Y_5, e Y_6 foram as que proporcionaram maiores valores de *F* quando se utilizou a matriz *R* para a obtenção dos CPs. Enquanto que, as variáveis Y_1, Y_8, e Y_9 apresentaram maior poder discriminatório quando se utilizou a matriz S^*.

Tabela 9. Valores da estatística *F* obtidos a partir da ANOVA para os grupos formados pelo método de agrupamento hierárquico da ligação média (UPGMA)

		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
	QMG	4.710,163	0,007	0,272	0,354	147,400	220,000	37.718,548	991,000	427,700
S	QMRes	58,012	0,004	0,018	0,020	51,600	65,800	465,572	178,000	75,000
	F	81,19	1,94	15,51	17,72	2,86	3,34	81,01	5,57	5,70
	QMG	4.795,977	0,005	0,383	0,538	338,300	276,100	20.459,579	2,240	2.106,200
R	QMRes	56,639	0,004	0,016	0,017	48,600	64,900	741,715	158,000	48,100
	F	84,68	1,36	24,36	31,44	6,96	4,25	27,58	14,18	43,79
	QMG	4.950,430	0,007	0,079	0,075	183,000	167,000	25.895,922	3,045	2.578,400
S*	QMRes	54,168	0,004	0,021	0,025	51,100	66,600	654,734	146	40,600
	F	91,39	1,89	3,82	3,07	3,58	2,51	39,55	20,92	63,52

As diferenças entre as discriminações dos grupos baseadas nas diferentes variáveis econômicas mostraram que as diferentes matrizes estudadas (*S*, *R* e S^*) priorizaram a formação dos mesmos baseadas nas variáveis mais importantes na constituição dos seus CPs selecionados. Isso confirmou, então, que dependendo da escolha da matriz, a formação dos grupos poderá não sofrer a maximização da heterogeneidade entre e da homogeneidade dentro dos mesmos. Portanto, é de fundamental importância priorizar a formação dos grupos a partir das variáveis que são mais importantes em discriminá-los. No presente estudo, foram as variáveis Y_1, Y_5, Y_7, Y_8 e Y_9 de maiores CVs (Tabela 1). Dentre elas, a matriz S^* conseguiu formar os melhores grupos para três variáveis, sendo, portanto, considerada como a melhor matriz para classificar os produtores.

Os resultados obtidos a partir dos cinco grupos formados através do método de agrupamento de Ward (Tabela 10) mostraram que a variável Y_7 foi a que possuiu maior poder discriminatório, quando se utilizou a matriz *S* para a obtenção dos CPs. Este resultado já era esperado, uma vez que a matriz *S* privilegiou variáveis que apresentaram maiores variâncias, como foi o caso da variável Y_7 (Tabela 1). Quando se considerou a matriz *R* para a obtenção dos CPs, observou-se que as variáveis Y_2, Y_3, Y_4, e Y_6 foram as que estiveram relacionadas aos grupos mais homogêneos dentro e heterogêneos entre. Já as variáveis Y_1, Y_5, Y_8, e Y_9 foram as que possuíram maiores valores de *F* pela matriz S^* e, consequentemente, com maior poder discriminatório neste caso.

Tabela 10. Valores da estatística *F* obtidos a partir da ANOVA para os grupos formados pelo método de agrupamento hierárquico de Ward

		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
	QMG	1.177,27	0,025	0,314	0,489	924,40	244,10	54.798,62	5.844,00	2.001,10
S	QMRes	114,54	0,003	0,017	0,018	39,20	65,40	192,29	101,00	49,80
	F	10,28	7,39	18,71	27,29	23,58	3,73	284,98	57,86	40,18
	QMG	1.538,60	0,064	0,859	0,990	1.287,80	1.706,10	21.543,75	7.408,00	1.950,80
R	QMRes	108,76	0,003	0,008	0,010	33,40	42,00	724,37	75,80	50,60
	F	14,15	22,73	105,99	100,56	38,56	40,62	29,74	97,73	38,55
	QMG	2.998,26	0,064	0,047	0,067	1.47,70	465,20	39.610,80	8.787,40	3.138,60
S*	QMRes	85,40	0,003	0,014	0,015	30,50	61,90	435,30	53,70	31,60
	F	35,11	22,42	32,70	45,43	48,26	7,52	91,00	163,69	99,23

De forma geral, pode-se observar que quando fez-se uso da matriz *S* para a obtenção dos CPs, as variáveis com maiores variâncias foram favorecidas. Quando os CPs foram obtidos através da matriz *R*, percebeu-se que o problema da escala foi solucionado. Entretanto, variáveis que foram de pouco poder discriminatório foram favorecidas. Isto implicou na utilização de pelo menos um CP influenciado por pelo menos uma variável de alto peso, mas com baixo poder de discriminação. A obtenção dos CPs a partir da matriz S^* favoreceu as variáveis que apresentaram maior variabilidade relativa (Tabela 1), resultado este que foi mais interessante como substituição às outras matrizes descritas anteriormente.

3.3. Análise de Regressão Múltipla

As variáveis econômicas mais representativas da variabilidade total dos dados deverão ser as mais importantes nos CPs selecionados, independentemente da matriz utilizada. Portanto, elas deverão ter uma relação de causa e efeito com as variáveis independentes (zootécnicas) e não simplesmente terem variâncias aleatórias.

Percebeu-se que as variáveis que apresentaram melhores ajustes foram Y_1 e Y_7, conforme explicado pelas maiores estimativas dos coeficientes de determinação (R^2), dadas respectivamente por 0,99 e 0,53 (Tabela 11).

Tabela 11. Constantes e coeficientes de regressão dos modelos ajustados para cada variável econômica (*Y*) em função das variáveis zootécnicas (*Xs*)

Variável	Estimativas dos coeficientes significativos										R^2
	Const.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	
Y_1	327,600	341,100	6,000	17,700	-	-	-	-	-	-	**0,99**
Y_2	0,869	-	0,017	-	0,014	-	-	0,015	0,008	-	0,32
Y_3	0,609	-	-0,031	0,063	-	-	-	-	-	-	0,12
Y_4	0,758	-	-0,020	0,043	-	-	-	-	-0,033	-	0,09
Y_5	11,400	-	-	1,650	-0,770	-	-	-1,570	-2,740	-	0,33
Y_6	32,670	-	-	-	-	-	-	1,560	-	-	0,04
Y_7	915,800	-	-	-166	-	-	-238	-	130	932	**0,53**
Y_8	16,990	-1,250	-	-	-	3,200	-	-	4,880	-	0,21
Y_9	8,200	-	-	-	-	1,920	1,030	-	2,120	-	0,18

Const.: Constante.

Dentre todas as três matrizes analisadas neste trabalho em relação à obtenção dos CPs, observou-se que aquela baseada na ponderação da variância pelo respectivo coeficiente de variação forneceu resultados mais concordantes com o critério da regressão. De acordo com a matriz S^*, as variáveis mais importantes foram: Y_1, Y_7 e Y_8.

Nos casos onde a maior parte da variabilidade foi devida às causas aleatórias, principalmente para as variáveis Y_3, Y_4 e Y_6 com os menores valores de R^2 e, consequentemente, pouco úteis para a discriminação dos produtores, foram consideradas, erroneamente, como importantes pela matriz *R*.

Dentro dessa estrutura de pouca relação de causa e efeito entre as variáveis econômicas e zootécnicas, para apenas duas variáveis o valor do coeficiente de determinação foi maior que 0,5 (Tabela 11), foi mais coerente trabalhar com os componentes que foram interpretados pelas variáveis econômicas mais relacionadas às zootécnicas, ou seja, pelos CPs obtidos a partir da matriz S^*.

4. CONCLUSÕES

As variáveis econômicas com maiores coeficientes de variação e, consequentemente, com maiores variabilidades não inerentes à escala, que são Y_1 - renda bruta da atividade leiteira (mil R$/ano) e Y_7 - margem bruta/ área (R$/ha), foram as mais relacionadas com as variáveis zootécnicas. Além disso, as variáveis econômicas, consideradas mais importantes do ponto de vista técnico Y_5 - gasto com mão-de-obra/ renda bruta do leite (%), Y_8 - taxa de remuneração do capital sem terra (% a.a.) e Y_9 - taxa de remuneração do capital com terra (% a.a.), foram aquelas que formaram grupos mais homegêneos quando se utilizou a matriz S^*. Isto fez com que a estrutura de dados utilizada na obtenção dos CPs via matriz S^* fosse considerada a mais indicada na discriminação dos produtores.

REFERÊNCIAS BIBLIOGRÁFICAS

ALEIXO, S.S.; SOUZA, J.G.; FERRAUDO, A.S. Técnicas de análise multivariada na determinação de grupos homogêneos de produtores de leite. **R. Bras. Zootec**. v. 36, p. 2168-2175, 2007.

CATTELL, R.B. The screen test for number of factors. **Multivarite Behavioral Research**, v.1, p.140-161, 1966.

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

EFROYMSON, M.A. “Multiple regression analysis”, in A. Ralston and H.S. Wilf (Eds.), **Mathematical Methods for Digital Computers**, New York, Wiley, 1960.

FERNANDES, E.N.; BRESSA, N.M.; VERNEQUE, R.S. Zoneamento da pecuária leiteira da região sul do Brasil. **Ciência Rural**, v.34, p.485-491, 2004.

GOMES, S.T. **Economia da Produção do Leite**. Belo Horizonte: Itambé, (2000).

HAIR, J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W. **Análise multivariada de dados.** Porto Alegre, Bookman, 2005, 600p.

MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte, Editora UFMG, 2007, 295p.

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

MONTGOMERY, D.C. **Design and analysis of experiments.** John Wiley & Sons, 1997. 704 p.

MONTGOMERY, D.C.; PECK, E.A. **Introduction to linear regression analysis.** New York: John Wiley & Sons, 1992. 704 p.

NETO, A.C.; CASTRO, G.P.C.; LIMA, J.E. Uso de análise estatística multivariada para tipificação de produtores de leite de Minas Gerais. **Organ. rurais agroind.**, v. 7, p.114-121, 2005.

REIS, E. **Estatística Multivariada Aplicada.** Lisboa: Edições Silabo, 1997. 343p.