Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Métodos multivariados para agrupamento de bovinos da raça Hereford em função dos parâmetros de curvas de crescimento

Luiz Ricardo Nakamura

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2011

Luiz Ricardo Nakamura
Estatístico

Métodos multivariados para agrupamento de bovinos da raça Hereford em função dos parâmetros de curvas de crescimento

Orientador:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2011

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - ESALQ/USP

Nakamura, Luiz Ricardo
Métodos multivariados para agrupamento de bovinos da raça Hereford em função dos parâmetros de curvas de crescimento / Luiz Ricardo Nakamura. - - Piracicaba, 2011.
95 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2011.

1. Análise de conglomerados 2. Análise multivariada 3. Bovinos de corte 4. Componentes principais 5. Curvas de crescimento - Parâmetros 6. Gado Hereford 7. Modelos não lineares I. Título

CDD 636.222
N163m

Dedicatória

Aos meus pais, **Edmilson** e **Neide Nakamura**,
pelo incondicional apoio em todas as etapas de minha vida.

AGRADECIMENTOS

Aos meus pais, Edmilson e Neide Nakamura, pelo imensurável apoio, afeto e dedicação em todos esses anos de caminhada. Com certeza não estaria aqui, dando mais este passo em minha vida, sem minha base, que são vocês.

Ao meu irmão, Cézar Augusto Nakamura, que apesar de todas as brigas e desavenças sempre esteve e continua ao meu lado.

À minha grande amiga, Ana Julia Righetto, cujo incentivo foi crucial nesses últimos anos de minha vida, especialmente em meu último ano de graduação. Pessoa de grande coração que aguentou todas as minhas crises de estupidez e mau humor, crises estas que pioraram na fase final desta dissertação.

Aos amigos que a ESALQ me proporcionou, com os quais tive a oportunidade de vivenciar essa nova etapa de vida. Em especial para Daniel, Edilan, Fabiane, Ítalo, João Vitor, Kuang, Mariana, Patrícia, Pedro, Rodrigo, Simone Sartório e Verónika.

Ao Prof. Dr. Antônio Assiz de Carvalho Filho, meu antigo orientador de graduação, que se tornou um grande amigo, sempre com palavras de conforto nas situações mais adversas possíveis, compartilhando sua sabedoria servindo de orientação para todos os profissionais que um dia foram seus alunos. Obrigado por acreditar em mim.

Ao Sr. Eduardo Cardoso de Oliveira, companheiro do grupo de pesquisa "Agronegócio, políticas públicas e desenvolvimento regional sustentável - APP&DR", por toda ajuda concedida ao longo de minha, ainda pequena, carreira acadêmica.

Ao meu orientador, Prof. Dr. Carlos Tadeu dos Santos Dias, pelo apoio concedido em todas as etapas de minha dissertação, pelos puxões de orelha nas horas corretas e, acima de tudo, por ter me aberto inúmeras portas, possibilitando assim um grande crescimento pessoal e profissional.

À Prof. Dra. Taciana Villela Savian pelo imensurável tempo empreendido na realização deste trabalho. Sua ajuda e paciência foram cruciais para que esta dissertação se concretizasse. Agradeço também pelas inúmeras horas de conversa e pelos vários cafézinhos.

Ao Prof. Dr. Luiz Roberto Martins Pinto pela importante colaboração neste trabalho e pelas horas de conversa sobre os mais diversos assuntos.

À todos os professores do Departamento de Ciências Exatas da ESALQ: Prof. Dra. Clarice Garcia Borges Demétrio, Prof. Dr. Edwin Moises Marcos Ortega, Prof. Dr. Gabriel Adrián Sarriés, Prof. Dra. Roseli Aparecida Leandro, Prof. Dr. Silvio Sandoval Zocchi e Prof. Dra. Sonia Maria De Stefano Piedade; por todo apoio, ensinamento, sabedoria e dedicação nesses dois anos.

À equipe de apoio do Departamento de Ciências Exatas da ESALQ: Eduardo Bonilha, Jorge Alexandre Wiendl, Luciane Brajão, Rosni Honofre Aparecido Pinto e Solange de Assis Paes Sabadin, por serem sempre tão solícitos nas mais diversas situações.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pelo apoio financeiro em forma de bolsa de estudos.

SUMÁRIO

RESUMO

Métodos multivariados para agrupamento de bovinos da raça Hereford em função dos parâmetros de curvas de crescimento

Após o ajuste individual das 55 vacas estudadas pelo modelo Gompertz difásico com estrutura de erros autorregressiva de ordem 1 (totalizando 7 parâmetros), notou-se que apenas 6 vacas tinham problemas nas estimativas de seus parâmetros (não convergentes ou não significativos), dessa forma continuou-se o trabalho proposto com 49 animais. Com as estimativas de cada um dos parâmetros (variáveis nessa etapa) foi realizada a análise de componentes principais e observação do gráfico biplot, sendo possível a constatação de que 2 dos parâmetros do modelo continham informações ambíguas com pelo menos um dos demais parâmetros e estes foram retirados da análise, restando 5 parâmetros para o estudo. A análise de componentes principais foi realizada novamente apenas com os 5 parâmetros restantes e os três primeiros componentes principais (escolhidos pelo critério da percentagem de variância original explicada) foram utilizados como variáveis em um processo de agrupamento hierárquico. Após a realização da análise de agrupamentos, observou-se que 5 grupos homogêneos de animais foram formados, cada um com caraterísticas distintas. Desta forma, foi possível identificar animais que se destacavam, positiva ou negativamente, no que tange ao seu peso assintótico e taxa de crescimento.

Palavras-chave: Modelo não-linear; Análise de componentes principais; Análise biplot; Análise de agrupamentos

ABSTRACT

Multivariate methods for grouping Hereford cattle breed against the parameters of growth curves

After individual adjustment of the 55 cows studied using the diphasic Gompertz model with autoregressive structure of errors (totalizing 7 parameters), it was noted that only 6 cows had problems on estimates of the parameters (not converged or not significant), then the proposed work continued with 49 animals. With each of the parameters estimates (variables at this stage) was performed a principal component analysis and observation of the biplot, and it was possible to find that two of the model parameters contained ambiguous information with at least one of the other parameters, then these 2 parameters were removed from the analysis, leaving 5 parameters for the study. The principal component analysis was performed again with only five remaining parameters and the first three principal components (chosen by the criterion of percentage of original explained variance) were used as variables in a process of hierarchical clustering. After performing the cluster analysis, we found that five homogeneous groups of animals were formed, each with distinct characteristics. Thus, it was possible to identify animals that stood out, positively or negatively, in terms of their asymptotic weight and growth rate.

Keywords: Nonlinear model; Principal component analysis; Biplot analysis; Cluster analysis

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Uma forma consistente de avaliar o peso de animais medidos repetidamente em intervalos pré-definidos de tempo é estudando suas curvas de crescimento. Para tal, diferentes modelos de regressão não-linear, como os modelos de Brody, Von Bertalanffy, Logístico, Gompertz, Richards, entre outros, vêm mostrando, em uma vasta literatura, boa adequação no que tange a descrição destas curvas, pois apresentam parâmetros biologicamente interpretáveis, fornecendo assim, importantes informações sobre variações genética e ambiental, que ocorrem entre as avaliações consecutivas do animal.

Geralmente, em estudos de curvas de crescimento, o pesquisador tem como principal interesse comparar os parâmetros das diferentes curvas, de modo que seja possível distinguir os processos de crescimento mais eficientes, ou seja, identificar animais com maior velocidade de crescimento e/ou maior ganho de peso em uma determinada fase da vida. Normalmente, os parâmetros são comparados através do ajuste individual para cada um dos animais e, posteriormente, ajusta-se uma curva com os dados médios, perdendo-se assim muita informação disponível, resultando, possivelmente, na subestimação ou superestimação dos pesos assintóticos e taxas de crescimentos de diversos animais.

Um problema que ocorre no estudo destas curvas é a heterogeneidade de variâncias, ou heterocedasticidade (GUEDES et al., 2004), pois à medida que a idade aumenta, a variância dos pesos corporais também aumenta. Uma solução para este caso é a utilização do inverso da variância como ponderador, o que resulta em melhores ajustes dos modelos.

Como as medidas são retiradas de um mesmo animal ao longo do tempo, ou seja, medidas repetidas, existe problema de autocorrelação nos dados. Ou seja, a medida retirada em um determinado tempo t depende da medida mensurada em $(t-1)$. Logo, a modelagem do erro, no caso a modelagem do erro em curvas de crescimento de bovinos, deve ser realizada levando-se em consideração os erros com estrutura autorregressiva.

Em determinadas situações, o crescimento animal apresenta um comportamento cíclico e é necessária a utilização de modelos de crescimento multifásico, que basicamente condensam em um único modelo, dois ou mais submodelos que explicam o comportamento dos dados em cada uma de suas fases, ou seja, este tipo de modelo possui parâmetros específicos (distintos) para cada uma das fases dos dados. Este comportamento cíclico é observado quando

se tem efeito sazonal, normalmente observado em decorrência das secas. A ausência de chuva restringe o crescimento do animal por um determinado período (primeira assíntota) e, após a mudança de estação, o determinado animal volta a crescer normalmente, posteriormente atingindo a segunda assíntota.

Os modelos multifásicos são de grande utilidade na modelagem do crescimento de bovinos de corte criados no sul do país (os bovinos neste estudo são provenientes de Bagé - RS), região na qual o inverno rigoroso proporciona estacionaridade do crescimento, fazendo com que a curva apresente fases significantemente distintas. Dentre as raças de gado de corte desta região, pode-se destacar a raça Hereford, originária da Inglaterra, que possui um esqueleto extremamente forte, boa massa muscular e chegam ao peso ideal de abate entre 20 e 26 meses quando são mantidos em boas condições alimentares, além de apresentarem boa adaptação aos mais diversos sistemas de produção.

Mendes (2007) modelou as curvas de crescimento de bovinos (fêmeas) da raça Hereford, considerando erros autorregressivos e heterogeneidade de variâncias, o que de fato resultou em ótimos ajustes individuais. A proposta deste trabalho foi a de reproduzir os resultados obtidos com o modelo melhor ajustado segundo a autora (modelo Gompertz difásico ponderado pelo inverso da variância dos pesos com estrutura de erros autorregressiva) e, posteriormente, agrupar os bovinos mais similares, de modo a caracterizar grupos de animais homogêneos, identificando aqueles que se destacam positiva e negativamente em relação ao seu crescimento com o auxílio do método de estatística multivariada análise de agrupamentos.

Os parâmetros de cada curva de crescimento, estimados individualmente em cada uma das 55 fêmeas em estudo serviram, após a aplicação da análise de componentes principais, como variáveis na aplicação da análise de agrupamentos, sendo possível assim, agrupar as vacas da raça Hereford de forma homogênea separando grupos compostos por vacas com maior ou menor taxa de crescimento e/ou peso assintótico.

2 REVISÃO BIBLIOGRÁFICA

Para o desenvolvimento deste trabalho foi realizada uma revisão sobre modelos de regressão não-linear, análise de componentes principais, análise biplot e análise de agrupamentos.

2.1 Raça Hereford

Segundo a Associação Brasileira de Hereford e Braford (2011), a raça Hereford é originária do condado inglês de Herefordshire, zona propicia a produção de pasto superior devido a suas condições climáticas, foi introduzida no Brasil no início do século XX e, atualmente, esta raça se encontra difundida em diversos países como Canadá, Austrália, Nova Zelândia, Argentina, Uruguai, entre outros; devido à sua alta adaptabilidade aos mais diversos tipos de ambiente.

Seguindo a mesma referência pode-se dizer que os animais da raça em estudo (Figura 1) caracterizam-se por uma pelagem vermelha com cara, ventre, extremidades da cauda e partes inferiores das patas totalmente brancas (esta pelagem recebe o nome de "pampa"). Ainda, possuem um esqueleto extremamente forte, boa massa muscular e chegam ao peso ideal de abate entre 20 e 26 meses quando são mantidos em boas condições alimentares, além de apresentarem boa adaptação aos mais diversos sistemas de produção. Esta raça tem como aptidão principal a produção de carne, sendo bastante eficiente em regime de pasto.

Figura 1 - Bovinos da raça Hereford. Fonte: Associação Brasileira de Hereford e Braford (2011)

A Associação Brasileira de Hereford e Braford(2011) afirma que a raça Hereford é a mais abundante em diversas regiões do mundo, pois combina performance, praticidade e lucratividade como características principais, além de possuir as características básicas de corte: fertilidade, rusticidade, eficiência alimentar, longevidade e, sobre tudo, adaptabilidade aos mais diversos locais. Devido a essas características de grande interesse, a mesma referência assegura que essa raça continuará a desempenhar papel de destaque na indústria de carne bovina brasileira.

2.2 Curvas de Crescimento

A evolução do peso vivo dos animais com a idade, em condições de crescimento contínuo, segue um padrão comum nas espécies e sua representação gráfica origina uma curva de forma sigmoide, designada curva de crescimento, que modela o padrão de resposta de dados peso-idade ao longo da vida do animal (SILVA et al., 2011).

Algumas das aplicações das curvas de crescimento na produção animal que Freitas (2005) cita são: i) resumir nos parâmetros de um determinado modelo as características de crescimento; ii) avaliar o perfil do animal ao longo do tempo com referência ao seu peso; iii) identificar animais com uma taxa de crescimento superior aos demais; iv) identificar animais com alto peso assintótico. Estas aplicações são de grande interesse no que tange ao estudo de curvas de crescimento animal e, normalmente, este estudo é conduzido por meio de uma abordagem frequentista, ajustando-se modelos não-lineares que sintetizam informações biológicas relevantes nos parâmetros do modelo (PAZ et al., 2004).

O ajuste de dados de peso-idade de cada animal ou de um grupo de animais permite obter informações descritivas da curva de crescimento e de prognósticos futuros para animais do mesmo grupo racial sob a mesma situação ambiental. Portanto, a função que é utilizada para descrever o crescimento do animal tanto pra fins de exigência nutricional, como para seleção genética, é de extrema importância (FITZHUGH JR., 1976).

As funções não-lineares com parâmetros biologicamente interpretáveis, desenvolvidas empiricamente para relacionar peso e idade, são úteis em estudos de crescimento, pois fornecem informações importantes sobre as variações genética e ambiental que ocorrem entre as avaliações consecutivas dos animais (LAIRD; HOWARD, 1967). Alguns exemplos de modelos

não-lineares que são frequentemente utilizadas na descrição de curvas de crescimento animal são o modelo logístico (NELDER, 1961), o modelo Gompertz (WINSOR, 1932), o modelo de von Bertalanffy (von BERTALANFFY, 1957), entre outras.

Nos diversos estudos de curvas de crescimento animal, observa-se que, por se tratarem de medidas repetidas, ou seja, medidas em um mesmo animal ao longo do tempo, os erros não são independentes e existe heterogeneidade de variâncias. Exemplos como Mendes et al. (2009), Mazzini et al. (2005) e outros, afirmam que, ao se levar em consideração essas duas informações, o ajuste de um determinado modelo não-linear é mais fiel aos dados.

Em determinadas situações, este crescimento apresenta um comportamento cíclico e é necessária a utilização de modelos de crescimento multifásico, que condensam em um único modelo, dois ou mais submodelos que explicam o comportamento dos dados em cada uma de suas fases, ou seja, este tipo de modelo possui parâmetros específicos (distintos) para cada uma das fases dos dados. Basicamente, um modelo multifásico busca combinar duas funções, no caso, não lineares para a melhor adequação aos dados. Koops(1986) mostrou o bom ajuste desses modelos nas mais diferentes espécies.

Este comportamento cíclico é observado quando se tem efeito sazonal, normalmente observado em regiões onde o frio é rigoroso ou as estações do ano são bem diferenciadas, como por exemplo, na região sul do Brasil. Durante o rigoroso inverno o alimento é mais escasso, resultando em uma estacionariedade no peso dos animais neste período (primeira assíntota) e, ao fim desta estação, o animal volta a engordar até alcançar seu peso máximo (segunda assíntota).

2.3 Modelo Não-Linear

A análise de regressão é uma das técnicas mais utilizadas na atualidade pois ela permite através da utilização de funções (equações), expressar a relação entre a variável de interesse (variável dependente ou resposta) Y com as variáveis ditas independentes, ou explicativas, $X_1, X_2, ..., X_n$ (MONTGOMERY et al., 2006). Segundo Draper e Smith (1998), os modelos de regressão podem ser subdivididos em dois grupos distintos: os lineares e os não-lineares. No caso linear, a partir de um conjunto de observações, busca-se um modelo que melhor explique a relação entre as variáveis inerentes a um dado fenômeno, ao passo que os

modelos não-lineares, na maioria das vezes, são baseados em considerações teóricas inerentes ao fenômeno que se tem interesse em estudar.

Na prática, quando é possível, um modelo não-linear é linearizado para facilitar a obtenção das estimativas dos parâmetros (BATES; WATTS, 1988). O inconveniente de uma transformação é que, além de o parâmetro perder sua interpretação intrínseca, pode-se alterar a estrutura e a distribuição do erro. Ou seja, se os erros do modelo original satisfazem às suposições usuais de normalidade, independência e homogeneidade da variância, os erros do novo modelo, em geral, não satisfarão tais suposições porque tais transformações podem mudar a forma da distribuição dos mesmos ou fazer com que a variância desta distribuição deixe de ser constante (CURRIE, 1982).

Ainda, segundo Draper e Smith (1998), pode-se classificar os modelos de regressão como:

1) Modelos Lineares: aqueles que são lineares em relação aos parâmetros.

$$\frac{\partial}{\partial \theta_j} f_i(X, \theta) = g(X),$$

para $i = 1, 2, ..., n$ e $j = 1, 2, ..., p$; em que n é o número total de observações e p é o número de parâmetros do modelo;

2) Modelos Linearizáveis: aqueles que podem ser transformados em lineares através de algum tipo de transformação. Tem-se o modelo:

$$Y = \theta^x.\epsilon,$$

em que o erro ϵ é dito multiplicativo. Aplicando-se o logaritmo natural à igualdade, tem-se:

$$\begin{aligned} \ln Y &= \ln(\theta^X.\epsilon) \\ &= \ln(\theta^X) + \ln(\epsilon) \\ &= X\ln(\theta) + \ln(\epsilon). \end{aligned}$$

Seja $Z = \ln(Y)$; $\lambda = \ln(\theta)$; e $\epsilon^* = \ln(\epsilon)$:

$$Z = \lambda X + \epsilon^*,$$

que é linear, pois

$$\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = X = g(X),$$

logo o modelo $Y = \theta^X.\epsilon$ é dito linearizável.

3) Modelos Não-Lineares: aqueles que não se enquadram nos dois primeiros casos. Considere o modelo:

$$Y = \theta^X + \epsilon,$$

em que o erro é dito aditivo e não existe transformação capaz de tornar o modelo linear. Verifica-se que:

$$\frac{\mathrm{d}Y}{\mathrm{d}\theta} = X\theta^{X-1} = g(X, \theta),$$

portanto o modelo é não-linear.

Formalmente, um modelo é dito não-linear quando uma ou mais derivadas parciais da variável dependente com relação a algum parâmetro presente no modelo, depende de algum parâmetro, ou seja, pode-se dizer que um modelo é não-linear quando ele não é linear nos parâmetros (DRAPER; SMITH, 1998).

Seja o modelo não-linear escrito como:

$$y_i = f(x_i, \boldsymbol{\theta}) + \epsilon_i \ , \ i = 1, \cdots, n, \tag{1}$$

em que y_i é a observação da variável dependente, $f(x_i, \boldsymbol{\theta})$ é a função resposta (ou esperança) não-linear em $\boldsymbol{\theta}$ e com forma funcional conhecida, x_i representa a variável independente, $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_p \end{bmatrix}$ é um vetor de parâmetros p-dimensional e ϵ_i é um erro experimental não observável diretamente, suposto independente com média 0 e variância σ^2 desconhecida.

É importante salientar que a estrutura dos erros envolvidos nos modelos nem sempre é considerada. A presença de heterogeneidade de variância (heterocedasticidade) dos pesos decorrentes do aumento da idade e/ou a autocorrelação entre os resíduos, tendo em vista que os dados são tomados longitudinalmente em cada animal, podem resultar em estimativas viesadas (PASTERNAK; SHALEV, 1994) e na subestimação das variâncias dos parâmetros (SOUZA, 1998). Como possível solução a estes problemas, considera-se a ponderação pelo inverso das variâncias em cada peso e a modelagem da correlação residual por meio de processos auto-regressivos.

2.3.1 Métodos de estimação dos parâmetros

Alguns métodos são propostos na literatura para a estimação dos parâmetros de um modelo não-linear e, como pode ser visto em Gallant (1987), o método dos mínimos quadrados pode ser utilizado assim como na estimação dos parâmetros em modelos lineares. O problema é que a solução do sistema de equações normais (SEN) não pode ser obtida de forma usual (RAWLINGS et al., 1998), ou seja, não possuem soluções explícitas e, sendo assim, é necessário algum tipo de artifício para estimar os parâmetros de interesse. A maioria dos algoritmos computacionais para as estimativas de mínimos quadrados $\hat{\theta}$ e a maioria dos métodos inferenciais para modelos não-lineares são baseados em métodos iterativos que consideram uma aproximação linear local para o modelo (BATES; WATTS, 1980). Dois dos processos mais citados e utilizados na literatura são: o método de Gauss-Newton/Gauss-Newton modificado (HARTLEY, 1961) e o método de Marquardt (MARQUARDT, 1963), que utilizam as derivadas parciais da função esperança $f(x_i, \theta)$ com relação a cada um dos parâmetros. Esses artifícios serão abordados, posteriormente, na seção de metodologia.

2.3.1.1 Método dos mínimos quadrados ordinários

Para a realização da estimação dos parâmetros através deste método, Souza (1998) reescreve o modelo não-linear da seguinte maneira:

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{\epsilon},$$

em que $\boldsymbol{Y}$ é um vetor composto por y_i, $\boldsymbol{f}(\boldsymbol{\theta})$ é um vetor composto por $f(x_i, \theta)$ e $\boldsymbol{\epsilon}$ é um vetor composto por ϵ_i.

A obtenção dos estimadores é realizada através da minimização da soma de quadrado dos erros aleatórios (SQE), logo a função de mínimos quadrados (soma de quadrados dos erros) para um modelo não-linear é dada por (DRAPER; SMITH, 1998):

$$\begin{aligned} SQE(\boldsymbol{\theta}) &= \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2 \\ &= (\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\theta}))^T (\boldsymbol{Y} - \boldsymbol{f}(\boldsymbol{\theta})), \end{aligned} \tag{2}$$

Para encontrar o estimador de mínimos quadrados $\hat{\boldsymbol{\theta}}$ é necessário diferenciar a

eq. (2) com relação a cada um dos parâmetros, ou seja,

$$\frac{\partial SQE(\theta)}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\boldsymbol{Y} - f(\boldsymbol{\theta})\right]' \left[\boldsymbol{Y} - f(\boldsymbol{\theta})\right].$$

Fazendo $\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{\theta}) = F(\boldsymbol{\theta})$, tem-se

$$\frac{\partial SQE(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2 \left[\boldsymbol{Y} - f(\boldsymbol{\theta})\right]' F(\boldsymbol{\theta}).$$

Igualando a equação acima à zero, tem-se:

$$\left[\boldsymbol{Y} - f(\hat{\boldsymbol{\theta}})\right]' F(\hat{\boldsymbol{\theta}}) = 0,$$

que é o sistema de equações normais não-linear. Neste ponto tem-se o problema da solução não trivial visto na seção 2.3.1 e o uso de algum algoritmo computacional para a resolução do SEN deve ser utilizado.

2.3.1.2 Método dos mínimos quadrados ponderados

Quando a pressuposição de homocedasticidade é violada, Hoffman e Vieira (1998) afirmam que o método dos mínimos quadrados ponderados é o mais adequado para a estimação dos parâmetros do modelo de regressão, pois fornece estimadores não tendenciosos e de mínima variância. Para explicar o procedimento realizado por este método, utilizar-se-á aqui, o modelo de regressão linear, dado por

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

em que $\boldsymbol{Y}$ é o vetor da variável resposta de dimensão $n\times 1$, $\boldsymbol{X}$ é a matriz de dimensão $n\times p$ que representa as variáveis independentes , $\boldsymbol{\beta}$ é o vetor de parâmetros de dimensão $p\times 1$ e $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{V}\sigma^2)$ é o vetor de erros de dimensão $n\times 1$, sendo que $\boldsymbol{V}$ é uma matriz diagonal positiva definida, que representa as variâncias associadas a cada um dos erros ϵ_i. Ainda, tem-se que

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{V}\sigma^2 = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_n \end{bmatrix} \sigma^2,$$

ou seja, a pressuposição de independência das observações é válida, uma vez que $E(\epsilon_i \epsilon_j) = 0$, para $i \neq j$.

Para realizar o procedimento do método, define-se uma matriz $\mathbf{\Lambda}$, diagonal, dada por

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \sigma^2,$$

em que $\lambda_j = \frac{1}{\sqrt{V_j}}$. Tem-se ainda que

$$\mathbf{\Lambda\Lambda} = \boldsymbol{V}^{-1} \quad \text{e} \quad \boldsymbol{V} = \mathbf{\Lambda}^{-1}\mathbf{\Lambda}^{-1}.$$

Pré-multiplicando a equação 3 pela matriz $\mathbf{\Lambda}$, obtém-se o seguinte modelo:

$$\mathbf{\Lambda}\boldsymbol{Y} = \mathbf{\Lambda}\boldsymbol{X}\boldsymbol{\beta} + \mathbf{\Lambda}\boldsymbol{\epsilon}.$$

Seja $\boldsymbol{\xi} = \mathbf{\Lambda}\boldsymbol{\epsilon}$, então $E(\boldsymbol{\xi}) = \mathbf{0}$ e

$$\begin{aligned} E(\boldsymbol{\xi}\boldsymbol{\xi}') &= E(\mathbf{\Lambda}\boldsymbol{\epsilon}\boldsymbol{\epsilon}\mathbf{\Lambda}) = \mathbf{\Lambda}\boldsymbol{V}\mathbf{\Lambda}\sigma^2 \\ &= \mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\sigma^2 = \boldsymbol{I}\sigma^2. \end{aligned}$$

Portanto, o modelo $\mathbf{\Lambda}\boldsymbol{Y} = \mathbf{\Lambda}\boldsymbol{X}\boldsymbol{\beta} + \mathbf{\Lambda}\boldsymbol{\epsilon}$ é homocedástico. Logo, o SEN fornecido pelo método dos mínimos quadrados ponderados é dado por:

$$\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$

e o estimador de mínimos quadrados ponderados pelo inverso da matriz de variâncias do vetor de parâmetros $\boldsymbol{\beta}$ é dado por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}.$$

2.3.1.3 Método dos mínimos quadrados generalizados

Quando há a constatação de heterogeneidade de variâncias e autocorrelação dos resíduos, isto é, o resíduo ϵ_i depende do resíduo ϵ_{i-1}, Hoffman e Vieira (1998) sugerem que o

método dos mínimos quadrados generalizados seja o utilizado na estimação dos parâmetros no modelo de regressão.

Neste método, basicamente tem-se que a matriz de variâncias dos resíduos é da forma $\boldsymbol{\Omega}\sigma^2$, com $\boldsymbol{\Omega} \neq \boldsymbol{I}$. Suponha o modelo linear

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}\sigma^2)$, sendo $\boldsymbol{\Omega}$ uma matriz simétrica positiva definida representando as variâncias e covariâncias dos erros. Se os erros são autocorrelacionados na forma de um processo autoregressivo de primeira ordem, ou seja AR(1), tem-se que (MORETTIN; TOLOI, 2004)

$$\epsilon_i = \phi_1 \epsilon_{i-1} + \xi_i,$$

em que ϵ_i é o resíduo no tempo i, ϕ_1 é o parâmetro autorregressivo de ordem 1, ϵ_{i-1} é o resíduo referente ao tempo $i-1$ e ξ_i é o ruído branco. Ainda,

$$E(\xi_i) = 0 \quad , \quad E(\xi_i^2) = \sigma_\xi^2 \quad \text{e} \quad E(\xi_i, \xi_{i-h}) = 0,$$

se $h \neq 0$.

A estimativa de $\boldsymbol{\beta}$ é encontrada da mesma forma como demonstrado no método dos mínimos quadrados ponderados, ou seja, $\hat{\boldsymbol{\beta}}$ é dado por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{Y}, \tag{4}$$

em que

$$\boldsymbol{\Omega} = \frac{\sigma_\xi^2}{1-\phi_1^2} \begin{bmatrix} 1 & \phi_1 & \phi_1^2 & \cdots & \phi_1^{n-1} \\ \phi_1 & 1 & \phi_1 & \cdots & \phi_1^{n-2} \\ \phi_1^2 & \phi_1 & 1 & \cdots & \phi_1^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_1^{n-1} & \phi_1^{n-2} & \phi_1^{n-3} & \cdots & 1 \end{bmatrix}.$$

Observa-se que para o caso em que $\phi = 0$ retorna-se ao caso de mínimos quadrados ordinários.

Como já mencionado na seção 2.3.1, a solução através do método dos mínimos quadrados não é obtida de forma trivial quando um modelo de regressão não-linear é considerado. Isso se deve ao fato de que a solução do SEN não-linear não pode ser obtida de forma usual, ou seja, não existe uma solução explícita e, sendo assim, é necessário a utilização de métodos iterativos na estimação dos parâmetros. Esses métodos iterativos são iniciados utilizando-se valores iniciais que provém de informações conhecidas do pesquisador à priori.

2.3.1.4 Valores iniciais

Os valores iniciais para $\boldsymbol{\theta}$ podem ser estimativas preliminares baseadas em alguma informação disponível. Por exemplo, a informação de um especialista em determinado assunto pode ser crucial na escolha do valor inicial a ser utilizado em determinado problema, de forma que necessite-se poucas iterações para estimar o vetor de parâmetros. Gallant (1987) afirma que a rapidez na convergência depende da estrutura do modelo em estudo e na qualidade dos valores iniciais utilizados.

Alguns métodos para a escolha dos valores iniciais são elucidados em Draper e Smith (1998) e Gallant (1987):

1) Interpretar o comportamento da função esperança em termos de parâmetros analiticamente ou graficamente;

2) Interpretar o comportamento das derivadas da função esperança em termos dos parâmetros analiticamente ou graficamente;

3) Transformar a função esperança analiticamente ou graficamente para obter comportamentos mais simples;

4) Reduzir dimensões substituindo valores para alguns parâmetros ou avaliando a função nos valores do delineamento específico;

5) Usar linearidade condicional.

Ainda, Gallant (1987) afirma que a falha no processo de convergência depende da distância do valor inicial à resposta correta e do grau de parametrização da função res-

posta relativamente ao conjunto de dados utilizados. Quando o método não converge, deve-se procurar valores iniciais melhores ou uma função resposta mais parcimoniosa.

Draper e Smith (1998) chamam atenção para para algumas problemáticas que podem ocorrer com o processo de convergência dos parâmetros:

1) A convergência ser atingida de forma muito lenta, ou seja, para atingí-la necessita-se um grande número de iterações;

2) No processo podem ocorrer oscilações, mudando a direção, que ocasionam o decrescimento da soma de quadrados $S(\boldsymbol{\theta})$;

3) O procedimento pode não convergir.

Desta forma, pode-se afirmar que uma escolha sensata dos valores iniciais para cada um dos parâmetros no método numérico utilizado, é de vital importância para se obter estimativas que convirjam para valores plausíveis e, no caso dos modelos não-lineares, interpretáveis.

2.4 Análise de Componentes Principais

A análise de componentes principais tem o poder de construir novas variáveis aleatórias (componentes principais) que são combinações lineares das variáveis originais e explicam a variância de um conjunto de p-variáveis. A variância original das variáveis em estudo é explicada pela construção de p componentes principais, todos independentes entre si. Este método vem sendo amplamente utilizado devido a facilidade de sua aplicação e por ser um método não-paramétrico capaz de extrair informações relevantes de um determinado conjunto de dados (JOHNSON; WICHERN, 2007).

O objetivo principal deste método é o de reduzir a dimensão de variáveis, sendo assim, ao invés de utilizar os p componentes principais possíveis de serem construídos, utiliza-se um número k de componentes, em que $k < p$, de forma a se explicar uma proporção de variância satisfatória para o problema em estudo. Esta proporção explicada dependerá do número de componentes mantidos após um dos critérios para a determinação do número k de componentes principais, que serão citados ainda neste texto.

A análise de componentes principais é usualmente confundida com a análise fatorial (BASILEVSKY, 1994), pois ambas têm a pretensão de explicar um determinado conjunto de dados em um número menor de dimensões. Como pode ser visto em Mardia et al. (1992), a principal diferença entre essas duas técnicas é que, a análise de componentes principais é um modelo matemático e nada mais é do que uma mera transformação dos dados sem nenhuma suposição pré-estabelecida, ao passo que a análise fatorial tem um modelo definido (estatístico), em que várias suposições devem ser aceitas para a aplicação correta do método.

Após a construção dos componentes principais é possível calcular escores e classificá-los através das suas respectivas pontuações. Frequentemente, esses escores são utilizados na aplicação de outras técnicas estatísticas (JOHNSON; WICHERN, 2007), como a análise de agrupamento, por exemplo. Esses escores são computados com base nos autovetores da matriz de covariâncias ou da matriz de correlação, como será elucidado a seguir.

2.4.1 Extração dos componentes principais

Segundo a literatura estudada, pode-se extrair os componentes principais por meio da matriz de covariâncias $\boldsymbol{\Sigma}$ ou da matriz de correlações $\boldsymbol{\rho}$, ambas de dimensão $p \times p$. Na prática, nenhuma dessas matrizes são conhecidas, sendo assim, a estimação dos componentes principais é realizada através da matriz de covariâncias amostral $\boldsymbol{S}_{p\times p}$ ou da matriz de correlações amostral $\boldsymbol{R}_{p\times p}$.

As notações desta subseção serão as mesmas utilizadas por Mingoti (2005) e Johnson e Wichern (2007). Vale ressaltar que a letra y não possui o mesmo significado da seção anterior como será ilustrado posteriormente.

2.4.1.1 Matriz de Covariâncias

Seja $\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}'$ um vetor de variáveis independentes com matriz de covariância amostral $\boldsymbol{S}_{p\times p}$ e matriz de correlações amostral $\boldsymbol{R}_{p\times p}$, o j-ésimo componente principal amostral estimado, em que $j = 1, \cdots, p$, é dado por (JOHNSON; WICHERN, 2007):

$$\hat{Y}_j = \hat{\boldsymbol{e}}_j' \boldsymbol{X} = \hat{e}_{j1} X_1 + \hat{e}_{j2} X_2 + \cdots + \hat{e}_{jp} X_p, \tag{5}$$

em que $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \cdots, \hat{\boldsymbol{e}}_p$ são os autovetores normalizados correspondentes aos autovalores

$\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_p$ extraídos da matriz de covariâncias amostral $\boldsymbol{S}_{p\times p}$. Neste texto, elementos dos autovetores em cada componente serão citados como pesos dos componentes, ou simplesmente pesos. Ainda, tem-se que

$$Var[\hat{Y}_j] = \hat{\lambda}_j \, , j = 1, 2, \cdots, p$$

e

$$Cov(\hat{Y}_j, \hat{Y}_l) = 0 \, , \forall l \neq j \ , \ l = 1, \cdots, p.$$

A proporção de variância total explicada (PVTE_j) pelo j-ésimo componente principal é dada por (MINGOTI, 2005):

$$\text{PVTE}_j = \frac{Var[\hat{Y}_j]}{\text{Variância Total Estimada de X}} = \frac{\hat{\lambda}_j}{\text{tr}[\boldsymbol{S}_{p\times p}]} = \frac{\hat{\lambda}_j}{\sum_{j=1}^{p} \hat{\lambda}_j}.$$

Usualmente, a PVTE é apresentada em forma de percentagem, ou seja, $\text{P}VTE_j \times 100\%$. A correlação estimada entre o j-ésimo componente amostral e a variável aleatória X_i, $i = 1, 2, \cdots, p$ é dada da seguinte maneira

$$r_{\hat{Y}_j, X_i} = \frac{\hat{e}_{ji}\sqrt{\hat{\lambda}_j}}{\sqrt{s_{ii}}},$$

em que s_{ii} é a variância amostral da variável aleatória X_i.

Finalmente, utilizando o teorema da decomposição espectral (GOLUB; REINSCH, 1970), a matriz de covariâncias $\boldsymbol{S}_{p\times p}$ pode ser expressa como

$$\boldsymbol{S}_{p\times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{\boldsymbol{e}}_j \hat{\boldsymbol{e}}_j'.$$

2.4.1.2 Matriz de Correlações

Quando a estimação dos componentes principais é realizada através da matriz de covariâncias, os resultados podem ser severamente influenciados se pelo menos uma das variáveis possuir uma variância numericamente muito discrepante das demais. Uma das formas de corrigir este problema é utilizar uma transformação nos dados originais. Usualmente, a

transformação utilizada é a padronização de cada variável pela sua respectiva média e desvio padrão, ou seja

$$Z_j = \frac{X_j - \mu_j}{\sigma_i}, \tag{6}$$

em que $j = 1, 2, \cdots, p$, μ_j é a média e σ_j^2 é a variância da j-ésima variável aleatória, respectivamente.

Trabalhar com esta padronização é o mesmo que trabalhar com a matriz de correlações amostral $\boldsymbol{R}_{p\times p}$ dos dados originais (MINGOTI, 2005). Lembrando que, os coeficientes obtidos a partir desta padronização são diferentes dos obtidos pela matriz de covariâncias.

Após a padronização, o j-ésimo componente principal amostral da matriz de correlação amostral $\boldsymbol{R}_{p\times p}$ é dado por (JOHNSON; WICHERN, 2007):

$$\hat{Y}_j = \hat{\boldsymbol{e}}_j' \boldsymbol{Z} = \hat{e}_{j1} Z_1 + \hat{e}_{j2} Z_2 + \cdots + \hat{e}_{jp} Z_p, \tag{7}$$

em que $\hat{\boldsymbol{e}}_1, \hat{\boldsymbol{e}}_2, \cdots, \hat{\boldsymbol{e}}_p$ são autovetores correspondentes aos autovalores $\hat{\lambda}_1, \hat{\lambda}_2, \cdots, \hat{\lambda}_p$ extraídos da matriz de correlações amostral $\boldsymbol{R}_{p\times p}$. Ainda, tem-se que

$$Var[\hat{Y}_j] = \hat{\lambda}_j \, , j = 1, 2, \cdots, p$$

e

$$Cov(\hat{Y}_j, \hat{Y}_l) = 0 \, , \forall l \neq j, \; l = 1, \cdots, p.$$

A proporção de variância total explicada (PVTE_j) pelo j-ésimo componente principal é dada por (MINGOTI, 2005):

$$\text{PVTE}_j = \frac{\hat{\lambda}_j}{p} \, , j = 1, 2, \cdots, p.$$

A correlação estimada entre o j-ésimo componente amostral $\hat{Y}_j$ e a variável padronizada Z_i, $i = 1, 2, \cdots, p$ é dada da seguinte maneira

$$r_{\hat{Y}_j, Z_i} = \hat{e}_{ij} \sqrt{\hat{\lambda}_j}.$$

Utilizando o teorema da decomposição espectral, a matriz de correlação $\boldsymbol{R}_{p\times p}$ pode ser expressa como

$$\boldsymbol{R}_{p\times p} \approx \sum_{j=1}^{k} \hat{\lambda}_j \hat{\boldsymbol{e}}_j \hat{\boldsymbol{e}}_j'.$$

2.4.2 Critério para escolha dos k componentes principais

Alguns critérios para escolha do número k de componentes principais citados em Johnson e Wichern (2007) são:

1) Análise da representatividade em relação a variância total:
Este critério pode ser subdividido em outros três critérios:

i) A análise baseia-se em manter um número k de componentes principais que representem uma percentagem $\gamma \times 100\%$ da variância total, em que $0 < \gamma < 1$ é pré-determinado pelo pesquisador. Na prática, busca-se

$$\frac{\sum_{i=1}^{k} \hat{\lambda}_i}{\sum_{j=1}^{p} \hat{\lambda}_j} = \gamma;$$

ii) Quando a matriz de correlações é utilizada para a extração dos componentes principais, um dos critérios para a retenção do número k de componentes principais é denominado Critério de Kaiser (KAISER, 1958), que se baseia na escolha de k por meio dos autovalores. Mantém-se no sistema as componentes referentes aos autovalores que sejam maiores que 1, ou seja, $\hat{\lambda}_j \geq 1$, para $j \in \{1, ..., p\}$.

Similarmente, quando a matriz de covariâncias é utilizada para a extração dos componentes principais, retém-se no sistema os componentes relacionados aos autovalores que sejam maiores ou iguais a $\hat{\lambda}_i$, dado por

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{p} \hat{\lambda}_j}{p};$$

iii) Análise do *scree-plot* (CATTELL, 1966): o gráfico *scree* tem como eixos os autovalores *versus* ordem do componente. Retém-se os componentes observando no gráfico o ponto onde os valores obtidos para cada autovalor tendem a se estabilizar. Por exemplo, observando-se a Figura 2, por este critério, apenas dois componentes principais seriam retidos no sistema.

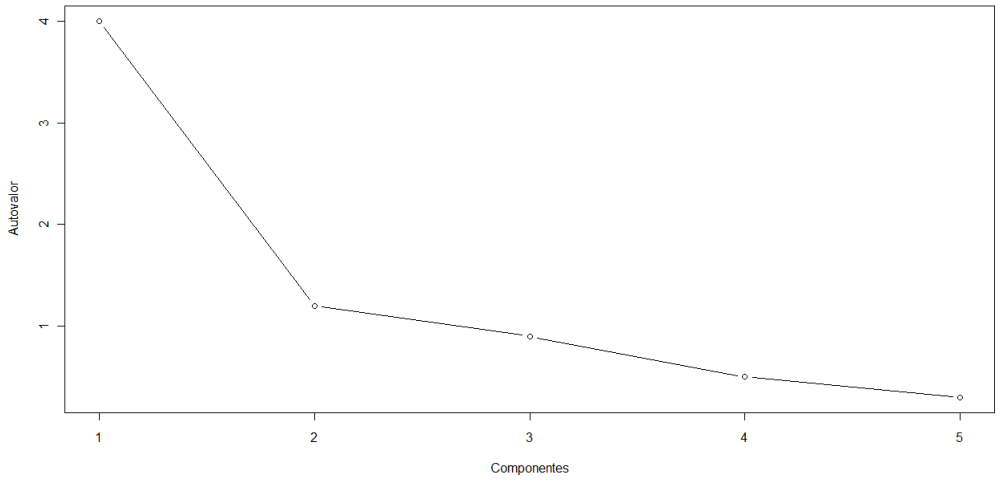

Figura 2 - Exemplo de gráfico *scree-plot*

2) Análise da qualidade de aproximação da matriz de correlação:
   Quando os componentes são extraídds, tanto da matriz de covariâncias, como da matriz de correlações, as seguintes aproximações podem ser realizadas, respectivamente:

$$\boldsymbol{S}_{p\times p} \approx \sum_{i=1}^{k} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i' \quad \text{e} \quad \boldsymbol{R}_{p\times p} \approx \sum_{i=1}^{k} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i',$$

   em que $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ são os respectivos autovalores e autovetores normalizados a partir das matrizes $\boldsymbol{S}_{p\times p}$ e $\boldsymbol{R}_{p\times p}$. O valor de k é escolhido determinando-se uma aproximação razoável.

3) Análise prática dos componentes:
   Os componentes retidos no sistema devem ser passíveis de interpretação, ou seja, o pesquisador deve ser capaz de entender e elucidar o significado de cada componente.

Após a seleção dos componentes principais a serem utilizados, realiza-se a nomeação e o cálculo dos escores de cada um dos componentes, que são não correlacionados uns com os outros e, desta forma, deve-se estudar, separadamente, cada um deles. As variáveis que possuem pesos elevados nos autovetores em determinado componente são as que mais contribuem com a rotulação do mesmo.

Em determinado componente, podem haver diversas variáveis com pesos elevados. Neste caso, tenta-se nomear o componente de modo que esse rótulo reflita todas essas

variáveis com pesos elevados; esse rótulo tem como objetivo, impor um significado para o componente principal em estudo. O rótulo dado para o componente depende, única e exclusivamente, da sensibilidade do pesquisador, uma vez que os pacotes computacionais não fornecem esses nomes.

Após a estimação e nomeação dos componentes selecionados por um dos critérios descritos neste trabalho, o escore é computado com base nos pesos de todas as variáveis no componente, ou seja, no cálculo dos escores são considerados todos os pesos de determinado componente principal, mesmo que sejam pesos de variáveis pouco influentes para o componente.

Por fim, Mingoti (2005) afirma que, normalmente, o número de componentes principais retidos para explicar uma percentagem $\gamma \times 100$ pela matriz de correlações amostral $\boldsymbol{R}_{p\times p}$ é superior àqueles retidos pela matriz de covariâncias amostral $\boldsymbol{S}_{p\times p}$. Isso se deve ao fato de que, como explicado anteriormente, quando uma ou mais variáveis possuem variância muito discrepante das demais e os componentes principais são retidos da matriz de covariâncias amostrais, os autovalores dos primeiros componentes tendem a reter mais informação.

Ainda, é possível construir intervalos de confiança para os autovalores populacionais se a distribuição das variáveis em estudo forem gaussianas, porém, como pode ser visto em Johnson e Wichern (2007), não há necessidade de verificar-se a distribuição de cada uma das variáveis se a aplicação do método tem caráter exploratório.

2.5 Análise biplot

O biplot é uma técnica multivariada que foi proposta por Gabriel (1971) e tem como principal objetivo realizar uma representação gráfica de uma matriz $\boldsymbol{X}$ de dados aproximada para uma matriz $\boldsymbol{Y}$ de menor dimensão (LAVORANTI et al., 2002). O biplot pode ser visto como uma generalização do gráfico de dispersão (GREENACRE, 2010). Essa representação gráfica é baseada na decomposição em valores singulares (ECKART; YOUNG, 1939) e pode ser expressa em duas ou três dimensões (ARAÚJO, 2009).

Gabriel (1978) afirma que qualquer matriz de posto r pode ser representada como um biplot, que segundo Lara et al. (2005) é de imenso interesse prático, uma vez que resume as informações que estão sendo estudadas e permite a análise e inspeção da estrutura entre as unidades amostrais e as variáveis. Ainda com base em Lara et al. (2005), qualquer matriz

$\boldsymbol{X}_{n\times m}$ de posto r pode ser fatorada como

$$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{H}',$$

em que $\boldsymbol{G}$ é uma matriz de dimensão $n\times r$ e $\boldsymbol{H}$ é uma matriz de dimensão $m\times r$, ambas necessariamente de posto r. O autor alerta que a fatoração não é única e que uma forma de aproximar a matriz $\boldsymbol{X}$ de dados para uma matriz $\boldsymbol{Y}$ é utilizar a decomposição por valores singulares ou a análise de componentes principais, já citada neste texto. Araújo (2009) detalha que, utilizando uma decomposição de duas dimensões, cada elemento $\hat{x}_{ij}$ de $\hat{\boldsymbol{X}}$ pode ser escrito como:

$$\hat{x}_{nm} = g_{n1}h_{m1} + g_{n2}h_{m2}, \tag{8}$$

que é o produto interno dos vetores linhas (g_{n1}, g_{n2}) e (h_{m1}, h_{m2}). O biplot é obtido então, representado-se cada linha como um ponto G_n com coordenadas (g_{n1}, g_{n2}) e cada coluna com um ponto H_m com coordenadas (h_{m1}, h_{m2}) em um gráfico bidimensional (Figura 3).

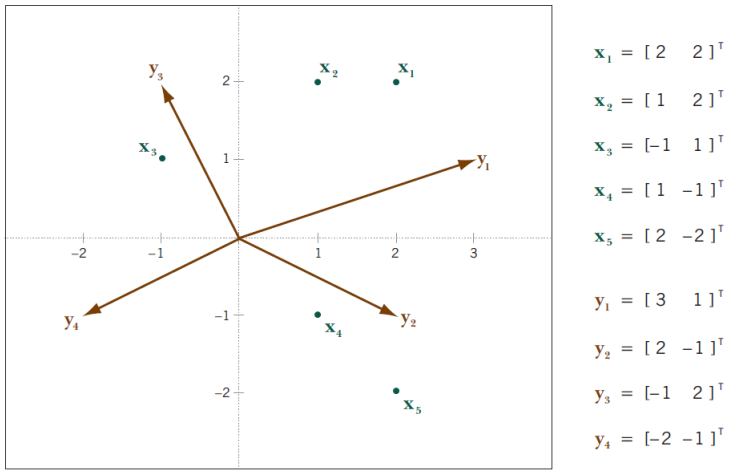

Figura 3 - Exemplo de um biplot, em que x_i e y_i representam, respectivamente, as linhas e colunas de uma matriz $\boldsymbol{X}$ qualquer em estudo. Fonte: Greenacre (2010)

Os vetores representam as variáveis em estudo e os pontos, os indivíduos em estudo. Quanto menor o ângulo entre dois vetores, mais relacionadas as variáveis estão, logo se o ângulo formado por dois vetores for próximo de zero, diz-se que a explicação dessas duas variáveis em um determinado modelo, são bem similares. Ainda, quanto maior a norma de

um vetor em relação a um determinado eixo (componente principal), mais influência a variável possui sobre essa determinado componente (GREENACRE, 2010).

É possível projetar perpendicularmente cada ponto do gráfico com relação à cada vetor e assim comparar a relação ou a interação entre dois objetos com a mesma variável avaliando-se o comprimento dessa projeção (ARAÚJO, 2009).

2.6 Análise de agrupamentos

A análise de agrupamentos, também conhecida como análise de cluster, como o próprio nome sugere, reúne objetos em grupos, de tal maneira que objetos no mesmo grupo sejam os mais homogêneos possíveis, com base em determinadas características (variáveis mensuradas). Por se tratar de um método não-inferencial, esta técnica é caracterizada, por muitos autores, como sendo de caráter exploratório.

Uma importante advertência salientada por Hair et al. (2005) é que, independentemente da real existência de uma estrutura nos dados, grupos serão formados ao longo de um dos procedimentos existentes na análise de agrupamentos. Logo, as variáveis utilizadas no método devem ser pertinentes com o objetivo do estudo.

Uma questão relevante que surge neste momento é de como se medir quão parecidos são dois objetos, ou seja, como medir a similaridade entre eles. A similaridade é um valor que mede a semelhança entre objetos, sendo que, quanto maior o valor da similaridade, mais parecidos os objetos são.

Johnson e Wichern (2007) sugerem alguns critérios baseados, em sua maioria, em medidas de distância para definir a similaridade entre um determinado objeto i com um objeto j. Algumas distâncias citadas na literatura são (MANLY, 2005):

1) Distância euclidiana generalizada:

$$d_{ij} = \left\{ \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right\}^{1/2},$$

em que $x_{.k}$ é a k-ésima variável, com $k = 1, 2, ..., p$; i é um objeto e j outro;

2) Distância de Mahalanobis:

$$d_{ij} = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)' \boldsymbol{S}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j)},$$

em que $\boldsymbol{X}_i$ corresponde ao vetor aleatório contendo as observações da variável i, $\boldsymbol{X}_j$ corresponde ao vetor aleatório contendo as observações da variável j e $\boldsymbol{S}$ é a matriz de variâncias e covariâncias amostral;

3) Métrica de Minkowski:

$$d_{ik} = \left[\sum_{k=1}^{p} |x_{ik} - x_{jk}|^m\right]^{\frac{1}{m}},$$

em que x_{ik} e x_{jk} possuem a mesma interpretação que na equação da distância euclidiana. Observe que para $m = 2$ tem-se a distância euclidiana;

4) Métrica de Canberra:

$$d_{ij} = \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})},$$

em que os elementos da equação são os mesmos dos vistos na distância euclidiana generalizada;

5) Coeficiente de Czekanowsli:

$$d_{ij} = 1 - \frac{2\sum_{k=1}^{p} \min(x_{ik}, x_{jk})}{\sum_{k=1}^{p} (x_{ik} + x_{jk})},$$

em que os elementos x_{ik} e x_{jk} possuem o mesmo significado no cálculo da distância euclidiana.

Em geral, as distâncias são calculadas por meio de variáveis padronizadas, como por exemplo, pela diferença de suas respectivas médias dividido pelos seus respectivos desvios padrões, como na eq. (6).

2.6.1 Métodos de agrupamento

Uma vez computadas as distâncias por meio de um dos métodos possíveis (como os citados na seção anterior), algum procedimento que auxilie na formação dos grupos de objetos deve ser utilizado. De um modo geral, na análise de agrupamento, esses procedimentos podem ser divididos em dois grupos principais: i) Métodos hierárquicos e ii) Métodos não-hierárquicos.

2.6.1.1 Métodos hierárquicos

Como o nome sugere, este método constrói, hierarquicamente, grupos de objetos homogêneos em diferentes etapas, de forma a construir uma árvore de classificação denominada dendrograma (Figura 4).

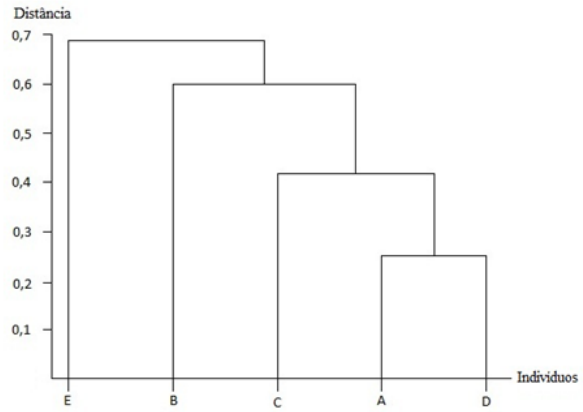

Figura 4 - Exemplo de um dendrograma

Segundo Hair et al. (2005), o dendrograma nada mais é do que uma representação gráfica dos resultados, que nos mostra como os agrupamentos são realizados em cada passo do processo hierárquico. Pode-se dizer que a utilização do dendrograma é muito interessante, pois facilita a visualização da formação dos grupos e, consequentemente, auxilia na interpretação dos mesmos.

Os métodos hierárquicos podem ser subdivididos em:

1) Métodos aglomerativos: o processo é iniciado com k grupos, ou seja, cada indíviduo está alocado em um grupo distinto. No primeiro passo, os objetos com maior similaridade são agregados em um único grupo, sendo assim, o número total de grupos nesta fase é de $k-1$. O processo continua até que apenas um grupo, que contenha todos os indivíduos seja formado;

2) Métodos divisivos: o processo é o oposto do método aglomerativo, ou seja, o algoritmo é iniciado com todos os objetos em apenas um grupo. Em etapas sucessivas, as observações menos homogêneas são divididas e transformadas em grupos menores até que cada observação esteja em um grupo unitário, ou seja, o processo se encerra quando atinge-se

o número k de grupos.

De modo ilustrativo, pode-se dizer que, na Figura 4, o método aglomerativo se move de baixo para cima e o método divisivo de cima para baixo. Na maioria da literatura disponível sobre o assunto, apenas o método aglomerativo é utilizado, uma vez que, os pacotes computacionais disponibilizam, em sua maioria, melhor formulados, apenas esses métodos. Alguns dos principais métodos hierárquicos aglomerativos são (FERREIRA, 2008): i) vizinho mais próximo; ii) vizinho mais distante; iii) ligação média; iv) método centróide; e v) método de Ward. Os procedimentos de cada um dos métodos aqui citados serão explanados posteriormente com um exemplo.

O método de agrupamento aglomerativo vem sendo utilizado nas mais diversas áreas de estudo, como, por exemplo, na economia aplicada. Firetti et al. (2010), com o intuito de proporcionar subsídios para o desenvolvimento regional da Região Administrativa de Presidente Prudente - SP, conseguiram bons resultados oferecendo ferramentas para a elaboração de políticas públicas em grupos distintos de municípios, utilizando os escores fatoriais de variáveis referentes ao número de vínculos empregatícios da região em estudo como variáveis em uma análise de agrupamentos.

Maia et al. (2009) ajustaram diversas curvas de crescimento em banananeiras do tipo anã e, posteriormente, agruparam todas as curvas (80 no total) utilizando como variáveis os parâmetros obtidos pelo modelo mais adequado ao conjunto de dados.

De maneira geral, o processo de agrupamento hierárquico aglomerativo se dá pelos seguintes passos, lembrando que esses passos são válidos para todos os critérios que aqui serão citados:

1) Calcular a matriz de distâncias $\boldsymbol{D}$ de dimensão $n \times n$;

2) Identificar na matriz calculada, qual a menor distância entre os pares de indivíduos. Agrupar esses dois objetos;

3) Recalcular a matriz de distâncias $\boldsymbol{D}$, agora com dimensão $(n-1) \times (n-1)$ utilizando um dos critérios que serão explicados nas próximas seções;

4) Repetir os passos 2 e 3 até que a matriz $\boldsymbol{D}$ tenha dimensão 1×1.

Para exemplificar cada método de agrupamento aglomerativo, será utilizada apenas uma matriz de distâncias como exemplo. Seja $\boldsymbol{D}$ uma matriz de distâncias dada por:

$$\boldsymbol{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0,2 & 0,4 & 0,4 \\ 0,2 & 0 & 0,5 & 0,3 \\ 0,4 & 0,5 & 0,0 & 0,3 \\ 0,4 & 0,3 & 0,3 & 0 \end{bmatrix} \end{array},$$

observa-se que, como a matriz $\boldsymbol{D}$ é uma matriz de distâncias, ela possui apenas elementos iguais a 0 em sua diagonal principal e é uma matriz simétrica.

2.6.1.1.1 Vizinho mais próximo

O método do vizinho mais próximo consiste em recalcular a matriz de distâncias (Passo 3) pela distância mínima. No exemplo anterior, a menor distância entre objetos encontra-se no elemento $[1,2]$ da matriz $\boldsymbol{D}$. Seguindo os passos elucidados anteriormente, agrupa-se os indivíduos 1 e 2. Agora que o primeiro grupo foi formado, deve-se atualizar a matriz $\boldsymbol{D}$. Por este método, na atualização, deve-se utilizar a menor distância entre os indivíduos $[1,2]$ e os demais, ou seja:

$$d_{(12)3} = \min(d_{13}; d_{23}) = \min(0,5; 0,4) = 0,4;$$

$$d_{(12)4} = \min(d_{14}; d_{24}) = \min(0,3; 0,4) = 0,3.$$

Logo, a matriz $\boldsymbol{D}$ atualizada fica da seguinte maneira:

$$\boldsymbol{D} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} 12 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 0,4 & 0 & \\ 0,3 & 0,3 & 0 \end{bmatrix} \end{array}.$$

Basta repetir este procedimento até que a matriz $\boldsymbol{D}$ tenha dimensão 1×1. Observe ainda que na última matriz apresentada (matriz $\boldsymbol{D}$ atualizada) existem dois pares de elementos com a mesma distância. Quando isto ocorre pode-se agrupar qualquer um dos pares, embora na literatura seja mais comum agrupar o primeiro par de distâncias encontrado.

2.6.1.1.2 Vizinho mais distante

Ao contrário do método anterior, o método do vizinho mais distante atualiza a matriz $\boldsymbol{D}$ através da distância máxima entre os indivíduos, ou seja, pelo exemplo, deve-se calcular as distâncias, após o agrupamento dos indivíduos $[1, 2]$, da seguinte maneira:

$$d_{(12)3} = \max(d_{13}; d_{23}) = \max(0,5; 0,4) = 0,5;$$

$$d_{(12)4} = \max(d_{14}; d_{24}) = \max(0,3; 0,4) = 0,4.$$

Sendo assim, a matriz $\boldsymbol{D}$ atualizada fica da seguinte maneira:

$$\boldsymbol{D} = \begin{array}{c} 12 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & & \\ 0,5 & 0 & \\ 0,4 & 0,3 & 0 \end{bmatrix}.$$

E repete-se o processo até que $\boldsymbol{D}$ tenha dimensão 1×1

2.6.1.1.3 Ligação média

Neste método, calcula-se a média entre as distâncias entre os indivíduos no processo de atualização da matriz de distâncias, ou seja, pelo exemplo, deve-se calcular as distâncias, após o agrupamento dos indivíduos $[1, 2]$, da seguinte maneira:

$$d_{(12)3} = \frac{d_{13} + d_{23}}{2} = \frac{0,5 + 0,4}{2} = 0,45;$$

$$d_{(12)4} = \frac{d_{14} + d_{24}}{2} = \frac{0,3 + 0,4}{2} = 0,35.$$

Portanto, a matriz de distâncias atualizada pelo método da ligação média é dada por:

$$\boldsymbol{D} = \begin{array}{c} 12 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & & \\ 0,45 & 0 & \\ 0,35 & 0,3 & 0 \end{bmatrix}.$$

Novamente, basta repetir o processo até que a matriz $\boldsymbol{D}$ seja de dimensão 1×1.

2.6.1.1.4 Método centróide

Como pode ser visto em Ferreira (2008), a distância entre dois grupos R e S, no método centróide, é definida como a distância euclidiana quadrática entre os vetores de médias, centróides, desses grupos. Suponha que R possui n_r elementos e S possui n_s elementos, então os centróides de ambos os grupos são dados, respectivamente, por:

$$\bar{\boldsymbol{y}}_r = \frac{\sum_{r=1}^{n_r} \boldsymbol{y}_r}{n_r} = \begin{bmatrix} \bar{y}_{r1} \\ \bar{y}_{r2} \\ \vdots \\ \bar{y}_{rp} \end{bmatrix} \quad \text{e} \quad \bar{\boldsymbol{y}}_s = \frac{\sum_{s=1}^{n_s} \boldsymbol{y}_s}{n_s} = \begin{bmatrix} \bar{y}_{s1} \\ \bar{y}_{s2} \\ \vdots \\ \bar{y}_{sp} \end{bmatrix},$$

em que $\boldsymbol{y}_r$ e $\boldsymbol{y}_s$ são os vetores de observações dos objetos pertencentes aos grupos R e S, respectivamente.

Uma vez calculados os centróides, calcula-se a distância euclidiana quadrada entre os grupos R e S da seguinte maneira

$$d_{rs}^2 = \sum_{u=1}^{p} (\bar{y}_{ru} - \bar{y}_{su})^2.$$

Quando o grupo formado a partir da junção de R e S - (RS), seu centróide é dado por

$$\bar{\boldsymbol{y}}_{(rs)} = \frac{n_r \bar{\boldsymbol{y}}_r + n_s \bar{\boldsymbol{y}}_s}{n_r + n_s},$$

que nada mais é do que uma média ponderada entre os grupos R e S.

Quando um novo grupo T qualquer é agregado ao grupo RS já formado, utiliza-se uma fórmula de atualização de distâncias quadradas, dada por

$$d_{(rs)t} = \frac{n_r}{n_r + n_s} d_{rt} + \frac{n_s}{n_r + n_s} d_{st} - \frac{n_r n_s}{(n_r + n_s)^2} d_{rs}, \tag{9}$$

em que as distâncias d_{rt}, d_{st} e d_{rs} provém da matriz de dissimilaridades antes da atualização referente ao agrupamento dos grupos R e S.

Utilizando o exemplo desta seção, novamente agrupa-se os indivíduos 1 e 2. Após a formação do grupo $[1, 2]$, a atualização da matriz de distâncias é realizada da seguinte forma (eq. 9):

$$d_{(12)3} = \frac{1}{1+1} d_{13} + \frac{1}{1+1} d_{23} - \frac{1 \times 1}{(1+1)^2} d_{12} = \frac{0,4}{2} + \frac{0,5}{2} - \frac{0,2}{4} = 0,4;$$

$$d_{(12)4} = \frac{1}{1+1}d_{14} + \frac{1}{1+1}d_{24} - \frac{1\times1}{(1+1)^2}d_{12} = \frac{0,4}{2} + \frac{0,3}{2} - \frac{0,2}{4} = 0,3.$$

Portanto, a matriz de distâncias atualizada pelo método centróide é dada por:

$$\boldsymbol{D} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & & \\ 0,4 & 0 & \\ 0,3 & 0,3 & 0 \end{bmatrix}.$$

Basta repetir este processo até que a matriz de distâncias $\boldsymbol{D}$ tenha dimensão 1×1.

2.6.1.1.5 Método de Ward

O método de Ward (WARD, 1963) é um dos mais utilizados, atualmente, em artigos científicos que utilizam a análise de agrupamentos devido ao forte apelo estatístico envolvido em seu processo que é baseado na análise de variância (FERREIRA, 2008). Basicamente, busca-se agregar os grupos R e S que minimizam a soma de quadrados do resíduo dentro dos grupos.

No contexto da análise de agrupamentos, a soma de quadrados dos erros dos grupos individuais em relação a todos os grupos é dada por (FERREIRA, 2008):

$$SQE = \sum_{l=1}^{k}\sum_{i=1}^{n_l}(\boldsymbol{y}_i^{(l)} - \bar{\boldsymbol{y}}^{(l)})^T(\boldsymbol{y}_i^{(l)} - \bar{\boldsymbol{y}}^{(l)}).$$

em que $\boldsymbol{y}_i^{(l)}$ corresponde ao vetor do i-ésimo objeto do l-ésimo grupo e $\bar{\boldsymbol{y}}^{(l)}$ é a o vetor de médias do l-ésimo grupo.

Após a junção dos grupos R e S a soma de quadrados é alterada da seguinte maneira:

$$\Delta SQE_{rs} = SQE_{rs} - SQE_r - SQE_s, \tag{10}$$

em que

$$SQE_{rs} = \sum_{i=1}^{n_{rs}}(\boldsymbol{y}_i^{(rs)} - \bar{\boldsymbol{y}}^{(rs)})^T(\boldsymbol{y}_i^{(rs)} - \bar{\boldsymbol{y}}^{(rs)})$$

em que o expoente rs corresponde ao agrupamento do grupo R com o grupo S e

$$\bar{\boldsymbol{y}}^{(rs)} = \frac{(n_r\bar{\boldsymbol{y}}^{(r)} + n_s\bar{\boldsymbol{y}}^{(s)})}{n_r + n_s},$$

é o centróide do grupo RS formado.

Quando o número de grupos k é o mesmo de n, Ferreira (2008), alerta que a matriz de distâncias deve ser transformada na forma $\boldsymbol{P} = [0,5d_{ij}]$, ou seja, o método de Ward deve ser aplicado na matriz de proximidades. A medida de proximidade do grupo RS com um grupo T qualquer é dada por Jain e Dubes (1988) como

$$p_{(rs)t} = \frac{1}{n_t + n_{rs}}\left[(n_t + n_r)p_{rt} + (n_t + n_s)p_{st} - n_t p_{rs}\right]. \tag{11}$$

Utilizando o exemplo desta seção, a matriz de proximidades é dada por

$$\boldsymbol{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0,1 & 0,2 & 0,2 \\ 0,1 & 0 & 0,25 & 0,15 \\ 0,2 & 0,25 & 0 & 0,15 \\ 0,2 & 0,15 & 0,15 & 0 \end{bmatrix} \end{array}.$$

Os indivíduos 1 e 2 possuem o menor valor de medida de proximidade ($p_{(12)} = 0,1$), logo são agrupados. Utilizando a medida de proximidade apresentada por Jain e Dubis (1998) tem-se:

$$\begin{aligned} p_{(12)3} &= \frac{1}{1+2}\left[(1+1)p_{13} + (1+1)p_{23} - p_{12}\right] \\ &= \frac{1}{3}\left[2\times 0,2 + 2\times 0,25 - 0,1\right] \\ &= 0,266; \end{aligned}$$

$$\begin{aligned} p_{(12)4} &= \frac{1}{1+2}\left[(1+1)p_{14} + (1+1)p_{24} - p_{12}\right] \\ &= \frac{1}{3}\left[2\times 0,2 + 2\times 0,15 - 0,1\right] \\ &= 0,2. \end{aligned}$$

Portanto, a matriz de proximidades atualizada após o passo 1 é dada por:

$$\boldsymbol{P} = \begin{array}{c} \\ 12 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} 12 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0,266 & 0,2 \\ 0,266 & 0 & 0,15 \\ 0,2 & 0,15 & 0 \end{bmatrix} \end{array}.$$

Com a matriz atualizada, observa-se novamente o elemento com a menor medida de proximidade (indivíduos 3 e 4). A medida de proximidade é dada por:

$$\begin{aligned} p_{(34)(12)} &= \frac{1}{2+2}\left[(2+1)p_{3(12)} + (2+1)p_{4(12)} - p_{34}\right] \\ &= \frac{1}{4}\left[3\times 0,2 + 3\times 0,15 - 0,15\right] \\ &= 0,225. \end{aligned}$$

Desta forma, a matriz de proximidades atualizada após o segundo passo é dada por

$$\boldsymbol{P} = \begin{array}{c} 12 \\ 34 \end{array} \begin{bmatrix} 0 & 0,225 \\ 0,225 & 0 \end{bmatrix}.$$

Finalmente, agrega-se os grupos (12) e (34) com medida de proximidade dada por 0,225.

2.6.1.2 Métodos não-hierárquicos

Diferentemente dos métodos hierárquicos, nos procedimentos não-hierárquicos já se sabe, a priori, o número k de grupos a serem formados antes mesmo de se iniciar a análise (FERREIRA, 2008) e o resultado final não é esboçado na forma de um dendrograma. Como pode ser visto em Hair et al. (2005), primeiramente deve-se selecionar uma semente de agrupamento, que nada mais é do que uma centróide inicial qualquer. O problema central em qualquer método não-hierárquico é, justamente, a escolha adequada dessas sementes de agrupamento. O objetivo neste método é particionar, por meio de uma maximização ou minimização de certas características, o conjunto de dados em subgrupos mutuamente exclusivos. O mais conhecido método não-hierárquico, recebe o nome de método *k-means*. Para maiores informações sobre os métodos não-hierárquicos consultar Ferreira (2008).

Alguns autores sugerem ainda uma terceira abordagem, na qual se utiliza uma combinação dos procedimentos hierárquicos com procedimentos não-hierárquicos. Em suma, aplica-se um procedimento hierárquico no conjunto de dados em estudo, a fim de se escolher o número k de grupos. Após a escolha de k, aplica-se um procedimento não hierárquico com os

centros de grupos dos resultados hierárquicos como pontos de sementes iniciais. Desta forma, as vantagens de ambos os métodos são condensadas em apenas uma análise (HAIR et al., 2005).

Uma importante advertência que deve ser citada com relação ao uso deste método é que ele não possui solução única, uma vez que cabe ao pesquisador selecionar o melhor número de grupos a ser utilizado (MANLY, 2008). Além disso, a análise de agrupamentos sempre criará grupos, independentemente da "verdadeira" existência de qualquer estrutura nos dados (HAIR et al., 2005). Obviamente, as variáveis utilizadas em uma análise de agrupamentos devem ser relevantes para a classificação desejada, uma vez que os agrupamentos são baseados exclusivamente nas variáveis que são fornecidas nos dados.

3 MATERIAIS E MÉTODOS

3.1 Dados em Estudo

Os dados que foram utilizados no trabalho são pesagens (kg) referentes a 55 fêmeas da raça Hereford, nascidas nos anos de 1999 a 2001, na Agropecuária Recreio (Bagé - RS). As pesagens foram realizadas de 15 em 15 dias, desde o nascimento até, aproximadamente, 675 dias. Para obter um melhor número de modelos com parâmetros convergentes e significativos, foram retirados, de todas as vacas, as quatro últimas pesagens, logo a última pesagem considerada foi a do dia 615.

Neste trabalho foi realizada a modelagem dos perfis de crescimento dos animais individualmente, assim como já feito em Mendes (2007) e, posteriormente, as estimativas dos parâmetros foram utilizadas em uma análise de componentes principais, que retorna combinações lineares (componentes principais) de todas as variáveis em estudo e estas foram utilizadas como novas variáveis estatísticas a fim de formar grupos de bovinos similares utilizando-se o método hierárquico de agrupamento.

3.2 Análise de Regressao Não-Linear

Como descrito em Mendes (2007), o melhor modelo que se ajustou aos dados em estudo foi o modelo Gompertz difásico com estrutura de erros autorregressiva ponderado pelo inverso da variância dos pesos, logo foi o utilizado nesta dissertação com algumas modificações em sua parametrização.

Os modelos difásicos têm por característica principal serem divididos em dois ciclos (duas fases). O primeiro ciclo é representado pelos parâmetros A_1, B_1 e K_1 e o segundo ciclo é representado por A_2, B_2 e K_2. O modelo Gompertz difásico é dado pela seguinte expressão:

$$Y_i = A_1 \exp\{-\exp\{B_1 - K_1 X_i\}\} + A_2 \exp\{-\exp\{B_2 - K_2 X_i\}\} + \phi_1 \epsilon_{i-1} + \xi_i \qquad (12)$$

em que:

Y_i é o peso do animal (kg) no tempo i, para $i = 1, ..., t$;

X_i é a idade do animal no tempo i;

ξ_i é o erro de regressão no tempo i, com distribuição $N(0, \sigma^2)$;

ϵ_{i-1} é o erro de regressão no tempo $i-1$;

ϕ_1 é o parâmetro autorregressivo de ordem 1;

A_1 é o parâmetro que estima o peso assintótico do animal na primeira fase, ou seja,

$$\lim_{X_i \to \infty} [A_1 \exp\{-\exp\{B_1 - K_1 X_t\}\}] = A_1;$$

$A_1 + A_2$ é uma combinação linear dos parâmetros que representa o peso assintótico do animal ao final da segunda fase, ou seja,

$$\lim_{X_i \to \infty} [A_1 \exp\{-\exp\{B_1 - K_1 X_i\}\} + A_2 \exp\{-\exp\{B_2 - K_2 X_i\}\}] = A_1 + A_2;$$

K_1 é o parâmetro referente à taxa de crescimento da primeira fase;

K_2 é o parâmetro referente à taxa de crescimento da segunda fase;

os parâmetros B_1 e B_2 não possuem interpretação biológica, estando ligados somente à característica sigmoidal da curva.

As derivadas parciais de primeira ordem em relação aos parâmetros do modelo ilustrado na eq. (12), desconsiderando o parâmetro referente a parte autorregressiva, são:

$$\frac{\partial Y}{\partial A_1} = \exp\{-\exp\{B_1 - K_1 X_t\}\}$$

$$\frac{\partial Y}{\partial B_1} = -A_1 \exp\{B_1 - K_1 X_t\} \exp\{-\exp\{B_1 - K_1 X_t\}\}$$

$$\frac{\partial Y}{\partial K_1} = A_1 X_t \exp\{B_1 - K_1 X_t\} \exp\{-\exp\{B_1 - K_1 X_t\}\}$$

$$\frac{\partial Y}{\partial A_2} = \exp\{-\exp\{B_2 - K_2 X_t\}\}$$

$$\frac{\partial Y}{\partial B_2} = -A_2 \exp\{B_2 - K_2 X_t\} \exp\{-\exp\{B_2 - K_2 X_t\}\}$$

$$\frac{\partial Y}{\partial K_2} = A_2 X_t \exp\{B_2 - K_2 X_t\} \exp\{-\exp\{B_2 - K_2 X_t\}\}$$

Observa-se então que, de fato, o modelo utilizado provém de uma função não-linear, pois as derivadas parciais em relação aos parâmetros em estudo dependem de um ou mais parâmetros.

3.2.1 Métodos numéricos para obtenção das estimativas

Como elucidado na seção 2.3.1, para resolver o sistema de equações normais de um modelo não-linear, ou seja, para estimar os parâmetros dos modelos não-lineares por meio do método dos mínimos quadrados, faz-se necessário a utilização de algum método numérico. Apresenta-se nas próximas seções os métodos de Gauss-Newton, do Gradiente e de Marquardt, este último utilizado nesta dissertação.

3.2.1.1 Método de Gauss-Newton

Como dito, a utilização de um artifício para auxiliar na estimação através do método dos mínimos quadrados é crucial para calcular as estimativas. Souza (1998) afirma que o método de Gauss-Newton é uma técnica extremamente popular e, acima de tudo, muito eficiente na maioria das aplicações. Este método nada mais é do que um processo iterativo que lineariza a função não-linear, que consiste em utilizar uma aproximação linear à função esperança. Seja a eq. (2) reescrita na forma

$$S(\boldsymbol{\theta}) = ||\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})||^2, \tag{13}$$

em que se pode extrair uma aproximação geométrica de ser o quadrado da distância (norma ao quadrado) entre o vetor resposta $\boldsymbol{y}$ e o vetor $\boldsymbol{\eta}(\boldsymbol{\theta})$ (função resposta) em um espaço N-dimensional. Em uma aproximação linear, o vetor $\boldsymbol{\eta}(\boldsymbol{\theta})$ pode ser escrito como

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \eta(\theta^0) + \boldsymbol{V}^0(\theta - \theta^0), \tag{14}$$

em que $\boldsymbol{V}^0$ é a matriz jacobiana cujos elementos são as derivadas parciais da função esperança utilizada em relação a cada um de seus parâmetros e θ^0 é o valor inicial dado à θ.

Com a aproximação linear dada pela eq. (14) reiterativamente melhorasse o θ^0 como valor inicial para estimar θ e continuasse melhorando as estimativas até não existir mais mudanças segundo um critério de convergência. Segundo Ratkowsky (1983) esse método expande uma série de Taylor sobre θ_t como na eq. (14):

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \cong \eta(\theta^0) + \boldsymbol{V}^0(\theta - \theta^0).$$

Rearranjando a eq.(14) com a eq. (13), tem-se:

$$
\begin{aligned}
S \quad (\boldsymbol{\theta}) &= ||\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})||^2 = [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]^T[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})] \cong \\
\cong \quad & [\boldsymbol{y} - (\boldsymbol{\eta}(\boldsymbol{\theta_t}) + \boldsymbol{V_t}(\boldsymbol{\theta} - \boldsymbol{\theta_t}))]^T[\boldsymbol{y} - (\boldsymbol{\eta}(\boldsymbol{\theta_t}) + \boldsymbol{V_t}(\boldsymbol{\theta} - \boldsymbol{\theta_t}))] = \\
= \quad & [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta_t}) - \boldsymbol{V_t}(\boldsymbol{\theta} - \boldsymbol{\theta_t}))]^T[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta_t}) - \boldsymbol{V_t}(\boldsymbol{\theta} - \boldsymbol{\theta_t}))] = \\
= \quad & [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta_t})]^T[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta_t})] - 2.[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta_t})]^T\boldsymbol{V_t}(\boldsymbol{\theta} - \boldsymbol{\theta_t}) + (\boldsymbol{\theta} - \boldsymbol{\theta_t})^T\boldsymbol{V_t}^T\boldsymbol{V_t}(\boldsymbol{\theta} - \boldsymbol{\theta_t})
\end{aligned}
$$

o vetor gradiente $\boldsymbol{g}(\boldsymbol{\theta}) = \left(\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_2}, ..., \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_P}\right)^T$ é então:

$$\boldsymbol{g}(\boldsymbol{\theta}) = -2\boldsymbol{V}_t^T[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta_t})] + 2\boldsymbol{V}_t^T\boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta_t})$$

Igualando a expressão acima à zero, teremos então:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + [\boldsymbol{V}_t^T\boldsymbol{V}_t]^{-1}\boldsymbol{V}_t^T[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)] \tag{15}$$

Com valores iniciais para θ com $t = 1$, o processo continua até a convergência, que ocorre quando $||\theta_{t+1} - \theta_t||$ é menor do que uma quantidade já pré-estabelecida.

3.2.1.2 Método do Gradiente

O método do gradiente (em inglês *Steepest Descent*), consiste em selecionar uma região do espaço paramétrico em que n somas são produzidas a partir de n combinações de diferentes valores dos parâmetros $\boldsymbol{\theta}$ (DRAPER; SMITH, 1998).

Utilizando essas n somas, denotadas por $S(\theta)$, como observações de uma variável dependente e as combinações dos parâmetros $\theta_1, \theta_2, \cdots, \theta_P$ como observações de variáveis independentes, ajusta-se o seguinte modelo:

$$S(\boldsymbol{\theta}) = b_0 + \sum_{i=1}^{P} \frac{bi(\theta_i - \bar{\theta}_i)}{s_i} + \epsilon$$

em que $\bar{\theta}_i$ representa o valor médio de θ_i das combinações distintas dos parâmetros e s_i é um fator de escala.

Assim, observando-se as estimativas $b_0, b_1, \cdots, b_P$, toma-se:

$$\theta_i = \bar{\theta}_i - \lambda b_i s_i \tag{16}$$

em que $\lambda > 0$ é um valor escolhido de forma a minimizar $S(\boldsymbol{\theta})$. Maiores informações sobre a obtenção e utilização da constante λ podem ser obtidas em SAS (1995).

3.2.1.3 Método de Marquardt

O método de Marquardt consiste na interpolação dos dois métodos elucidados anteriormente. A estratégia utilizada baseia-se na interpolação dos parâmetros de correção δ_{GN} e δ_G, provenientes do método de Gauss-Newton e do Método do Gradiente, respectivamente, dados por:

$$\delta_{GN} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} \tag{17}$$

e

$$\delta_G = -\boldsymbol{X}'\boldsymbol{\epsilon} \tag{18}$$

Da interpolação da eq. (17) e eq. (18), retira-se o parâmetro de correção do método de Marquardt (δ):

$$\delta = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} \tag{19}$$

em que λ é conhecida como constante de Marquardt. Marquardt (1963), demonstra que

$$\delta \to \delta_{GN} \quad se \quad \lambda \to 0 \quad e \quad \delta \to \delta_G \quad se \quad \lambda \to \infty$$

Finalmente, o algoritmo de Marquardt é dado por:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} \tag{20}$$

3.2.2 Teste de Durbin-Watson

O teste de Durbin-Watson testa, como hipótese nula, se os resíduos de um determinado modelo ajustado a um determinado conjunto de dados não são autocorrelacionados *versus* a hipótese alternativa de que os resíduos possuem caráter autorregressivo de ordem 1. Em suma, este teste auxilia na constatação da presença de autocorrelação no modelo ajustado.

A estatística de Durbin-Watson é definida por (DURBIN; WATSON, 1950):

$$D = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2} \tag{21}$$

em que ϵ_i é o resíduo para o tempo i. Observa-se que o numerador apresenta a diferença ao quadrado entre os resíduos do tempo i e do tempo $(i-1)$ somados a partir da segunda observação. O denominador conota a soma de quadrados dos resíduos.

O teste é realizado comparando-se a estatística D obtida com os valores (limites) tabelados provenientes de uma tabela específica. Se $D < D_{\text{inferior}}$ existe indícios de que existe autocorrelação positiva entre os resíduos, ao passo que se $D > (4 - D_{\text{inferior}})$ existem evidências de autocorreçalação positiva nos resíduos do modelo ajustado. A hipótese nula não é rejeitada quando $D_{\text{superior}} < D < (4 - D_{\text{superior}})$. Para outros casos, o teste é inconclusivo.

3.3 Análise de Agrupamentos

As variáveis que serão utilizadas neste trabalho para a aplicação da análise de agrupamentos serão os componentes principais retidos das sete variáveis (três parâmetros da primeira fase, três parâmetros da segunda fase e um parâmetro referente a parte autorregressiva do modelo) retiradas do modelo da eq. (12). Neste trabalho as distâncias das observações de cada variável foram calculadas utilizando-se a distância euclidiana.

Atualmente, um dos métodos mais utilizados na tarefa de agrupar objetos similares é o Método de Ward, o qual será utilizado neste trabalho, devido ao seu forte apelo estatístico envolvido em seu processo. O método é um procedimento no qual a similaridade é usada para juntar agrupamentos somados sobre todas as variáveis. Formalmente, a distância entre dois grupos é a soma dos desvios ao quadrado dos pontos aos centróides. A atribuição de um elemento a um grupo é feita de modo a minimizar a soma de quadrados dentro dos grupos.

Segundo a literatura estudada, não existe um método robusto para a escolha do número de grupos a serem formados e sendo assim, essa escolha fica a cargo do pesquisador da área de interesse dos dados, que deve buscar uma solução aceitável para um determinado problema proposto, ou seja, os grupos formados devem ser passíveis de interpretação. Faria (2009) estudou diferentes metodologias para a escolha do número k de grupos e concluiu que a utilização do índice RMSSTD (do inglês *Root mean square standard deviation*) utilizado em conjunto com o método da máxima curvatura possuía um poder maior de discriminação de indivíduos quando comparado com as demais propostas para escolha do número de agrupamentos.

O índice RMSSTD mensura a homogeneidade dos agrupamentos, ou seja, quanto

menor o valor de RMMSTD, mais homogêneos os grupos estão, ao passo que quanto maior o valor, maior a heterogeneidade. Este índice é dado por:

$$RMSSTD_k = \sqrt{\frac{SQ_1 + SQ_2 + \cdots + SQ_p}{gl_1 + gl_2 + \cdots + gl_p}} \tag{22}$$

em que $SQ_i = \sum_{j=1}^{n}(x_{ji} - x_j)^2$ é a soma de quadrados da i-ésima variável calculada considerando as n observações em cada novo agrupamento k.

Após o cálculo dos valores de RMSSTD é possível realizar a construção de um gráfico cujos eixos são compostos pelo número de grupos *versus* o próprio valor do índice. O comportamento deste gráfico pode ser descrito na forma $RMSSTD = a(Nc)^{-b}$ em que a e b são os parâmetros do modelo e Nc o número de grupos formados (SILVEIRA, 2010). O número k de agrupamentos, através deste método, seria a maior distância entre o gráfico citado e a curva ajustada aos valores do RMSSTD.

4 RESULTADOS E DISCUSSÃO

Primeiramente, as curvas de cada uma das vacas foram esboçadas em um gráfico cujos eixos representam o peso *versus* idade (Figura 5). A curva esboçada em vermelho na Figura 5, representa os dados médios, ou seja, a média de pesos das 55 vacas em cada uma das mensuracões no decorrer dos 615 dias.

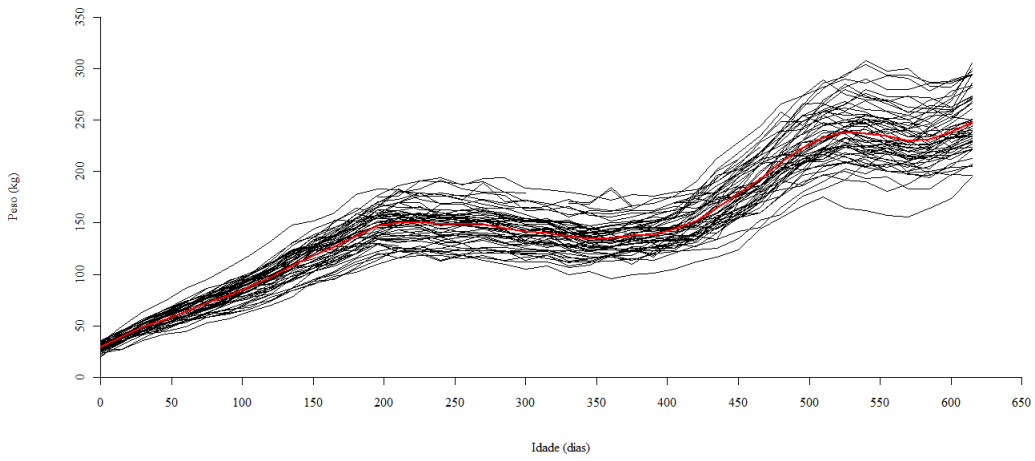

Figura 5 - Perfis de crescimento das 55 vacas em estudo. A curva destacada em vermelho corresponde ao peso médio dos animais

Nota-se claramente pela Figura 5 que, de fato, as curvas de crescimento dos animais em estudo podem ser repartidas em duas fases distintas. A primeira fase corresponde ao período dos dias 0 à 400, aproximadamente; ao passo que a segunda fase corresponde ao intervalo dos dias 400 e 615. Ainda, observa-se que nos primeiros dias de pesagens, a variância entre os animais é pequena e no decorrer dos dias essa variância vai aumentando consideravelmente, o que pode indicar a presença de heterogeneidade de variâncias. Observado esse fato, optou-se pelo ajuste do modelo Gompertz difásico com estrutura de erros autorregressiva ponderado pelo inverso da variância ao conjunto de dados. Mendes (2007) afirma que a utilização deste modelo, em curvas individuais, é o mais indicado devido à sua alta qualidade de ajuste aos bovinos fêmeas da raça Hereford.

Mazzini et al. (2003, 2005), ajustaram, respectivamente, cinco e quatro modelos não-lineares distintos para curvas de crescimento de machos da mesma raça. Como conclusão, os autores afirmam que o modelo Gompertz é o mais indicado para descrever o crescimento dos

novilhos por se ajustar extremamente bem aos perfis de crescimento dessa raça. Silva et al. (2004), ao ajustarem curvas de crescimento de bovinos da raça Nelore, chegaram nas mesmas conclusões, comprovando que o modelo Gompertz também é útil para descrever as curvas de crescimento outras raças bovinas.

As curvas de crescimento de outros animais também são bem ajustadas pelo modelo Gompertz. Sarmento et al. (2006), concluíram que, de fato, o modelo Gompertz era o mais adequado na descrição de perfis de crescimento de ovinos da raça Santa Inês.

Os estudos atuais de curvas de crescimento em animais ajustam, além das curvas individuais, um modelo para a curva de dados médios, perdendo-se muita informação, pois a modelagem desta única curva, considera tanto os animais com maiores pesos assintóticos e/ou maiores taxa de crescimento, similares com os animais que possuem essas características de modo menos favorecido. A curva destacada em vermelho da Figura 5, comparativamente às outras curvas de todos os animais em estudo, mostra claramente esse fato.

4.1 Ajuste do modelo Gompertz difásico

Inicialmente foi realizado o ajuste do modelo Gompertz difásico considerando estrutura de erros independentes. Após a realização do teste de Durbin-Watson, foi constatada a presença de resíduos autocorrelacionados, demonstrando assim a necessidade da inclusão de um parâmetro autorregressivo de primeira ordem.

Com a constatação da presença de autocorrelação, a estimação dos parâmetros do modelo Gompertz difásico com estrutura de erros autorregressiva de ordem 1, ponderado pelo inverso da variância dos pesos foi feita utilizando-se o método dos mínimos quadrados generalizados. Após esse ajuste, novamente foi realizado o teste de Durbin-Watson, onde nenhuma curva modelada apresentou resíduos correlacionados.

Como citado, neste trabalho a parametrização do modelo Gompertz difásico foi diferente da apresentada em Mendes (2007). Provavelmente, devido a esse fato e a exclusão das quatro últimas pesagens de cada animal, somente um ajuste não convergiu, ao passo que a autora citada conseguiu que apenas 69,09% dos ajustes, convergissem.

Como mencionado na seção de metodologia, após o ajuste dos modelos em todas as curvas de crescimento, a próxima etapa a ser realizada seria a aplicação da análise de

componentes principais utilizando como variáveis as estimativas de cada um dos parâmetros. Para a utilização deste método todos os indivíduos devem ter observações em todas as variáveis. Das 54 vacas restantes, cinco apresentaram pelo menos uma estimativa de determinado parâmetro não significativa pelo teste t e, por isso, foram excluídas da análise. Na Tabela 1 são apresentadas as estimativas dos parâmetros dos 49 animais restantes.

Tabela 1 - Estimativas dos parâmetros do modelo Gompertz difásico ajustado à cada uma das vacas em estudo

(continua)

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1
1	80,84703	21,68257	0,04558	149,51940	0,44271	0,01181	0,88718
2	103,38290	15,71899	0,03305	170,37890	0,62464	0,01231	0,84404
3	91,16059	17,24500	0,03678	147,48160	0,47661	0,01176	0,88271
4	133,32940	11,85520	0,02550	191,55370	0,71223	0,00982	1,06198
5	135,64990	13,78181	0,02952	220,67190	0,73440	0,01083	1,02570
6	74,32972	15,56787	0,03291	134,14090	0,29696	0,00765	0,81954
7	96,76313	16,39863	0,03439	173,95990	0,54238	0,01201	0,90062
8	119,77460	17,61013	0,03719	176,24680	0,62260	0,01166	0,8836
9	73,92454	21,70737	0,04518	152,00840	0,46271	0,01038	0,83527
10	107,48560	17,79472	0,03678	186,58500	0,54150	0,00988	0,79350
11	78,57852	16,79271	0,03730	137,43750	0,38704	0,01022	0,86563
12	97,25458	16,25377	0,03651	162,11260	0,55249	0,01191	0,87658
13	111,95250	13,80863	0,02996	174,28070	0,58633	0,01112	0,85450
14	118,36790	13,18548	0,02917	151,21420	0,42189	0,00878	0,84933
15	105,90850	9,97851	0,02178	143,28670	0,39682	0,00944	0,91516
16	110,42140	14,58517	0,03262	136,33890	0,49294	0,01068	0,92161
17	80,10375	13,57617	0,03038	150,54300	0,51459	0,01343	0,85926
18	86,35460	18,18088	0,04049	136,80090	0,44396	0,01297	0,80744
19	98,25995	11,18714	0,02455	150,90240	0,46726	0,01210	0,85425
20	107,25300	11,59170	0,02568	129,06160	0,58144	0,01261	0,74041

Tabela 1 - Estimativas dos parâmetros do modelo Gompertz difásico ajustado à cada uma das vacas em estudo

(continuação)

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1
21	89,36866	14,38542	0,03175	136,45970	0,58316	0,00904	0,76904
22	104,12880	14,05545	0,03114	140,47210	0,35493	0,00900	0,84647
23	112,41470	15,28183	0,03381	164,41090	0,44722	0,01081	0,8466
24	93,85148	13,41928	0,03009	155,44960	0,49402	0,01233	0,89055
25	93,73013	18,22190	0,04047	138,96330	0,54994	0,01468	0,73868
26	90,57353	12,18910	0,02634	130,66480	0,48296	0,00988	0,74568
27	105,47430	8,77222	0,01882	120,17010	0,32384	0,01092	0,85439
28	95,71713	13,70035	0,03003	152,21860	0,44003	0,01351	0,88206
29	123,04990	15,96120	0,03541	177,19090	0,57334	0,01076	0,89981
30	108,87870	11,70588	0,02585	131,07290	0,43157	0,01002	0,72637
31	100,46920	13,72610	0,03056	142,87620	0,35267	0,01302	0,92897
32	90,39831	13,82238	0,03059	135,94440	0,38322	0,01060	0,91111
33	110,26280	15,79819	0,03546	160,54100	0,51973	0,00955	0,77079
34	96,61646	13,95661	0,03220	127,44270	0,50006	0,01345	0,71023
35	103,29470	11,18499	0,02573	155,06880	0,60196	0,01183	0,85014
36	105,90880	14,82210	0,03445	156,91770	0,59911	0,01474	0,81044
37	135,96350	12,20040	0,02738	145,16710	0,43339	0,01087	0,81438
38	82,36584	19,82905	0,04551	131,89320	0,48386	0,01294	0,79575
39	117,90600	13,43990	0,03091	146,97490	0,50438	0,01279	0,91539
40	101,62940	16,95616	0,03920	160,18100	0,51440	0,01033	0,85810
41	83,95020	18,58329	0,04304	143,29510	0,49742	0,01389	0,85818
42	78,54047	12,39375	0,02879	121,19330	0,41225	0,01375	0,66904
43	87,55383	18,96227	0,04407	158,08300	0,52443	0,01236	0,83880
44	114,95320	15,62992	0,03590	173,40580	0,49797	0,01509	0,81805

Tabela 1 - Estimativas dos parâmetros do modelo Gompertz difásico ajustado à cada uma das vacas em estudo

(conclusão)

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1
45	103,38870	9,29304	0,02084	127,65360	0,30540	0,00796	0,78235
46	118,67280	12,53342	0,02889	150,68150	0,62959	0,01441	0,75213
47	106,21830	19,30339	0,04480	165,24850	0,42252	0,01034	0,81801
48	103,36050	13,02092	0,02949	145,88180	0,41328	0,01385	0,82835
49	97,58026	13,63745	0,03102	152,71550	0,38790	0,01217	0,87633

Nas Figuras 6 e 7 são apresentados gráficos de dispersão dos resíduos. Observa-se que todos os gráficos apresentam resíduos aleatórios próximos de zero e, como confirmado pelo teste de Durbin-Watson, são independentes entre si. Esse ajuste só foi possível graças a utilização do inverso da variância dos pesos como ponderador e a inclusão de um parâmetro autorregressivo ao modelo.

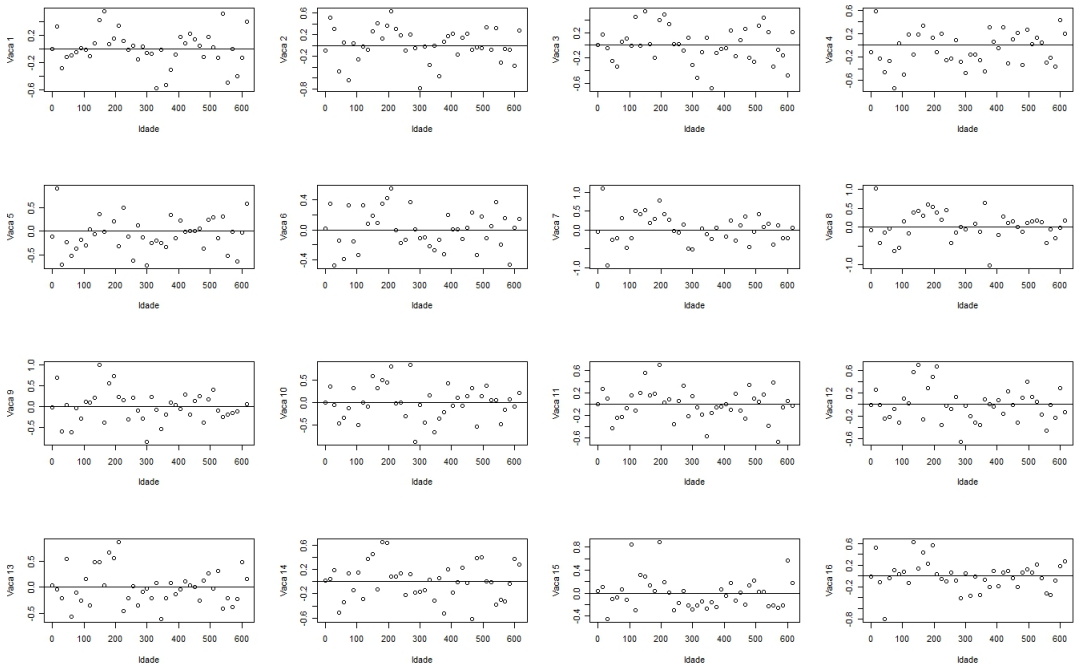

Figura 6 - Gráfico de resíduos do modelo Gompertz difásico ajustado as primeiras 16 vacas em estudo

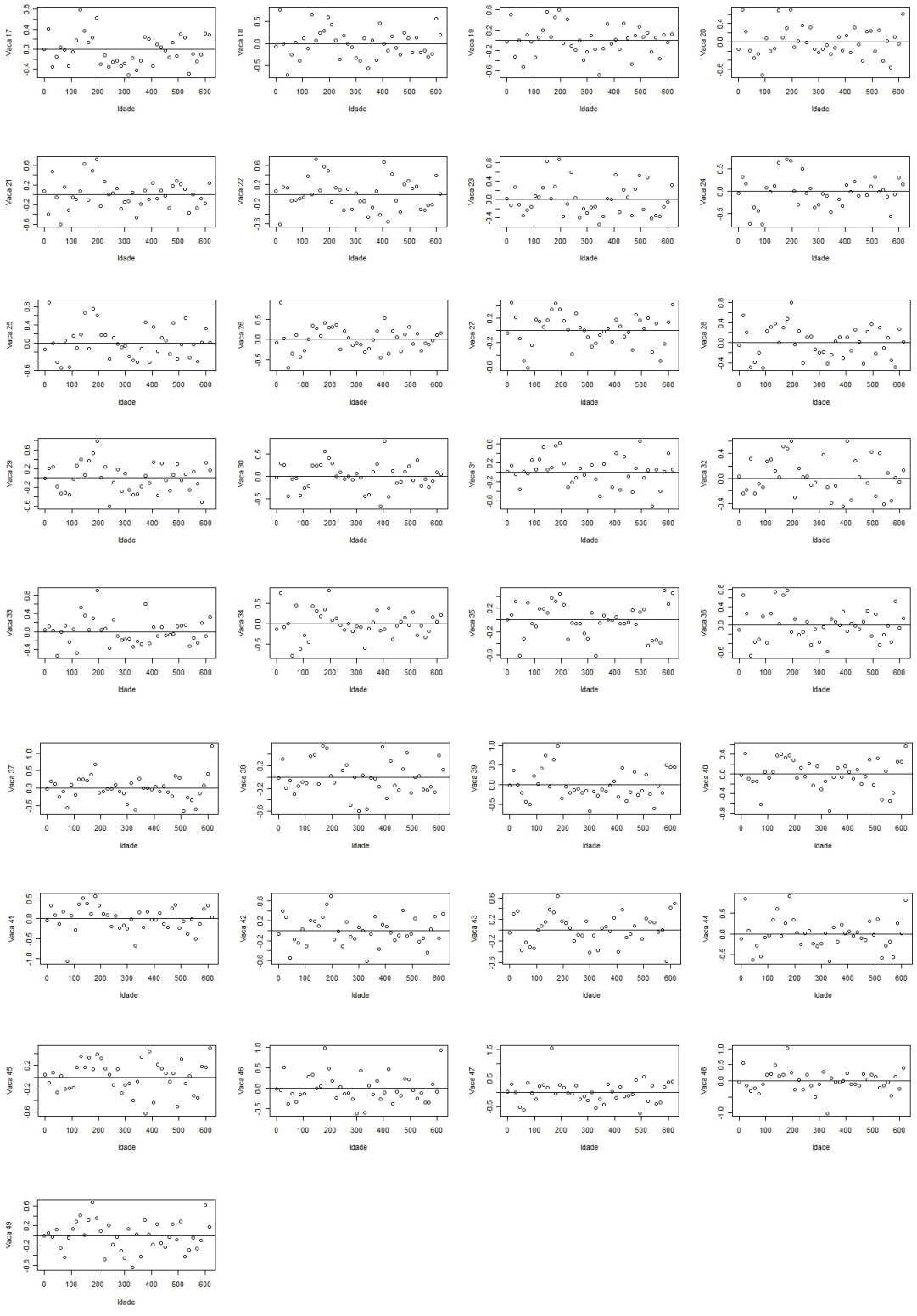

Figura 7 - Gráfico de resíduos do modelo Gompertz difásico ajustado as últimas 33 vacas em estudo

4.2 Análise descritiva das variáveis

Na Tabela 2 são fornecidas algumas estatísticas descritivas univariadas (média, desvio padrão, assimetria e curtose) acerca dos parâmetros do modelo Gompertz difásico com estrutura de erros autorregressiva de primeira ordem, que são imprescindíveis em qualquer estudo estatístico.

As médias elucidadas na Tabela 2, no geral, condizem com a interpretação prática de cada um dos parâmetros. Mendes (2007) ao ajustar as curvas individuais, com todas as pesagens, utilizando uma parametrização diferente do modelo Gompertz difásico com estrutura de erros autorregressiva de primeira ordem, encontrou resultados que superestimavam o peso adulto dos animais. Ainda, os desvios padrões não são numericamente grandes.

Tabela 2 - Estatísticas descritivas das estimativas dos parâmetros do modelo Gompertz difásico ponderado pelo inverso da variância dos pesos com estrutura de erros autorregressiva ajustado

Parâmetro	Média	Desvio Padrão	Assimetria	Curtose
A_1	101,374	15,189	0,283	-0,125
B_1	14,802	2,974	0,333	-0,133
K_1	0,033	0,007	0,251	-0,254
A_2	151,485	19,517	1,080	2,078
B_2	0,489	0,097	0,265	0,045
K_2	0,012	0,002	-0,027	-0,616
ϕ_1	0,843	0,073	0,312	1,441

Pode-se afirmar pela observação da Tabela 2 e pela Figura 8 que apenas a variável A_2 possui leve assimetria positiva, ou seja, as observações dessa variável estão concentradas, suavemente, à esquerda da distribuição. Com relação à curtose, novamente, apenas a variável A_2 apresenta um valor razoavelmente elevado, mostrando um certo acúmulo de observações em um determinado ponto de sua distribuição, ou seja, existem indícios de que as observações dessa variável estejam acumuladas do lado esquerdo da distribuição.

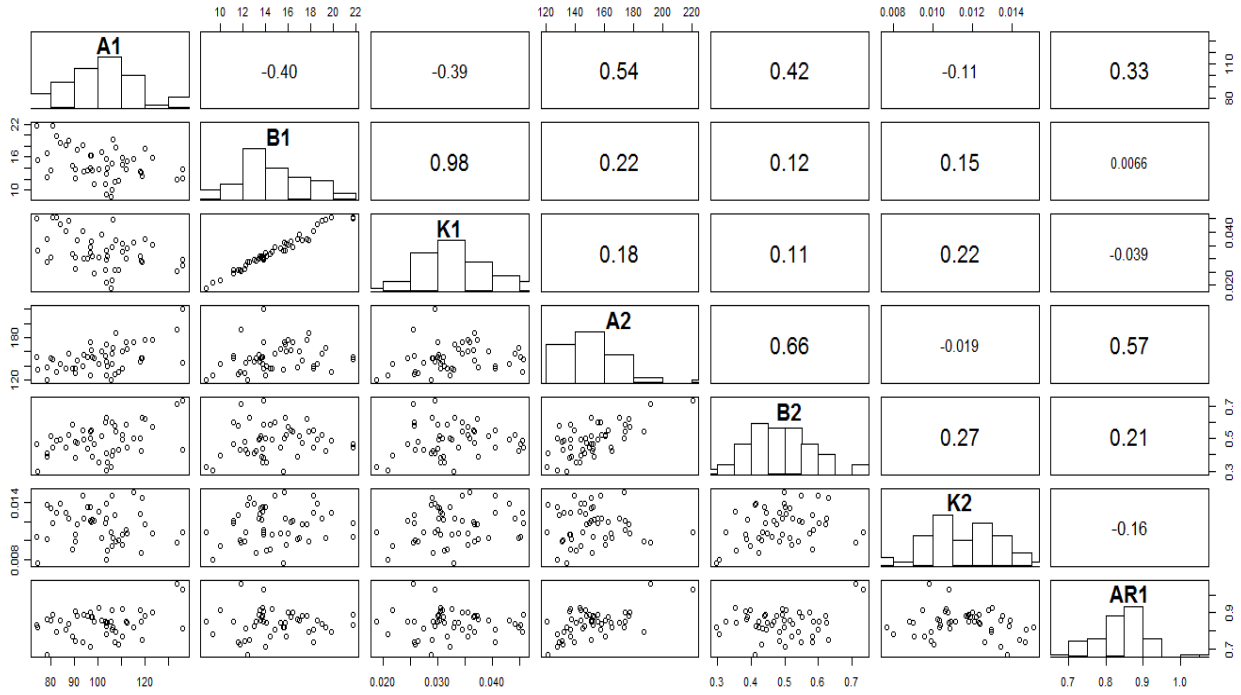

Figura 8 - Matriz de correlação das variáveis em estudo

Na Figura 8, apresenta-se a matriz de dispersão das variáveis em estudo. Valores acima da diagonal principal correspondem as correlações bivariadas; abaixo da diagonal principal observa-se os gráficos de dispersão; e a diagonal principal representa a distribuição de cada variável em estudo (histograma).

Observa-se que, em concordância com a Tabela 2, apenas a variável A_2 possui leve assimetria positiva. Nota-se ainda que as correlações não são numericamente elevadas, excetuando a correlação entre as variáveis B_1 e K_1 ($\rho = 0,98$) que possuem alta relação linear. Estes resultados corroboram as afirmações de Koops e Grossman (1991), que afirmam que as estimativas dos parâmetros de modelos multifásicos não possuem correlações muito fortes.

A normalidade das variáveis é estudada na Tabela 3 através do teste de Shapiro-Wilk (SHAPIRO; WILK, 1965), em que a hipótese nula testada refere-se a aceitação da normalidade da variável em estudo e a hipótese alternativa corresponde a não-normalidade da variável.

Tabela 3 - Teste de normalidade de Shapiro-Wilk para as estimativas dos parâmetros do modelo Gompertz difásico ajustado

Parâmetro	Estatística do teste	Valor p
A_1	0,977	0,448
B_1	0,981	0,618
K_1	0,971	0,267
A_2	0,940	0,015
B_2	0,987	0,872
K_2	0,985	0,787
ϕ_1	0,964	0,138

Pela Tabela 3, pode-se observar claramente que apenas a variável A_2 (valor p= $0,015$) não possui normalidade, pois à um nível de significância de 0,05, a hipótese de nulidade é rejeitada. Este fato está de acordo com o que foi encontrado na Tabela 2 e na Figura 8 anteriormente, em que foi constatada leve assimetria positiva (1,080) nesta variável.

4.3 Análise de componentes principais (preliminar)

Como a aplicação da análise de componentes principais não necessita da confirmação da distribuição gaussiana nas variáveis utilizadas, segue-se agora, com a aplicação dos métodos multivariados estudados.

Para a escolha do número k de componentes principais a serem retidos, será utilizado aqui, o critério de Kaiser (1958) e a observação do gráfico *scree*, além da observação da quantidade de variância explicada pelo número de componentes selecionados.

Pode-se observar na Tabela 4 que o primeiro, segundo e terceiro componentes principais explicam, respectivamente, 34,3%, 32,9% e 16,3% da variância original dos dados. Ainda pela Tabela 4, existem três autovalores maiores do que 1, logo, pelo critério de Kaiser ($\lambda_j \geq 1$), três componentes principais seriam retidas no sistema. Observando o gráfico *scree* (Figura 9) chega-se a mesma conclusão, pois a partir do terceiro componente principal a curva começa a se estabilizar. Retendo-se esses três componentes alcança-se um percentual de

explicação de 83,50% da variância original.

Tabela 4 - Autovalores e variância explicada dos componentes (preliminar)

Componente	Autovalor	Variância Explicada (%)	Variância Acumulada (%)
1	2,399	0,343	0,343
2	2,301	0,329	0,672
3	1,139	0,163	0,835
4	0,626	0,090	0,925
5	0,353	0,051	0,976
6	0,170	0,023	0,999
7	0,012	0,001	1,000

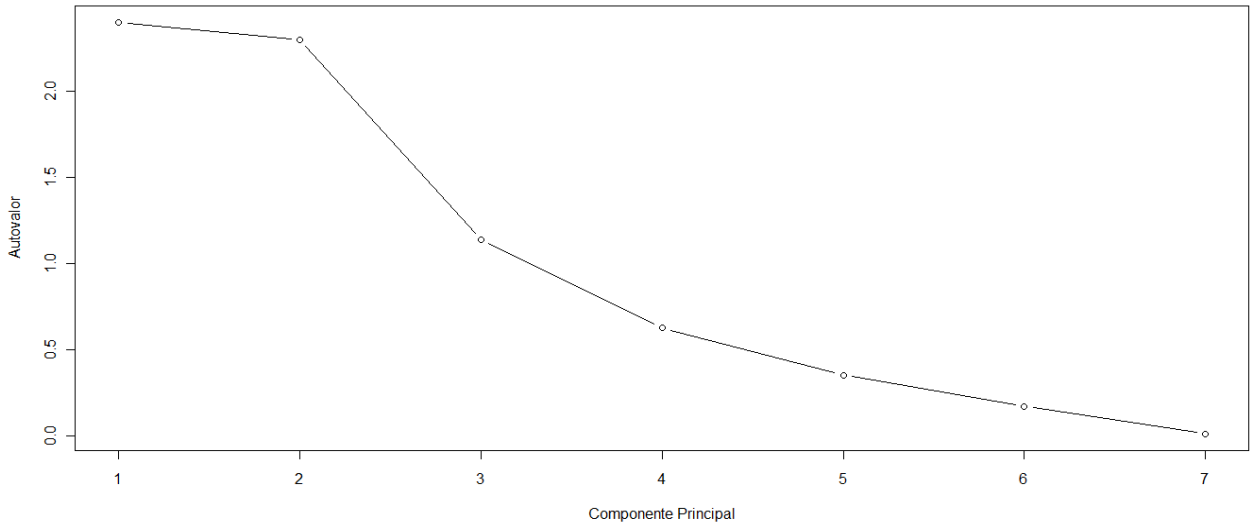

Figura 9 - *Scree-plot* dos componentes principais (preliminar)

De forma geral, Mardia et al. (1992) afirma que a retenção de componentes que expliquem mais de 70% da variância original de um conjunto de dados é plenamente satisfatória. Essa margem é refletida em diversas aplicações como, por exemplo, Pereira-da-Silva e Pezzato (2000), Sodré et al. (2007), Paiva et al. (2010), entre outros. Alguns autores trabalham com percentagens mais baixas como, por exemplo, Barbosa et al. (2006) que se dizem satisfeitos

com a retenção de três componentes principais que totalizam 60,65% da variância total original dos dados trabalhados por eles.

A Tabela 5 apresenta os coeficientes dos componentes principais retidos e, logo abaixo, entre parênteses, são apresentadas as correlações entre cada componente com cada variável inicial em estudo. Estudar estes coeficientes é de suma importância, pois como já elucidado anteriormente, eles auxiliam na nomeação de cada um dos componentes principais.

Tabela 5 - Autovetores e correlações entre componente principal e variável (preliminar)

	PC1	PC2	PC3
A_1	0,531	-0,177	0,101
	(0,823)	(-0,268)	(0,108)
B_1	-0,143	0,620	-0,193
	(-0,336)	(0,940)	(-0,206)
K_1	-0,164	0,619	-0,125
	(-0,271)	(0,939)	(-0,133)
A_2	0,542	0,278	-0,100
	(0,840)	(0,422)	(-0,107)
B_2	0,444	0,256	0,389
	(0,688)	(0,389)	(0,415)
K_2	-0,051	0,230	0,769
	(-0,079)	(0,349)	(0,821)
ϕ_1	0,422	0,076	-0,428
	(0,654)	(0,115)	(-0,457)

De acordo com a Tabela 5, as variáveis que mais contribuem para o primeiro componente principal são A_1 e A_2, cujos coeficientes são dados por, respectivamente, 0,531 (com correlação $r_{\hat{Y}_1,A_1} = 0,823$) e 0,542 (com correlação $r_{\hat{Y}_1,A_2} = 0,840$). Na ótica dos modelos não-lineares, esses parâmetros quando observados dentro do modelo Gompertz difásico com estrutura de erros autoregressiva de ordem 1, correspondem aos valores assintóticos em cada um dos ciclos do modelo, portanto esse componente será denominado como valor assintótico.

O segundo componente possui como maiores contribuidoras as variáveis B_1 e

K_1, cujos coeficientes são dados por 0,620 (com correlação $r_{\hat{Y}_2,B_1} = 0,940$) e 0,619 (com correlação $r_{\hat{Y}_2,K_1} = 0,939$). Como observado na seção de metodologia, a variável B_1 não possui interpretação prática quando observada como um parâmetro do modelo ajustado neste trabalho, logo resolveu-se utilizar o nome de taxa de crescimento da primeira fase, pois esta é a interpretação prática de K_1 no modelo Gompertz difásico.

As demais variáveis: B_2, K_2 e ϕ_1, são todas alocadas no último componente principal com coeficientes 0,389 (com correlação $r_{\hat{Y}_3,B_2} = 0,415$), 0,769 (com correlação $r_{\hat{Y}_3,K_2} = 0,821$) e -0,428 (com correlação $r_{\hat{Y}_3,\phi_1} = -0,457$), respectivamente. No modelo Gompertz difásico, assim como B_1, B_2 também não possui significado prático, K_2 é o parâmetro referente à taxa de crescimento da segunda fase e ϕ_1 é o parâmetro autorregressivo de primeira ordem. Logo, denominou-se este componente como taxa de crescimento da segunda fase corrigida pelo parâmetro autorregressivo.

Após a escolha de três componentes principais para explicação da variância original dos dados, foram calculados os escores referentes a cada um desses componentes (Tabela 6), por meio das seguintes equações:

1) Componente principal 1:

$$\hat{Y}_1 = 0,531A_1 - 0,143B_1 - 0,164K_1 + 0,542A_2 + 0,444B_2 - 0,051K_2 + 0,422\phi_1;$$

2) Componente principal 2:

$$\hat{Y}_2 = -0,177A_1 + 0,620B_1 + 0,619K_1 + 0,278A_2 + 0,256B_2 + 0,230K_2 + 0,076\phi_1;$$

3) Componente principal 3:

$$\hat{Y}_3 = 0,101A_1 - 0,193B_1 - 0,125K_1 - 0,100A_2 + 0,389B_2 + 0,769K_2 - 0,428\phi_1.$$

Como mencionado na seção de revisão bibliográfica, independentemente do valor do peso de uma determinada variável em um certo componente, ela será utilizada na obtenção dos escores. Por exemplo, pode-se observar que o peso da variável K_2 no primeiro componente principal é de somente $-0,051$, porém a variável também é utilizada na obtenção do escore da componente.

Tabela 6 - Escores de cada componente principal selecionado (preliminar)

Animal	PC1	PC2	PC3	Animal	PC1	PC2	PC3
1	-1,383	2,813	-1,161	28	-1,209	-1,658	0,163
2	1,151	0,913	0,708	29	-0,753	-3,581	-0,153
3	-0,515	0,984	-0,480	31	-0,105	-0,258	0,486
4	4,894	-0,516	-0,798	32	2,089	0,782	-0,450
5	5,452	0,803	-0,404	33	-0,873	-2,151	0,290
6	-2,365	-0,784	-2,434	35	-0,326	-0,633	-0,281
7	0,913	1,115	-0,214	36	-0,771	-0,836	-1,121
8	1,930	1,544	0,065	37	0,228	0,236	-0,408
10	-1,716	2,714	-1,435	38	-1,548	-0,391	1,784
11	0,947	1,304	-0,624	40	1,070	-1,059	0,887
12	-1,689	0,494	-1,406	41	0,493	0,841	1,949
13	0,465	1,099	-0,046	42	0,896	-1,814	0,189
14	1,647	-0,063	0,208	43	-2,103	2,308	0,246
15	0,566	-1,406	-1,200	44	1,022	-0,449	0,416
16	0,499	-2,656	-1,097	45	0,227	1,104	-0,816
17	0,412	-0,404	-0,668	46	-1,212	2,179	0,418
19	-0,489	0,070	0,790	47	-2,840	-1,154	1,881
20	-1,737	1,422	0,233	48	-0,665	2,381	-0,108
21	0,209	-1,486	0,450	49	0,764	1,062	1,543
22	-0,284	-1,464	1,919	50	-1,112	-3,647	-1,184
23	-0,714	-0,409	-0,215	51	0,826	-0,423	2,647
24	-0,649	-1,175	-1,489	52	-0,378	1,860	-1,200
25	0,550	0,050	-0,552	54	-0,411	-0,714	0,975
26	0,261	-0,242	0,137	55	-0,282	-0,504	-0,263
27	-1,384	1,798	1,827				

4.4 Análise biplot

Uma vez calculados os escores de cada componente principal, foi possível realizar a construção dos gráficos biplot bidimensionais (PC1×PC2, PC1×PC3 e PC2×PC3) como a seguir (Figura 10)

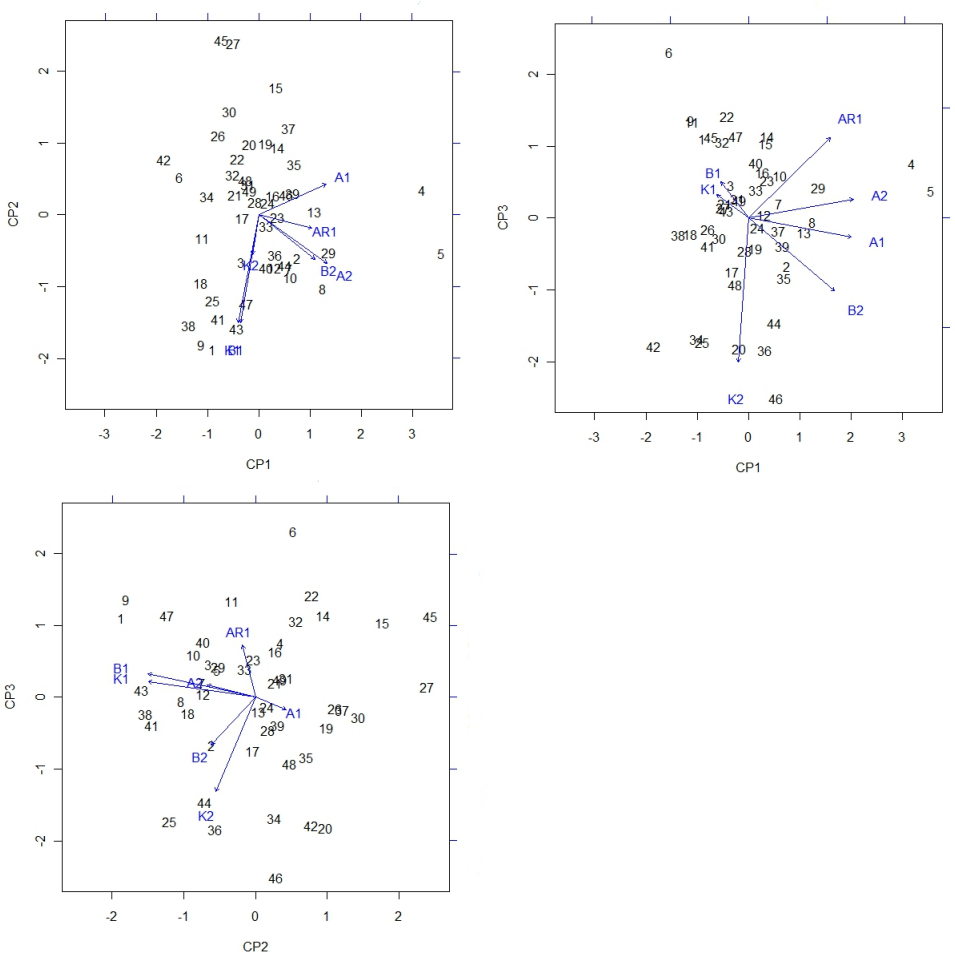

Figura 10 - Biplots bidimensionais dos três primeiros componentes principais

Pela Figura 10 pode-se observar que a informação contida na variável K_1 é a mesma da variável B_1 que não possui significado prático no modelo Gompertz ajustado, portanto decidiu-se nesta etapa retirá-la da análise. Da mesma maneira, a variável B_2 também possui a mesma explicação de outras variáveis e, portanto, também é retirada. Esses resultados obtidos a partir da observação do gráfico biplot, estão em concordância com as correlações observadas na Figura 8. Com a exclusão dessas duas variáveis, ou seja, com as cinco variáveis restantes (A_1, K_1, A_2, K_2, ϕ_1) realizou-se novamente a análise de componentes principais.

4.5 Análise de componentes principais (final)

Os mesmos critérios para a escolha do número k de componentes principais a serem retidos na análise de componentes principais preliminar serão utilizados aqui (critério de Kaiser (1958), observação do gráfico *scree* e observação da quantidade de variância explicada pelo número de componentes selecionados) observados na Tabela 7 e Figura 11.

Tabela 7 - Autovalores e variância explicada dos componentes (final)

Componente	Autovalor	Variância Explicada (%)	Variância Acumulada (%)
1	2,017	0,404	0,404
2	1,323	0,265	0,669
3	0,907	0,181	0,850
4	0,547	0,109	0,959
5	0,206	0,041	1,000

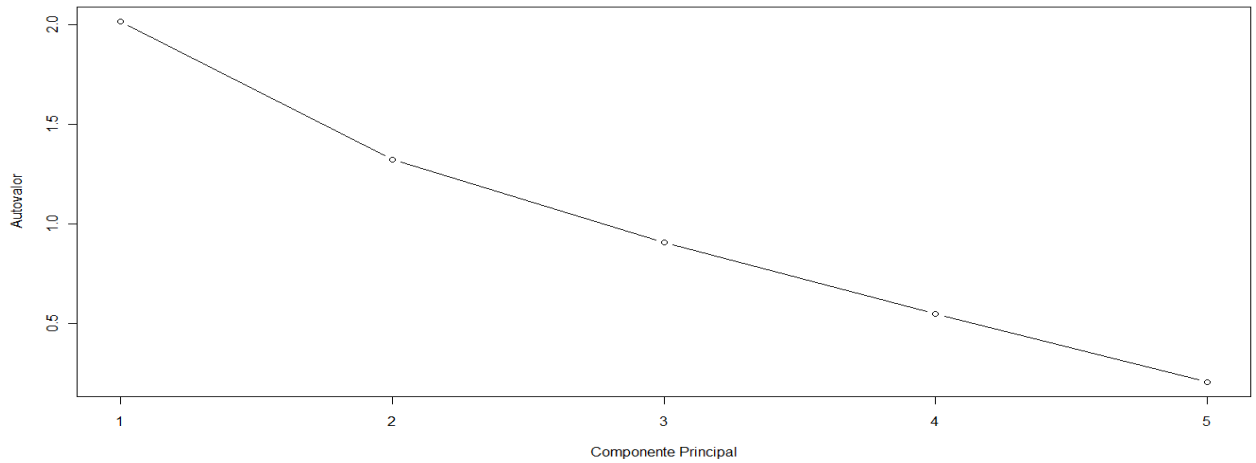

Figura 11 - *Scree-plot* dos componentes principais (final)

Observa-se que o *scree-plot* (Figura 11) é inconclusivo pois não possui um ponto específico onde os autovalores tendem a se estabilizar. Pelo prisma do critério de Kaiser (1958) retería-se no sistema os dois primeiros componentes principais com uma percentagem

de explicação da variância de $66,9\%$. Porém a inclusão de um terceiro componente principal, cujo autovalor é $0,907$, aumenta a percentagem de variância original explicada para 85%, que vai de acordo com o critério estabelecido por Mardia et al. (1992). Ainda, esse último componente carrega a informação de uma importante variável como será mostrado a seguir, logo decide-se aqui trabalhar com os três primeiros componentes principais obtidas.

A Tabela 8 apresenta os coeficientes dos componentes principais retidos e, logo abaixo, entre parênteses, são apresentadas as correlações entre cada componente com cada variável em estudo.

Tabela 8 - Autovetores e correlações entre o componente principal e a variável após a retirada das variáveis B_1 e B_2

	PC1	PC2	PC3
A_1	0,562	-0,199	0,406
	(0,798)	(-0,229)	(0,387)
K_1	-0,179	0,746	-0,326
	(-0,254)	(0,858)	(-0,310)
A_2	0,570	0,400	0,010
	(0,810)	(0,460)	(0,010)
K_2	-0,192	0,455	0,814
	(-0,273)	(0,523)	(0,775)
ϕ_1	0,538	0,195	-0,255
	(0,764)	(0,224)	(-0,243)

Pela Tabela 8 observa-se que alocaram-se no primeiro componente principal as variáveis A_1 e A_2, cujos coeficientes são dados por 0,562 (com correlação $r_{\hat{Y}_1,A_1} = 0,798$) e 0,570 (com correlação $r_{\hat{Y}_1,A_2} = 0,810$), respectivamente, que, como mencionado anteriormente, referem-se as assíntotas de cada uma das fases do modelo, além da variável referente ao processo autorregressivo do modelo ajustado ϕ_1, cujo coeficiente é dado por 0,538 (com correlação $r_{\hat{Y}_1,\phi_1} = 0,764$).

No segundo componente principal alocou-se apenas a variável K_1 com coeficiente dado por 0,746 (com correlação $r_{\hat{Y}_2,K_1} = 0,858$), cujo significado prático refere-se à taxa de crescimento do animal na primeira fase do modelo.

O terceiro e último componente principal retido no sistema tem como principal variável influente K_2 cujo coeficiente é dado por 0,814 (com correlação $r_{\hat{Y}_3,K_2} = 0,775$) que denota à taxa de crescimento do animal na segunda fase do modelo ajustado. Como mencionado anteriormente, a retenção do terceiro componente principal, mesmo que não necessária segundo os critérios de Kaiser e da observação do *scree-plot* é de grande importância uma vez que reteve a variável referente à taxa de crescimento da segunda fase do modelo ajustado.

Desta forma, os rótulos dados aos componentes principais retidos no sistema e suas respectivas equações são dadas por:

1) Primeiro componente principal: peso assintótico. O cálculo de seus escores é dado como

$$\hat{Y}_1 = 0,562A_1 - 0,179K_1 + 0,570A_2 - 0,192K_2 + 0,538\phi_1,$$

2) Segundo componente principal: taxa de crescimento referente à primeira fase do modelo. O cálculo de seus escores é dado como

$$\hat{Y}_2 = -0,199A_1 + 0,746K_1 + 0,400A_2 + 0,455K_2 + 0,195\phi_1;$$

3) Terceiro componente principal: taxa de crescimento referente à segunda fase do modelo. O cálculo de seus escores é dado como

$$\hat{Y}_3 = 0,406A_1 - 0,326K_1 + 0,010A_2 + 0,814K_2 - 0,255\phi_1.$$

Utilizando as equações apresentadas anteriormente, foram calculados os escores dos três componentes principais retidos a partir das cinco variáveis restantes (Tabela 9). Por meio dos escores obtidos, pode-se afimar, por exemplo, que os animais 4 e 5 se destacam positivamente com relação ao primeiro componente principal, ou seja, esses dois indivíduos possuem peso assintótico muito superior quando comparado aos demais. Por outro lado, o animal 47 se destaca negativamente com referência ao mesmo componente, ou seja, esse indivíduo possui baixo peso assintótico quando comparado aos demais animais em estudo.

Tabela 9 - Escores de cada componente principal selecionado após a retirada das variáveis B_1 e B_2

Animal	PC1	PC2	PC3	Animal	PC1	PC2	PC3
1	-0,865	1,868	-1,234	28	-1,368	-1,705	-0,391
2	0,551	0,582	0,387	29	-0,224	-2,422	0,466
3	-0,330	0,665	-0,522	31	-0,027	0,368	0,729
4	4,361	-0,282	-0,306	32	1,990	0,494	-0,093
5	4,812	0,901	0,150	33	-0,822	-2,009	0,255
6	-1,267	-1,032	-2,413	35	0,260	0,177	0,441
7	0,823	0,972	-0,189	36	-0,191	-0,486	-0,862
8	1,578	0,902	0,189	37	0,203	-0,323	-0,538
10	-1,269	1,466	-1,859	38	-2,040	-0,379	1,204
11	0,958	0,541	-0,598	40	0,398	-0,673	0,502
12	-1,066	0,250	-1,523	41	-0,293	0,951	1,584
13	0,271	0,873	-0,252	42	1,110	-1,449	0,988
14	1,271	-0,074	0,202	43	-2,118	1,520	-0,369
15	1,067	-1,323	-0,635	44	0,940	-0,020	0,837
16	0,993	-1,824	-0,534	45	0,333	0,639	-0,913
17	0,575	-0,460	-0,427	46	-1,298	1,857	0,012
19	-0,824	0,496	0,335	47	-3,134	-0,697	1,164
20	-1,605	1,032	-0,038	48	-0,741	1,793	-0,557
21	0,124	-0,748	0,532	49	0,503	1,444	1,893
22	-1,108	-1,359	1,335	50	-0,356	-2,945	-0,763
23	-1,130	-1,101	-1,151	51	-0,245	-0,218	2,257
24	0,126	-1,083	-1,009	52	0,200	1,217	-0,923
25	0,868	0,058	-0,097	54	-0,347	0,017	1,296
26	0,185	0,191	0,116	55	0,129	0,116	0,148
27	-1,959	1,225	1,174				

4.6 Análise de agrupamentos

Os três componentes principais, retidos pelos critérios de Kaiser (1958), análise do gráfico *scree* e análise da percentagem de variância original explicada das cinco variáveis selecionadas pela análise do gráfico biplot, foram utilizados como variáveis em uma análise de agrupamento a fim de se agregar grupos de animais homogêneos.

Metodologia similar foi utilizada em Rodrigues et al. (2002) e Moura et al. (2006) que utilizaram os escores obtidos através da análise de componentes principais como observações em uma análise de agrupamento e obtiveram ótimos resultados com seus objetivos propostos.

Como em Faria (2009), Fiorini et al. (2010) e Silveira (2010), para a aplicação da análise de agrupamento, calculou-se o número k ótimo de grupos a serem formados pelo método RMSSTD (Figura 12).

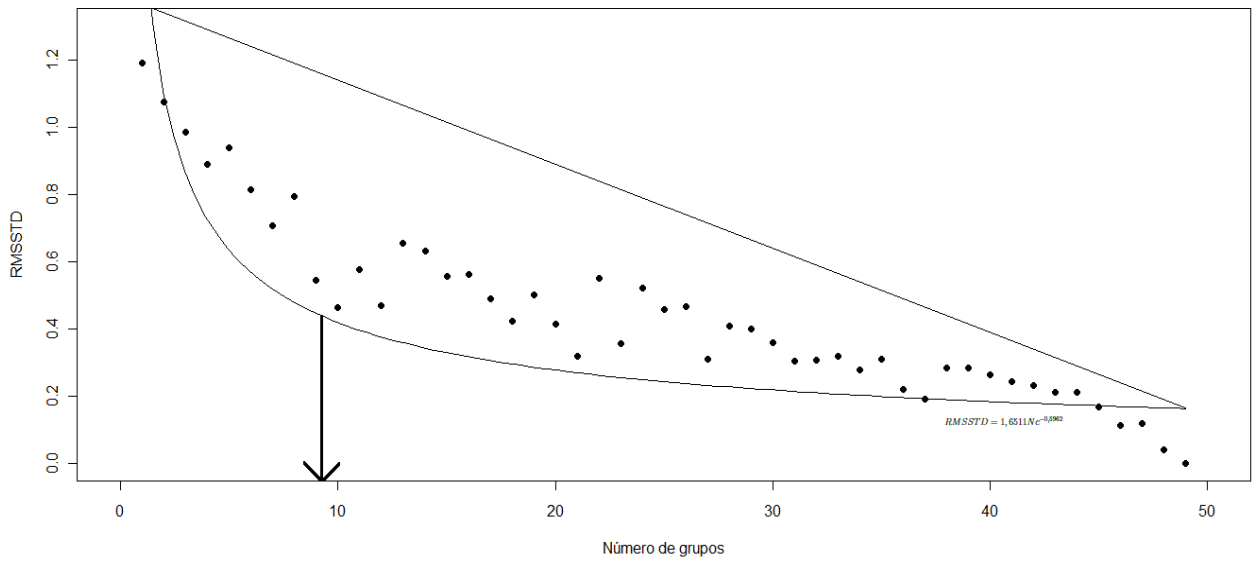

Figura 12 - Escolha do número de grupos pelo critério RMSSTD

Nota-se pelo gráfico referente ao RMSSTD (Figura 12) que o número k ótimo de grupos a serem formados é de 9 (indicado pela seta). Porém, como elucidado na seção de metodologia, não existe um método robusto para escolha do número ótimo de grupos, ficando a cargo do pesquisador agregar objetos homogêneos dentro de determinados grupos.

Com a construção do dendrograma (Figura 13), notou-se que a formação de apenas cinco grupos distintos era suficiente para dividir os animais em categorias diferentes passíveis de interpretação prática.

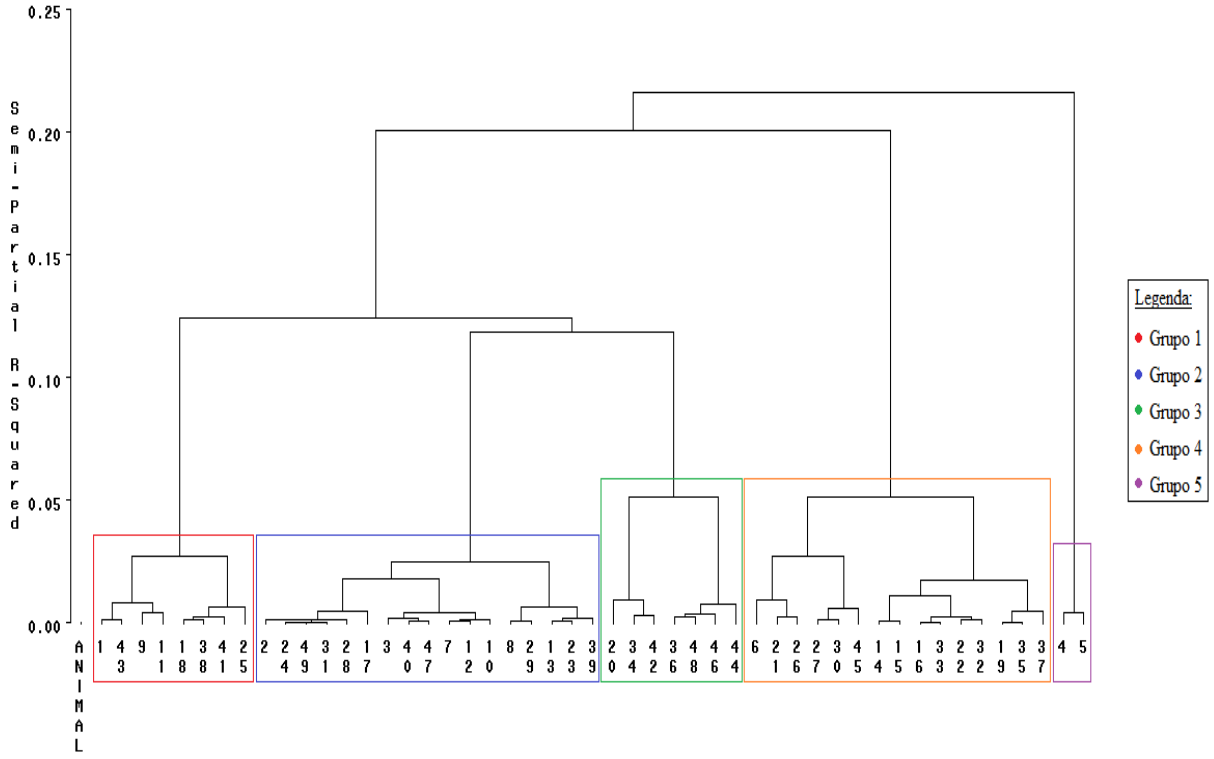

Figura 13 - Dendrograma obtido referente ao agrupamento dos animais em estudo

Observa-se pelo dendrograma construído (Figura 13) que a análise de agrupamentos separou os 49 animais em estudo em cinco grupos distintos homogêneos:

O primeiro grupo, composto por oito animais, identificados pelos números 1, 9, 11, 18, 25, 38, 41 e 43; tem como características principais baixo peso assintótico com alta taxa de crescimento na primeira fase e crescimento moderado na segunda parte do modelo quando comparados aos demais animais em estudo (Tabela 10).

Os grupo de número 2, composto por dezessete animais, identificados pelos números 2, 3, 7, 8, 10, 12, 13, 17, 23, 24, 28, 29, 31, 39, 40, 47, 49; tem como características peso assintótico e taxas de crescimento da primeira e segunda fases moderados quando comparado as demais vacas (Tabela 11).

O terceiro grupo, composto por sete animais, identificados pelos números 20, 34,

36, 42, 44, 46 e 48; é formado por animais que possuem moderado peso assintótico, taxa de crescimento da primeira fase do modelo moderada e alta taxa de crescimento na segunda fase do modelo quando comparados com os demais animais (Tabela 12).

O quarto grupo, composto por quinze animais, identificados pelos números 6, 14, 15, 16, 19, 21, 22, 26, 27, 30, 32, 33, 35, 37 e 45; agrega animais com peso assintótico moderado e taxas de crescimento em ambas as fases baixas quando comparados aos demais (Tabela 13).

O quinto e último grupo, composto por apenas dois animais, identificados pelos números 4 e 5, agrupa animais com alto peso assintótico e taxas de crescimento moderadas nas duas fases do modelo multifásico ajustado quando comparado aos demais animais em estudo (Tabela 14).

Desta forma, pôde-se ajustar cinco curvas distintas para os grupos formados a partir da média das 49 vacas (Figura 14). Observa-se que, de fato, essas curvas médias atendem as necessidade de diferentes perfis de crescimento, diferentemente da curva média traçada na Figura 5.

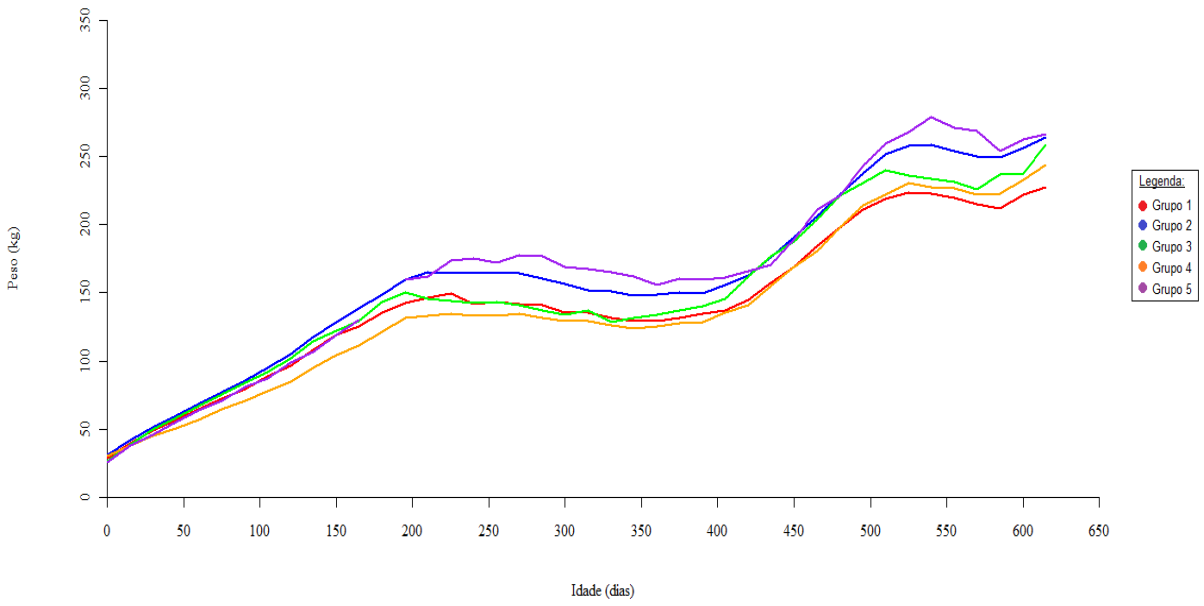

Figura 14 - Curvas médias dos cinco grupos de animais formados por meio da análise de agrupamento hierárquico pelo método de Ward

Com a análise pela Figura 14 e das Tabelas de 10 à 14, pode-se notar que, de um modo geral, os grupos 2, 3 e 4 são compostos por animais com perfil de crescimento intermediário. Os animais do primeiro grupo são os menos interessantes, pois possuem perfil de crescimento ruim quando comparado aos demais e, finalmente, o grupo de número 5, composto por apenas dois animais, é o que possui melhor perfil de crescimento.

Pela observação da Figura 14 e das Tabelas de 10 a 14, pode-se afirmar que o grupo criado de maior destaque com relação ao ganho de peso é o grupo 5 (Tabela 14), composto pelos animais de número 4 e 5, pois este grupo possui como característica principal um peso assintótico extremamente superior às demais vacas em estudo, sendo que o peso assintótico da primeira fase do perfil de crescimento médio deste grupo é dado por $A_1 = 134,49$ e o peso médio assintótico dos animais é de $A_1 + A_2 = 340,60$.

Tabela 10 - Estimativas dos parâmetros do modelo dos animais pertencentes ao grupo 1

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1	$A_1 + A_2$
1	80,85	21,68	0,046	149,52	0,44	0,012	0,89	230,37
9	73,92	21,71	0,045	152,01	0,46	0,010	0,84	225,93
11	78,58	16,79	0,037	137,44	0,39	0,010	0,87	216,02
18	86,35	18,18	0,040	136,80	0,44	0,013	0,81	223,15
25	93,73	18,22	0,040	138,96	0,55	0,015	0,74	232,69
38	82,37	19,83	0,046	131,89	0,48	0,013	0,80	214,26
41	83,95	18,58	0,043	143,30	0,50	0,014	0,86	227,25
43	87,55	18,96	0,044	158,08	0,52	0,012	0,84	245,63
Média	83,41	19,25	0,043	143,50	0,47	0,012	0,83	226,91

Com relação à taxa de crescimento animal, o primeiro grupo (Tabela 10), formado pelos animais 1, 9, 11, 18, 25, 38, 41 e 43 e o grupo 3 (Tabela 12), formado pelos animais 20, 34, 36, 42, 44, 46 e 48; são os grupos que mais se destacam uma vez que possuem alta taxa de crescimento na primeira fase do modelo (média de K_1 do grupo 1 é de $0,043$) e alta taxa de crescimento na segunda fase do modelo (média de K_2 do grupo 3 é de 0,014), respectivamente.

Tabela 11 - Estimativas dos parâmetros do modelo dos animais pertencentes ao grupo 2

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1	$A_1 + A_2$
2	103,38	15,72	0,033	170,38	0,63	0,012	0,84	273,76
3	91,16	17,25	0,037	147,48	0,48	0,012	0,88	238,64
7	96,76	16,40	0,034	173,96	0,54	0,012	0,90	270,72
8	119,77	17,61	0,037	176,25	0,62	0,012	0,88	296,02
10	107,49	17,79	0,037	186,59	0,54	0,010	0,79	294,07
12	97,25	16,25	0,037	162,11	0,55	0,012	0,88	259,37
13	111,95	13,81	0,030	174,28	0,59	0,011	0,85	286,23
17	80,10	13,58	0,030	150,54	0,51	0,013	0,86	230,65
23	112,41	15,28	0,034	164,41	0,45	0,011	0,85	276,83
24	93,85	13,42	0,030	155,45	0,49	0,012	0,89	249,30
28	95,72	13,70	0,030	152,22	0,44	0,014	0,88	247,94
29	123,05	15,96	0,035	177,19	0,57	0,011	0,90	300,24
31	100,47	13,73	0,030	142,88	0,35	0,013	0,93	243,35
39	117,91	13,44	0,031	146,97	0,50	0,013	0,92	264,88
40	101,63	16,96	0,039	160,18	0,51	0,010	0,86	261,81
47	106,22	19,30	0,045	165,25	0,42	0,010	0,82	271,47
49	97,58	13,64	0,031	152,72	0,39	0,012	0,88	250,30
Média	103,34	15,52	0,034	162,29	0,51	0,012	0,87	265,62

As conclusões realizadas acerca de cada um dos grupos foram possíveis após a realização da análise de agrupamentos pelo método hierárquico de Ward. Conclusões essas que se confirmam com a observação do gráfico das curvas médias de cada um dos cinco grupos formados pelas 49 vacas em estudo (Figura 14) confrontada com as informações contidas nas Tabelas de 10 à 14.

Tabela 12 - Estimativas dos parâmetros do modelo dos animais pertencentes ao grupo 3

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1	$A_1 + A_2$
20	107,25	11,59	0,026	129,06	0,58	0,013	0,74	236,31
34	96,62	13,96	0,032	127,44	0,50	0,013	0,71	224,06
36	105,91	14,82	0,034	156,92	0,60	0,015	0,81	262,83
42	78,54	12,39	0,029	121,19	0,41	0,014	0,81	199,73
44	114,95	15,63	0,036	173,41	0,50	0,015	0,82	288,36
46	118,67	12,53	0,029	150,68	0,63	0,014	0,75	269,35
48	103,36	13,02	0,030	145,88	0,41	0,014	0,83	249,24
Média	103,62	13,42	0,031	143,51	0,52	0,014	0,78	247,13

Tabela 13 - Estimativas dos parâmetros do modelo dos animais pertencentes ao grupo 4

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1	$A_1 + A_2$
6	74,33	15,57	0,033	134,14	0,30	0,008	0,82	208,47
14	118,37	13,16	0,029	151,21	0,42	0,009	0,85	269,58
15	105,91	9,98	0,021	143,29	0,40	0,009	0,92	249,20
16	110,42	14,59	0,032	136,34	0,49	0,011	0,92	246,76
19	98,26	11,19	0,025	150,90	0,47	0,012	0,85	249,16
21	89,37	14,39	0,032	136,46	0,58	0,009	0,77	225,83
22	104,13	14,06	0,031	140,47	0,35	0,009	0,85	244,6
26	90,57	12,19	0,026	130,66	0,48	0,010	0,75	221,24
27	105,47	8,77	0,019	120,17	0,32	0,011	0,85	225,64
30	108,88	11,71	0,026	131,07	0,43	0,010	0,73	239,95
32	90,40	13,82	0,031	135,94	0,38	0,011	0,91	226,34
33	110,26	15,80	0,035	160,54	0,52	0,010	0,77	270,80
35	103,29	11,18	0,026	155,07	0,60	0,012	0,85	258,36
37	135,96	12,20	0,027	145,17	0,43	0,011	0,81	281,13
45	103,39	9,29	0,021	127,65	0,31	0,008	0,78	231,04
Média	103,27	12,53	0,028	139,94	0,43	0,010	0,83	243,21

Tabela 14 - Estimativas dos parâmetros do modelo dos animais pertencentes ao grupo 5

Animal	A_1	B_1	K_1	A_2	B_2	K_2	ϕ_1	$A_1 + A_2$
4	133,33	11,86	0,026	191,55	0,71	0,010	1,06	324,88
5	135,65	13,78	0,030	220,67	0,73	0,011	1,03	356,32
Média	134,49	12,82	0,028	206,11	0,72	0,010	1,04	340,60

5 CONCLUSÕES

A utilização da análise de agrupamentos, tendo como variáveis os componentes principais extraídos das estimativas dos parâmetros do modelo Gompertz difásico com estrutura de erros autorregressiva de ordem 1, permitiu alocar vacas com características similares em 5 grupos distintos.

No que se refere ao objetivo proposto pelo trabalho, ou seja, agrupar animais de acordo com as estimativas dos parâmetros de suas respectivas curvas de crescimento, notou-se que a utilização de parâmetros do modelo não-linear ajustado auxiliou de forma positiva no agrupamento dos mesmos, pois esses parâmetros possuem interpretação biológica e esse fato foi refletido nos agrupamentos obtidos, que separaram os animais em estudo, identificando os animais com maiores/menores pesos assintóticos e taxas de crescimento. Portanto, o objetivo proposto pelo trabalho, ou seja, agrupar animais com características similares foi alcançado por meio da metologia proposta.

REFERÊNCIAS

ARAÚJO, L. B. **Seleção e análise dos modelos PARAFAC e Tucker e gráfico triplot com aplicação em interação tripla**. 2009. 111 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2009.

ASSOCIAÇÃO BRASILEIRA DE HEREFORD E BRAFORD. **Hereford - Carne de qualidade tipo exportação**. Disponível em: <http://www.hereford.com.br/?bW9kdWxvPTEmbWVudT0xJmFycXVpdm89Y29udGV1ZG8ucGhwICAgICAgICAgICAgICAgICAg>. Acesso em: 13 de fevereiro de 2011.

BARBOSA, L.; LOPES, P. S.; REGAZZI, A. J.; GUIMARÃES, S. E. F.; TORRES, R. A. Avaliação de características de qualidade da carne de suínos por meio de componentes principais. **Revista Brasileira de Zootecnia**, Viçosa, v. 35, n. 4, p. 1693-1645, 2006.

BASILEVSKY, A. **Statistical factor analysis and related methods**: theory and applications. New York: John Wiley & Sons, 1994. 737 p.

BATES, D. M.; WATTS, D. G. Relative curvature measures of nonlinearity (with discussion). **Journal of the Royal Statistical Society**, New York, v. 42, n. 1, p. 1-25, 1980.

_____. **Nonlinear regression analysis and its applications**. New York: John Wiley & Sons, 1988. 365 p. Wiley Series in Probability and Mathematical Statistics.

CATTELL, R. B. The scree test for the number of factors. **Multivariate Behavioural Research**, London, v. 1, n. 2, p. 245-276, 1966.

CURRIE, D. J. Estimating Michaelis-Menten parameters: bias, variance and the experimental design. **Biometrics**, Washington, v. 38, n. 4, p. 907-919, 1982.

DURBIN, J.; WATSON, G. S. Testing for serial correlation in least squares regression I. **Biometrika**, Oxford, v. 37, p. 409-428, 1950.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3rd ed. New York: John Wiley & Sons, 1998. 706 p.

ECKART, C.; YOUNG, G. A. principal axis transformation for non-hermitian matrices. **Bulletin of the American Mathematical Society**, London, v. 45, n. 2, p. 118-121, 1939.

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

FERREIRA, D. F. **Estatística multivariada**. Lavras: Ed. UFLA, 2008. 662 p.

FIORINI, C. V. A.; SILVA, D. J. H.; SILVA, F. F.; MIZUBUTI, E. S. G.; ALVES, D. P.; CARDOSO, T. S. Agrupamento de curvas de progresso de requeima, em tomateiro originado de cruzamento interespecífico. **Pesquisa Agropecuária Brasileira**, Brasília, v. 45, n. 10, p. 1095-1101, 2010.

FIRETTI, R.; NAKAMURA, L. R.; OLIVEIRA, E. C.; CARVALHO FILHO, A. A. Agrupamentos de municípios da 10ª Região Administrativa do estado de São Paulo em função dos vínculos empregatícios rurais e urbanos. In: Congresso da Sociedade Brasileira de Economia, Administração e Sociologia Rural, 48., 2010, Campo Grande. **Anais...** Campo Grande: UCDB, 2010. p. 1-20.

FITZHUGH JR., H. A. Analysis of Growth Curves and Strategies for Altering Their Shape. **Journal of Animal Science**, Champaign, v. 42, n. 4, p. 1036-1051, 1976.

FREITAS, A. R. Curvas de crescimento na produção animal. **Revista Brasileira de Zootecnia**, Viçosa, v. 34, n. 3, p. 786-795, 2005.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometrika**, Oxford, v. 58, n. 3, p. 453-467, 1971.

_____. Least squares approximation of matrices by additive and multiplicative models. **Journal of the Royal Statistical Society**, New York, v. 40, n. 2, p. 186-196, 1978.

GALLANT, A. R. **Nonlinear statistical models**. New York: John Wiley & Sons, 1987. 610 p.

GOLUB, G. H.; REINSCH, C. Singular value decomposition and least squares solutions. **Numerische Mathematik**, New York, v. 14, n. 5, p. 403-420, 1970.

GOWER, C.; HAND, D. J. **Biplots**. London: Chapman & Hill, 1996. 152 p.

GREENACRE, M. **Biplots in practice**. Bilbao: Fundación BBVA, 2010. 219 p.

GUEDES, M. F.; MUNIZ, J. A.; PEREZ, J. R. O.; SILVA, F. F.; AQUINO, L. H.; SANTOS, C. L. Estudo das curvas de crescimento de cordeiros das raças Santa Inês e Bergamácia considerando heterogeneidade de variâncias. **Ciência e Agrotecnologia**, Lavras, v. 28, n. 2, p. 381-388, 2004.

GROSSMAN, M.; KOOPS, W. J. Multiphasic analysis of growth curves in chickens. **Poultry Science**, Stanford, v. 67, n. 1, p. 33-42, 1988.

HAIR, J. F.; TATHAM, R.L.; ANDERSON, R.E.; BLACK, W.C. **Análise multivariada de dados**. Tradução de Adonai Schlup Sant'Anna e Anselmo Chaves Neto. 5.ed. Porto Alegre: Bookman, 2005. 593 p.

HARTLEY, H. O., The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. **Technometrics**, Alexandria, v. 3, n. 2, p. 269-280, 1961.

HOFFMAN, R.; VIEIRA, S. **Análise de regressão**: uma introdução à econometria. 3.ed. São Paulo: Hucitec, 1998. 379 p.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. New Jersey: Prentice Hall, 1988. 320 p.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 6th ed. New Jersey: Prentice Hall, 2007. 773 p.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, New York, v. 23, n. 3, p. 187-200, 1958.

KOOPS, W. J. Multiphasic growth curve analysis. **Growth Development & Aging**, Bar Harbor, v. 50, n. 2, p. 169-177, 1986.

KOOPS, W. J.; GROSSMAN, M. Multiphasic growth and allometry. **Growth Development & Aging**, Bar Harbor, v. 55, n. 3, p. 203-212, 1991.

LAIRD, A. K.; HOWARD, A. Growth curves in inbred mice. **Nature**, London, v. 213, n. 5078, p. 786-788, 1967.

LARA, I.A.R.; CORRÊA, A.M.C.J; DIAS, C.T.S. Perfil da desigualdade entre as pessoas ocupadas na agricultura brasileira: uma abordagem multivariada. **Cadernos da FACECA**, Campinas, v.14, n.2, p. 149-155. 2005.

LAVORANTI, O. J.; DIAS, C. T. S.; VENCOVSKY, R. Estudo da adaptabilidade e estabilidade fenotípica de progênies de *Eucalyptus grandis* via metodologia AMMI. **Boletim de Pesquisa Florestal**, Colombo, v. 44, n.44, p. 107-124, 2002.

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

MANLY, B. F. J. **Métodos estatísticos multivariados**. Tradução de Sara Ianda Correa Carmona. 3.ed. Porto Alegre: Bookman, 2008. 229 p.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Academic Press, 1992. 518 p.

MARQUARDT, D. W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. **Journal of the Society for Industrial and Applied Mathematics**, Philadelphia, v. 11, n. 2, p. 431-441, 1963.

MAZZINI, A. R. A.; MUNIZ, J. A.; AQUINO, L. H; SILVA, F. F. Análise da curva de crescimento de machos Hereford. **Ciência e Agrotecnologia**, Lavras, v. 27, n. 5, p. 1105-1112, 2003.

MAZZINI, A. R. A.; MUNIZ, J. A.; SILVA, F. F.; AQUINO, L. H. Curva de crescimento de novilhos Hereford: heterocedasticidade e residuos autorregressivos. **Ciência Rural**, Santa Maria, v. 35, n. 2, p. 422-427, 2005.

MENDES, P. N. **Curvas de crescimento difásicas de fêmeas Hereford com erros autoregressivos e heterogeneidade de variâncias**. 2007. 84 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2007.

MENDES, P. N.; MUNIZ, J. A.; SILVA, F. F.; MAZZINI, A. R. A.; SILVA, N. A. M. Análise da curva de crescimento difásica de fêmeas Hereford por meio da função não linear de Gompertz. **Ciência Animal Brasileira**, Goiânia, v. 10, n. 2, p. 454-461, 2009.

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

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 4th ed. New York: Wiley, 2006. 621 p.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. 2.ed. São Paulo: Editora Edgarg Blucher, 2004.

MOURA, M. C. S.; LOPES, A. N. C.; MOITA, G. C. M.; NETO, J. M. M. Estudo multivariado de solos urbanos da cidade de Teresina. **Química Nova**, São Paulo, v. 29, n. 3, p. 429-435, 2006.

NELDER, J. A. The fitting of a generalization of the logistic curve. **Biometrics**, Washington, v. 17, n. 1, p. 89-110, 1961.

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

PASTERNAK, H.; SHALEV, B. A. The effect of a feature of regression disturbance on the efficiency of fitting growth curves. **Growth, Development & Aging**, Bar Harbor, v. 58, n. 1, p. 33-39, 1994.

PAZ, C. C. P.; FREITAS, A. R.;PACKER, I. U.; TAMBASCO-TALHARI, D.; REGITANO, L. C. A; ALENCAR, M. M. Ajuste de modelos não-lineares em estudos de associação entre polimorfismos genéticos e crescimento em bovino de corte. **Revista Brasileira de Zootecnia**, Viçosa, v. 33, n. 6, p. 1416-1425, 2004.

PEREIRA-DA-SILVA, E. M.; PEZZATO, L. E. Respostas da tilápia do nilo (*Oreochromis niloticus*) à atratividade e palatabilidade de ingredientes utilizados na alimentação de peixes. **Revista Brasileira de Zootecnia**, Viçosa, v. 29, n. 5, p. 1273-1280, 2000.

RATKOWSKY, D. A. **Nonlinear regression modeling, a unified practical approach**. New York: Marcel Dekker, 1983. 241 p.

RAWLINGS, J. O.; PANTULA, S. G.; DICKEY, D. A. **Applied regression analysis**. 2nd ed. New York: Springer, 1998. 659 p.

RODRIGUES, L. S.; ANTUNES, I. F.; TEIXEIRA, M. G.; SILVA, J. B. Divergência genética entre cultivares locais e cultivares melhoradas de feijão. **Pesquisa Agropecuária Brasileira**, Brasília, v. 37, n. 9, p. 1275-1284, 2002.

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

SAS INSTITUTE. **SAS/ETS User's Guide**. Version 6. 2nd ed. Cary: SAS Institute Inc., 1995.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, Oxford, v. 52, n. 3/4, p. 591-611, 1965.

SILVA, F, L.; ALENCAR, M. M.; FREITAS, A. R.; PACKER, I. U.; MOURÃO, G. B. Curvas de crescimento em vacas de corte de diferentes tipos biológicos. **Pesquisa Agropecuária Brasileira**, Brasília, v. 46, n. 3, p. 262-271, 2011.

SILVA, N. A. M.; AQUINO, L. H.; SILVA, F. F.; OLIVEIRA, A. I. G. Curvas de crescimento e influência de fatores não-genéticos sobre as taxas de crescimento de bovinos da raça Nelore. **Ciência e Agrotecnologia**, Lavras, v. 28, n. 3, p. 647-654, 2004.

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

SODRÉ, G. S.; MARCHINI, L. C.; MORETI, A. C. C. C.; OTSUK, I. P.; CARVALHO, C. A. L. Caracterização físico-química de amostras de méis de *Apis millifera L.* (Hymenoptera: Apidae) do estado do Ceará. **Ciência Rural**, Santa Maria, v. 37, n. 4, p. 1139-1144, 2007.

SOUZA, G. S. **Introdução aos modelos de regressão linear e não-linear**. Brasília: Embrapa, 1998. 505 p.

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

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, New York, v. 58, n. 301, p. 236-244, 1963.

WINSOR, C. P. The Gompertz curve as a growth curve. **Proceedings of the National Academy of Science**, Stanford, v. 18, n. 1, p. 1-8, 1932.

APÊNDICES

APÊNDICE A - Programação desenvolvida no *SAS*

```
DATA DISSERTACAO;
INPUT IDADE  V1 V2 ... V55 INVAR @@;
DATALINES;
0 31.5 26 ... 35 0.068273508
15 42 38 39 ... 45 0.0537985
.
.
615 231 273 ... 250 0.00121417
;
/*Ajuste do modelo Gompertz Difásico com estrutura de erros
 autorregressivos de ordem 1*/
PROC MODEL DATA=DISSERTACAO MAXITER=32000;
V1 = A1*EXP(-EXP(B1-K1*IDADE))+A2*EXP(-EXP(B2-K2*IDADE));
PARMS A1=140 B1=3 K1=0.0001 A2=100 B2=2 K2=0;
FIT V1 / OUTALL OUT=SAIDA DW DWPROB PRL=WALD OUTEST=PRED;
%AR(V1,1);
WEIGHT INVAR;
RUN;

/*Análise de componentes principais*/
DATA DISSERTACAO_PCA;
INPUT ANIMAL $ A1   B1  K1  A2  B2  K2  AR1;
DATALINES;
v1 80.84703 21.68257 0.045582 149.5194 0.442716 0.01181 0.887181
v2 103.3829 15.71899 0.033057 170.3789 0.624644 0.01231 0.844044
.
.
.
v55 97.58026 13.63745 0.031027 152.7155 0.387908 0.01217 0.876333
```

```
;
PROC PRINCOMP DATA=DISSERTACAO_PCA OUT=SCORES;
VAR A1 B1  K1  A2  B2 K2 AR1;
RUN;

/*Análise de Agrupamentos*/
DATA AGRUP_PCA;
INPUT ANIMAL $ Prin1 Prin2 Prin3;
DATALINES;
1 -0.865 1.868 -1.234
2 0.551 0.582 0.387
.
.
.
55 0.129 0.116 0.148
;
PROC CLUSTER DATA=AGRUP_PCA METHOD=WARD RMSSTD OUT=SAIDA1;
VAR Prin1 Prin2 Prin3;
ID ANIMAL;
RUN;
/*Plot do RMSSTD*/
DATA SAIDA2;
SET SAIDA1;
KEEP _NCL_ _RMSSTD_;
PROC SORT;
BY _NCL_;RUN;
PROC GPLOT DATA=SAIDA2;
PLOT _RMSSTD_*_NCL_;
RUN;
/*Ajuste da curva para escolha do k ótimo de grupos*/
```

```
PROC MODEL DATA=SAIDA2;
_RMSSTD_=a*_NCL_**(-b);
PARMS a=5 b=1;
fit _RMSSTD_/outall;
RUN;
/*Dendrograma*/
PROC TREE GRAPHICS;
ID ANIMAL;
RUN;
QUIT;
```

APÊNDICE B - Programação desenvolvida no R

```
# Gráfico das curvas ajustadas
dados = read.csv2("dados3.csv",h=T)
par(family = "serif")
plot(c(0,650,50),c(0,350,50), xlab= "Idade (dias)",
ylab="Peso (kg)", type = "n", axes=F)
axis(1, pos=0, at=seq(0,650,50))
axis(2, pos=0,  at=seq(0,350,50))
t<- seq(0,615,15)
for(i in 2:56)
{
lines(t, dados[,i],  col="black", type="l",lwd="1.8")
}
lines(t, media,  col="red", type="l",lwd="2.5")
# Construção do biplot
parametros_ = read.csv2("parametros.csv",h=T)
pca <- princomp(parametros_)
princomp(parametros_, cor = TRUE)
summary(pca <- princomp(parametros_, cor = TRUE))
par(mfrow=c(2,2))
biplot(pca,choices=c(1,2),col=c("black","blue"),
xlab="CP1",ylab="CP2",scale=1,pc.biplot=T,xlim=c(-3.5,3.5),
ylim=c(-2.5,2.5),arrow.len=0.05)
biplot(pca,choices=c(1,3),col=c("black","blue"),
xlab="CP1",ylab="CP3",scale=1,pc.biplot=T,xlim=c(-3.5,3.5),
ylim=c(-2.5,2.5),arrow.len=0.05)
biplot(pca,choices=c(2,3),col=c("black","blue"),
xlab="CP2",ylab="CP3",scale=1,pc.biplot=T,xlim=c(-2.5,2.5),
ylim=c(-2.5,2.5),arrow.len=0.05)
```