FILIPE RIBEIRO FORMIGA TEIXEIRA

GENÔMICA E MODELOS NÃO-LINEARES MISTOS NO AJUSTE DE CURVAS DE LACTAÇÃO DE BOVINOS DA RAÇA GIROLANDO

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

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

T266g
2018

Teixeira, Filipe Ribeiro Formiga, 1989-
Genômica e modelos não-lineares mistos no ajuste de curvas de lactação de bovinos da raça Girolando / Filipe Ribeiro Formiga Teixeira. - Viçosa, MG, 2018.
xii, 99 f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Moysés Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Estatística matemática. 2. Genômica. 3. Teoria bayesiana de decisão estatística. 4. Genética molecular. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

FILIPE RIBEIRO FORMIGA TEIXEIRA

GENÔMINCA E MODELOS NÃO-LINEARES MISTOS NO AJUSTE DE CURVAS DE LACTAÇÃO DE BOVINOS DA RAÇA GIROLANDO

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

APROVADA: 05 de outubro de 2018.

Débora Martins Paixão

Antônio Policarpo Souza Carneiro

Daniele B. D. Marques

Paulo Roberto Cecon

Moyses Nascimento
(Orientador)

À minha família.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado força e determinação para mais essa conquista, além de ter sempre posto coisas boas no meu caminho.

Aos meus pais, Manuel Benício e Luíza Helena por estarem sempre presentes e atuantes na minha vida, prestando sempre o suporte necessário em todos os aspectos, pelo incentivo que recebo todo dia, por me propiciarem uma boa educação e formação e por serem, acima de tudo, meus exemplos.

A minha irmã Mariana, pela companhia, conselhos, amizade, apoio e por se manter sempre próxima a mim.

A toda a minha família, incluindo tios, primos e avós, pelo apoio e suporte, além de serem meus exemplos.

Aos meus grandes amigos que considero como irmãos José Luíz, Lúcio, Sérgio, Gabriel, Ithalo e Eduardo pela amizade e parceria.

Ao meu orientador Moysés Nascimento pelos ensinamentos, pela paciência, preocupação e por ter sempre me incentivado desde os primeiros dias de curso, além da amizade e consideração que me fez tê-lo como um grande amigo. Agradeço por ser uma pessoa com quem posso contar, além de um exemplo como pessoa e como profissional.

Aos meus coorientadores Ana Carolina Campana Nascimento, Fabyano Fonseca e Silva, Antônio Policarpo, Marcos Vinícius e também à professora Camila Azevedo, por contribuírem diretamente no meu aprendizado e pelas sugestões nos trabalhos até aqui realizados.

Aos professores e funcionários do departamento de estatística da Universidade Federal de Viçosa, que sempre se empenharam e se mostraram acessíveis, dispostos a compartilharem conhecimento e suporte para com os alunos.

Aos membros da banca, Paulo Cecon, Débora Paixão, Antônio Policarpo e Daniele Diniz por aceitarem o convite e por estarem dispostos a dar suas contribuições para este trabalho.

Aos meus mestres, amigos e colegas de trabalho da Universidade Federal do Piauí, pelo companheirismo, pelo empenho que sempre tiveram para tornar cada vez melhor o curso de graduação, por me ensinarem os primeiros conceitos da estatística e por serem excelentes colegas de trabalho, se mostrando sempre disponíveis a ajudar em qualquer dificuldade.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

À CAPES, pela concessão da bolsa de estudos.

Por fim, a todos os que contribuíram direta ou indiretamente para a concretização deste trabalho.

BIOGRAFIA

FILIPE RIBEIRO FORMIGA TEIXEIRA, filho de Luíza Helena Ribeiro Formiga Teixeira e Manuel Benício Teixeira Neto, nasceu em Teresina, Piauí, em 20 de Novembro de 1989.

Em março de 2009, ingressou no curso de Bacharelado em Estatística na Universidade Federal do Piauí, Teresina – PI, graduando-se em maio de 2013.

Em agosto do mesmo ano, iniciou o curso de mestrado do Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da dissertação em 26 de fevereiro de 2015.

Em março do mesmo ano ingressou no curso de Doutorado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa da tese de doutorado em 5 de outubro de 2018.

SUMÁRIO

RESUMO

TEIXEIRA, Filipe Ribeiro Formiga, D.Sc., Universidade Federal de Viçosa, outubro de 2018. **Genômica e modelos não-lineares mistos no ajuste de curvas de lactação de bovinos da raça Girolando**. Orientador: Moysés Nascimento. Coorientador: Marcos Vinícius Gualberto Barbosa da Silva.

Pesquisas que visam a construção de curvas de lactação de diferentes raças de gado de leite têm sido realizadas com frequência nos últimos anos. Esses trabalhos usualmente têm como objetivo identificar o comportamento da produtividade leiteira do rebanho, a identificação de indivíduos superiores segundo suas características de lactação ou estudar as associações fenotípicas/genéticas entre diferentes variáveis de lactação. O presente trabalho teve como objetivo propor a utilização da Seleção Genômica Ampla para estimar os valores genéticos genômicos das características de lactação e construir as curvas de lactação genômicas de bovinos da raça Girolando (responsável por 80% da produção do leite brasileiro) baseadas em informações estimadas pelo ajuste de modelos não-lineares mistos. Objetivou-se também a identificar o melhor modelo para o ajuste de curvas de lactação dessa raça, sendo escolhido dentre nove propostos na literatura. Os dados foram fornecidos pela Embrapa Gado de Leite (Juíz de Fora-MG), referentes a 1.822 registros de controle leiteiro correspondente a 226 bovinos Girolando, juntamente com a informação de 37.673 marcadores SNPs associados aos animais em estudo. Comparando nove modelos não-lineares (Brody, Cappio-Borlino, Cobby & Le Du, Dhanoa, Nelder, Papajscik e Bodero, Rook, Sikka e Wood) com a abordagem de modelos mistos, constatou-se que a melhor equação segundo os critérios de AIC e BIC, com valores de 10.013,79 e 10.101,92, respectivamente, foi a de Wood. Através das características de lactação estimadas pelo modelo de Wood foi possível identificar um grupo seleto de 8 animais com maior produtividade (média de 10.584 Kg/lactação). A estimação dos valores genéticos genômicos (*Estimated Genomic Breeding Values – EGBV*) das características genômicas (produção inicial – *a*, taxa de ascensão – *b*, taxa de declínio – *c*, produção total, pico de lactação, persistência e tempo até o pico) através do BLASSO (*Bayesian LASSO*) permitiram o conhecimento genético dessas características. As herdabilidades das mesmas variaram de 0,09 para a taxa de declínio até 0,29 para a persistência. As correlações entre seus valores genéticos genômicos apresentaram resultados de -0,90 (entre a taxa de declínio e a persistência) a 0,98 (entre o pico de lactação e a produção total). Os coeficientes de Kappa para concordância entre os

indivíduos selecionados de acordo com diferentes variáveis variou de 0 a 0,95. As análises realizadas permitiram o conhecimento do melhor entre os modelos não-lineares para ajuste das curvas de lactação de bovinos da raça Girolando. Com a construção das curvas genômicas foi possível identificar diferenças genéticas entre os indivíduos, estas livres dos efeitos ambientais. A abordagem proposta foi capaz de produzir resultados relevantes e pode ser aplicada para outras raças e situações diferentes.

ABSTRACT

TEIXEIRA, Filipe Ribeiro Formiga, D.Sc., Universidade Federal de Viçosa, October, 2018. **Genomic and nonlinear mixed models in the adjustment of lactation curves of Girolando cattle**. Adviser: Moysés Nascimento. Co-adviser: Marcos Vinícius Gualberto Barbosa da Silva.

Researches aiming at the construction of lactation curves of different dairy cattle breeds have been carried out frequently in recent years. These studies usually aim to identify the dairy productivity behavior of the herd, the identification of superior individuals according to their lactation characteristics or to study the phenotypic / genetic associations between different lactation variables. The objective of the present work was to propose the use of the Genome Wide Selection to estimate genomic genetic values of lactation characteristics and to construct genomic lactation curves of Girolando cattle (responsible for 80% of Brazilian milk production) based on estimated information by the adjustment of mixed nonlinear models. The objective was also to identify the best model for the adjustment of lactation curves of this breed, being chosen among nine proposed in the literature. The data were provided by Embrapa Gado de Leite (Juíz de Fora-MG), referring to 1,822 records of dairy control corresponding to 226 Girolando cattle, together with the information of 37,673 markers SNPs associated to the animals under study. Comparing nine non-linear models (Brody, Cappio-Borlino, Cobby & Le Du, Dhanoa, Nelder, Papajscik and Bodero, Rook, Sikka and Wood) with the mixed models approach, it was verified that the best equation according to the AIC and BIC, with values of 10,013.79 and 10.101.92, respectively, was that of Wood. Through the lactation characteristics estimated by the Wood model it was possible to identify a select group of 8 animals with higher productivity (mean of 10,584 kg / lactation). The estimated genomic breeding values (EGBVs) of the genomic characteristics (initial production - a, ascent rate - b, rate of decline - c, total yield, peak lactation, persistence and time to peak) of BLASSO (Bayesian LASSO) allowed the genetic knowledge of these characteristics. Heritabilities ranged from 0.09 for the rate of decline to 0.29 for persistence. Correlations between their genomic genetic values showed results of -0.90 (between the rate of decline and persistence) to 0.98 (between peak lactation and total production). The Kappa coefficients for agreement between the individuals selected according to different variables ranged from 0 to 0.95. The analyzes allowed the knowledge of the best among the non-linear models to adjust the lactation curves of Girolando cattle. With the

construction of genomic curves, it was possible to identify genetic differences between individuals, which are free of environmental effects. The proposed approach was able to produce relevant results and can be applied to other races and different situations.

INTRODUÇÃO GERAL

A pecuária leiteira exerce um papel fundamental no contexto econômico e nutricional. Segundo a Produção Pecuária Municipal – PPM, realizada pelo IBGE (2016), o Brasil é um dos maiores produtores de leite do mundo, ficando em quinto colocado ao produzir cerca de 34 bilhões de litros de leite no ano de 2016. Isso corresponde a cerca de 6,8% da produção mundial, sendo superado apenas pela União Europeia (30,0%), Estados Unidos (19,0%), Índia (13,4%), e China (7,0%). No mesmo ano, a produtividade média por animal foi de 1.709 litros/vaca/ano, correspondendo ao terceiro maior efetivo de vacas leiteiras.

O crescimento recente da produção (72,3% na produção leiteira e de 28,7% do rebanho entre os anos de 2000 e 2015) e aumento da produtividade por animal são resultados da demanda de consumo e de pesquisas associadas à caracterização e melhoramento genético das raças produtoras de leite. Para este propósito, diversos estudos têm sido realizados e grande parte deles faz o uso de técnicas para construção de curvas de lactação para diferentes raças e espécies. Como exemplo, podemos citar Tekerli *et al*. (2000), Farhangfar e Rolinson (2012) e Boujenane e Hilal (2012), os quais buscaram a construção de curvas de lactação para identificar características leiteiras e identificar as correlações genéticas entre essas características.

A abordagem tradicional da construção de curvas de lactação por modelos não-lineares consiste na estimação dos parâmetros por indivíduo, não permitindo a inclusão de efeitos adicionais e é realizado um ajuste por animal. Quando temos diferenças de estações, idade ou grupos contemporâneos, através dessa metodologia é necessário que seja realizado um ajuste de acordo com cada situação de maneira separada. Ainda, segundo Macciotta *et al*. (2011), o principal interesse nesse tipo de modelagem não se trata da construção do comportamento do fenômeno, mas no ajuste de desvios individuais de uma curva média.

Uma maneira eficiente para atingir este objetivo é a utilização de modelos não lineares mistos (MNLM). Essa metodologia permite a inclusão de efeitos fixos (condições ambientais ou físicas citadas anteriormente) e aleatórios (efeitos individuais por animal, em torno da sua média) em um único ajuste, permitindo a modelagem conjunta dos resíduos e eliminando a necessidade da realização de vários ajustes de acordo com diferentes condições. Além disso, estes modelos permitem a inserção de uma estrutura de correlação residual específica, melhorando a qualidade do ajuste, e como a modelagem é

feita com base no conjunto total dos indivíduos, em cada ajuste se utiliza mais informações e raramente existe a necessidade de exclusão de indivíduos da análise.

A extração de informações individuais por meio de efeitos aleatórios permite a obtenção de informações importantes a respeito da produção de cada animal, os quais podem ser analisados geneticamente visando a identificação e seleção dos indivíduos superiores. Com este objetivo, Macciotta *et al*. (2015) buscaram a associação entre marcadores SNPs e características de curvas de lactação em bovinos da raça *Italian Simmental*; Já Cardona *et al*. (2015) objetivaram a identificação de marcadores SNPs relacionados com curvas de produção e qualidade do leite. Mesmo com a grande gama de aplicações dos modelos não-lineares mistos encontrada na literatura, ainda é desconhecida a construção de curvas de lactação genômicas baseadas em marcadores SNPs analisando parâmetros obtidos por meio de modelos não-lineares mistos. Isso é possível com a utilização da Seleção Genômica Ampla – SGA (Meuwissen *et al*., 2001), pois essa metodologia permite e estimação de efeitos de marcadores por meio de um ajuste em relação às características de lactação, representadas pelos parâmetros da curva estimada e por medidas obtidas por meio destes.

A aplicação da SGA tendo como variáveis resposta as características de lactação obtidas por meio de modelos não-lineares mistos permitem o conhecimento genômico da lactação, este sem a interferência de fatores externos que influenciam a lactação, como por exemplo, grupos contemporâneos, idade, etc. Este procedimento busca a construção de modelos estatísticos baseado em informações estimadas diretamente das informações do genoma destes animais.

Diante do exposto, o presente trabalho teve como objetivo propor a utilização da Seleção Genômica Ampla para a construção de curvas de lactação genômicas para identificar o comportamento genético da lactação de bovinos da raça Girolando com base em variáveis estimadas via MNLM, bem como a relação entre essas características. Objetiva-se também a identificação da equação que melhor ajusta os dados para essa raça, dentre nove modelos propostos na literatura.

Na revisão de literatura foram abordados os temas que foram posteriormente utilizados nos capítulos 1 e 2. Inicialmente é apresentado a produção leiteira, juntamente com conceitos importantes sobre modelos não-lineares (para efeitos fixos e mistos). Em seguida são abordadas a análise de agrupamento e algumas considerações importantes sobre Seleção Genômica Ampla, realizada através do BLASSO (*Bayesian LASSO*), juntamente com considerações sobre as medidas usadas posteriormente.

No primeiro capítulo são comparados nove modelos não-lineares mistos para verificar qual das equações melhor ajustou os dados de acordo com critérios comumente utilizados na literatura. Como efeito fixo, foram utilizados níveis de produção obtidos por meio de análise de agrupamento, e como efeito aleatório foram considerados efeitos individuais de cada animal.

No segundo capítulo a Seleção Genômica Ampla, através do Lasso Bayesiano (BLASSO), foi utilizada para a obtenção dos valores genéticos genômicos das variáveis que caracterizam a lactação (produção inicial – *a*, taxa de ascensão – *b*, taxa de declínio – *c*, produção total, pico de lactação, persistência e tempo até o pico). Através destes valores genômicos, foram obtidas as curvas genômicas por animal, as correlações entre os valores genômicos estimados e foi verificada a concordância entre os melhores indivíduos segundo cada variável.

Após os capítulos, finalmente foram apresentadas as considerações finais do trabalho.

REVISÃO DE LITERATURA

1. Produção leiteira

1.1. Introdução

A pecuária leiteira no Brasil teve início em meados do século XVI, sendo inserida no litoral paulista, onde foram trazidos animais da Europa para a então colônia portuguesa. Segundo Vilela *et al*. (2017), a pecuária permaneceu insignificante por mais de três séculos, mas a partir de 1870, quando o cenário político brasileiro favoreceu a tendência agrária e permitiu a modernização das fazendas, surgiu o momento propício para desenvolver a pecuária. Ainda segundo Vilela *et al*. (2017), entre os anos de 1975 e 1985 houve um salto expressivo na produção de leite, de 7,9 milhões de toneladas para 12 milhões (crescimento de mais de 50% em relação à data inicial).

Considerando uma situação mais recente, segundo a Produção Pecuária Municipal – PPM, realizada pelo IBGE (2016), no ano de 2016 o Brasil foi o quinto maior produtor mundial de leite, produzindo pouco mais de 34 bilhões de litros, correspondentes a aproximadamente quase 6,8% da produção mundial. Tal produtividade é superada apenas pela União Europeia (30,0%), Estados Unidos (19,0%), Índia (13,4%), e China (7,0%).

Ainda segundo o IBGE, entre os anos de 2000 e 2015 houve um aumento de 72,3% na produção, juntamente com um crescimento de 28,7% do rebanho. Esse aumento ocorreu principalmente por conta da alta demanda nacional, pois o consumo nacional também apresenta destaque, sendo de aproximadamente 180 litros por habitante no ano de 2015. Outra razão é a importância econômica da produção de leite e seus derivados no cenário nacional e internacional. No Brasil, o agronegócio do leite e seus derivados desempenham um papel relevante no suprimento de alimentos e na geração de emprego e renda para a população.

O Brasil também se destaca em relação à efetividade da produção leiteira por animal, produzindo 1.709 litros/vaca/ano em 2016, terceiro índice considerando os maiores produtores de leite. Isso ocorre devido a maximização da produtividade por animal, resultando no aumento da produção leiteira nacional. Com essa finalidade, o melhoramento genético de gado de leite tem se tornado cada vez mais importante, pois por meio dessa ferramenta é possível modificar a frequência de alelos favoráveis a alguma

característica de interesse, por meio da seleção dos melhores indivíduos. Isso tem como resultado uma nova geração de bovinos de leite mais produtiva que a anterior.

1.2. Melhoramento genético de gado de leite

Devido ao benefício e suas grandes vantagens, empresas especializadas, por meio de programas de melhoramento genético de diversas raças de gado leiteiro, têm surgido e se desenvolvido no âmbito nacional e mundial, viabilizando soluções de pesquisa, desenvolvendo inovação para a sustentabilidade da agricultura em benefício da economia e da sociedade. Dentre essas, podemos citar a Embrapa Gado de Leite, atualmente localizada em Juíz de Fora – MG, por meio do Programa de Melhoramento Genético da raça Girolando (PMGG), o qual tem como objetivos: a identificação de indivíduos superiores, a multiplicação genética de forma orientada, a avaliação de várias características econômicas e a promoção da sustentabilidade da produtividade leiteira.

Além de iniciativas como a da Embrapa Gado de Leite, diversas pesquisas são realizadas com o intuito de caracterizar curvas de lactação de determinadas raças produtoras de leite, bem como o melhoramento da quantidade e qualidade do leite produzido. Podemos citar, por exemplo, estudos como Ledic *et al.* (2002), que teve por objetivo verificar a viabilidade de utilização da produção de leite no dia de controle baseado em informações obtidas de 2082 vacas da raça Gir; ou Esteves *et al.* (2004), os quais estimaram correlações genéticas e fenotípicas entre 21 características de tipo e produção de leite em bovinos da raça holandesa.

Em ambos os estudos destacados, podemos ressaltar duas raças importantes para o melhoramento genético de gado de leite: o Gir e o Holandês. A primeira é uma raça zebuína, que teve origem na Índia, sendo inserida no Brasil por volta de 1906, e por ser uma raça indiana, se adapta bem as condições climáticas do Brasil, além de se destacar pela produção de leite. Já a segunda se refere a uma raça pura europeia, que é bastante difundida em todo o mundo também por conta da sua alta produtividade (superior aos gados Gir) e do alto retorno financeiro decorrente dessa produtividade. Porém, gados da raça Holandesa são adaptados ao clima europeu, que é diferente do clima tropical brasileiro.

Com o objetivo de unir as características importantes de diferentes raças produtoras de leite, algumas raças chamadas de mestiças vem sendo criadas. Os

cruzamentos entre raças de origens diferentes, são uma maneira recomendada para aumentar a tolerância às condições ambientais adversas.

Trazendo para o atual contexto, podemos citar como exemplo o cruzamento entre gados da raça Gir e Holandês, resultando numa raça conhecida como Girolando, objetivando unir a adaptabilidade climática de regiões tropicais encontrada nos gados Gir com a produtividade do gado Holandês. Esta raça é fundamentalmente produzida pelo cruzamento do Holandês com o Gir, passando por variados graus de sangue. O direcionamento dos acasalamentos busca a fixação do padrão racial, no grau de 5/8 Holandês + 3/8 Gir, objetivando um gado produtivo e padronizado, buscando a consolidação do Puro Sintético da Raça Girolando (PS), a raça propriamente dita.

Para que o melhoramento genético de raças de gado produtor de leite possa ser quantificado e avaliado, utiliza-se um procedimento chamado de controle leiteiro. Neste processo, são medidas informações importantes sobre a produção de leite em diferentes dias de lactação, que são posteriormente utilizadas para fazer análises, previsões e principalmente para seleção dos melhores indivíduos, ou seja, aqueles que mais produzem.

1.3. Controle leiteiro

O controle leiteiro é uma ferramenta utilizada para se verificar a evolução produtiva de cada animal integrante do rebanho. Trata-se da coleta e análise de informações a respeito da produtividade de cada animal com o objetivo de caracterizar a produção leiteira do rebanho. Segundo Teodoro e Verneque (2000), dentre as finalidades do controle leiteiro, destacam-se: i) o fornecimento de alimentos, principalmente o concentrado, de acordo com a produção de leite. Conhecendo-se, portanto, o potencial de produção de uma determinada vaca, não estaremos fornecendo alimentos além do necessário para algumas e aquém para outras; ii) o provimento de informações que auxiliem no melhoramento genético animal. Conhecendo-se a produção dos animais e seu valor genético, pode-se então selecionar os melhores e usá-los intensivamente nos acasalamentos; iii) uso das informações do controle leiteiro para propaganda do rebanho, e esta utilização comercial certamente induzirá a uma maior disseminação dos genótipos superiores, principalmente por meio da venda de tourinhos ou de sêmen de touros provados.

A análise de dados obtidos pela coleta de informações do controle leiteiro é usualmente realizada através do ajuste de modelos não-lineares das curvas de lactação de gados de leite, aplicados, por exemplo, em Ghavi Hossein-Zadeh (2015), Tekerli *et al*. (2000) e Piccardi *et al*. (2017). O conhecimento do comportamento das curvas dos animais que compõe qualquer rebanho é de grande importância para a caracterização da sua produtividade.

1.4. Curvas de lactação

De acordo com Cobuci *et al*. (2001), a curva de lactação é uma representação gráfica da produção de leite de uma vaca no decorrer de sua lactação. O conhecimento do comportamento das curvas de lactação de um rebanho auxilia no descarte e na seleção de animais de acordo com um padrão desejável, preestabelecido conforme a capacidade de produção. Dessa maneira, a comparação da forma da curva entre grupos distintos de animais, com diferentes composições raciais, idades ao parto, rebanhos e outros tratamentos de interesse é de grande importância, pois mediante essas comparações, podem ser obtidas informações sobre a eficiência desses grupos propiciando um melhor controle da produção (Groenewald & Viljoen, 2003).

Nesse contexto e com o objetivo de caracterizar a produção leiteira dos correspondentes rebanhos, vários estudos sugerem a utilização de curvas de lactação. Podemos citar como exemplo o estudo desenvolvido por Molento *et al.* (2004), que teve como objetivo a construção de curvas de lactação de um rebanho de vacas Holandesas no estado do Paraná por meio de usuais modelos não-lineares. Já Oliveira *et al*. (2007) utilizaram a função gama incompleta para ter acesso a curvas de lactação de vacas Holandês-Gir baseadas em um rebanho de 5.368 vacas, onde foram estimados parâmetros importantes como pico e persistência de lactação.

A curva de lactação é composta por basicamente três fases: i) fase crescente, em que a produção de leite aumenta até atingir a fase seguinte; ii) fase de pico, representada pela produção máxima observada, seguida de; iii) fase de declínio contínuo até o final da lactação, ou previamente pela secagem completa aos 305 dias de lactação. O processo de produção leiteira pode ser descrito por importantes características que configuram a lactação de cada animal em estudo. Algumas variáveis importantes são usualmente mensuradas para que se tenha um conhecimento mais amplo sobre a lactação dos animais

do rebanho. Tais medidas são calculadas para cada indivíduo, geralmente por meio do uso de procedimentos estatísticos utilizando modelos não-lineares.

1.5. Medidas importantes obtidas pelas curvas de lactação

Para o entendimento e a caracterização da produção leiteira de um rebanho, é imprescindível o conhecimento dos parâmetros estimados pelo modelo, bem como de quatro funções paramétricas: a produção total, o pico de lactação, a persistência e o tempo até o pico, que também são variáveis associadas a lactação do rebanho. Tais atributos são responsáveis pelo formato das curvas de lactação, auxiliando na identificação dos animais menos ou mais produtivos, facilitando a seleção e o direcionamento dos cruzamentos entre os melhores indivíduos.

A produção total de um gado leiteiro é a quantidade (em kg) de leite que o mesmo produz no decorrer de cada lactação, geralmente calculado considerando 305 dias após o início da lactação. O pico de lactação é a produção máxima observada de cada animal, ou seja, é o ponto máximo que a curva de lactação irá atingir. Esse tempo varia de acordo com a espécie e/ou raça em questão. Gonçalves *et al.* (2002) encontraram um valor de pico de lactação variando aproximadamente de 24,8 a 25,1 dias após o parto de acordo com diferentes equações. Já Cobuci *et al.* (2004), também trabalhando com lactações de animais da raça Holandesa, relataram que o pico de lactação ocorreu entre 60 e 90 dias de lactação, constatando uma variação mesmo considerando dados de uma mesma raça. Considerando-se ainda vacas zebuínas ou mestiças, tal pico pode-se apresentar no primeiro dia da lactação, ou seja, iniciando na produção máxima com ausência da fase de inclinação do parto ao pico (Papajcsik & Bodero, 1988). Tal fato também foi reportado na raça Gir por Rebouças *et al*. (2008). O tempo até o pico, que pode ser mensurado em dias, semanas ou meses, é o intervalo de tempo que cada animal leva para atingir o pico de lactação.

Já o conceito de persistência é descrito de maneiras diferentes por alguns autores. Wood (1967) afirma que é a extensão pela qual a produção máxima na lactação é mantida. Gengler (1996) considera a habilidade do animal em manter mais ou menos constante a produção de leite durante a lactação; para Tekerli (2000), é a expressão da capacidade da vaca em continuar a produzir leite nos níveis de produção do pico em toda a lactação. As definições formais apresentam pequenas diferenças, mas a ideia é a mesma. A persistência na lactação mede, de maneira geral, o decaimento da produção de leite de

cada vaca após o pico de lactação. Esse parâmetro é responsável por mensurar a última fase da lactação: o declínio até o final da produção de leite. Para Wood (1967), a persistência é o principal componente da curva de lactação.

As variáveis descritas acima são de grande importância em estudos de curvas de lactação. Na literatura usualmente o objetivo se trata da construção de curvas de lactação e de estudos de associação (genéticas e/ou fenotípicas) entre essas características. Como exemplos envolvendo este objetivo, podemos citar os estudos de Boujenane e Hilal (2012), El-Awady (2013), Farhangfar e Rowlinson (2007), Muir *et al*. (2004), Saghanezhad *et al*. (2017), Rekaya *et al*. (2000) e Canaza-Cayo *et al*. (2015).

A estimação de importantes parâmetros, como os que foram citados anteriormente, é comumente feita através de parâmetros estimados por modelos estatísticos não-lineares, que fazem uso do estimador de máxima verossimilhança e de processos iterativos para obtenção de estimativas precisas dos parâmetros. Na seguinte seção alguns destes procedimentos são descritos para o entendimento dos modelos não-lineares.

2. Modelos Não-Lineares e suas abordagens

2.1. Introdução

Nas diversas áreas de pesquisa, quando se deseja verificar a associação entre uma variável resposta (também conhecida como variável dependente, geralmente denotada por Y) e uma ou mais variáveis explicativas (condicionantes ou independentes, representadas por X), é comum o uso de técnicas de análise de regressão. Tais metodologias consistem em encontrar, por meio de uma equação, uma relação funcional entre a variável resposta e a (s) variável (is) independente (s).

A construção de modelos de regressão pode ser útil quando temos como objetivo quantificar a influência das variáveis explicativas na variável de interesse Y, selecionar as condicionantes que exercem maior influência sobre a resposta ou predizer o comportamento de Y diante de uma observação de X, sendo ou não um valor observado dentro do intervalo considerado.

Segundo Mazucheli e Archar (2002), pode-se atingir este objetivo por meio dos bem conhecidos modelos de regressão, os quais se dividem em duas classes distintas: os lineares e os não-lineares. O modelo estatístico é linear se a quantidade de interesse,

geralmente a média de Y, é função linear dos parâmetros, caso contrário é não linear, o que caracteriza a principal diferença entre esses dois tipos de modelos. Tal divergência também está presente nas suas respectivas formulações.

No caso dos modelos lineares, podemos ter inúmeras variáveis diferentes para explicar a variável resposta, e caso exista alguma relação entre elas, a associação será quantificada por meio de uma equação. Para essa situação, inicialmente desconhecemos tanto a equação quanto os parâmetros nela contidos. Em contrapartida, na situação em que a relação entre as variáveis explicativa e resposta é não-linear, outras metodologias devem ser utilizadas, fazendo-se necessária a aplicação da abordagem dos modelos não-lineares. Neste caso, em geral a relação entre as variáveis é descrita através de uma curva.

A escolha pela utilização de modelos não-lineares para estimação de parâmetros normalmente vem de um conhecimento prévio que se tem acerca do fenômeno em estudo. Brody (1945), por exemplo, reitera que as funções não-lineares se ajustam melhor às diferentes informações relacionadas a peso-idade durante o crescimento, permitindo o agrupamento dessas informações em poucos parâmetros biologicamente interpretáveis (como taxas de crescimento, curvas de lactação, etc.), facilitando o entendimento do fenômeno.

Por essa razão, essa classe de modelos é usual para descrever processos como crescimento, decaimento, competição, mortalidade, lactação, dentre outros processos que, em situações usuais, esporadicamente possuem parâmetros relacionados linearmente com a variável de interesse. Este fato caracteriza uma das principais vantagens sobre o uso de regressão não-linear. Dentre outras, podemos citar a pequena quantidade de parâmetros a serem estimados, sendo modelos parcimoniosos; a interpretação prática (física, química, biológica, etc.) de cada um desses parâmetros, facilitando a descrição da relação entre as variáveis, entre outras vantagens.

Diferentemente dos modelos lineares, para estimação dos parâmetros são requeridos métodos computacionais iterativos, visto que a resolução dos sistemas de equações de estimação dos parâmetros se torna bastante complicada no caso de uma relação não-linear entre as variáveis. Procedimentos computacionais buscam, a cada iteração, aproximar valores iniciais sugeridos das reais estimativas dos parâmetros de interesse. As aproximações são avaliadas por meio de critérios de convergência.

2.2. Formulação dos modelos não-lineares

A relação entre a variável resposta Y e a (s) variável (is) explicativa (s) é descrita por meio de uma equação, denotada, de maneira geral, por Mazucheli e Archar (2002):

$$\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

em que $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ é o vetor numérico que representa as observações $1, 2, \dots, n$ da variável resposta Y; $f(\mathbf{x}; \boldsymbol{\theta}) = [f(x_1; \boldsymbol{\theta}), f(x_2; \boldsymbol{\theta}), \dots, f(x_n; \boldsymbol{\theta})]$ é a função de regressão (ou função esperança), formada por uma combinação linear entre o vetor dos coeficientes $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)'$ e pelas observações (constantes) da (s) variável (s) explicativa (s) X ('s); e finalmente $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ é o vetor dos erros aleatórios, que representa a proporção de variação em Y devido ao acaso. Assume-se usualmente que os erros são variáveis aleatórias independentes e identicamente distribuídas. O erro relativo da i-ésima observação de Y (ε_i) possui distribuição normal com média 0 e variância σ^2, ou seja, $\varepsilon_i \sim N(0, \sigma^2)$. Na notação matricial, temos que a distribuição do vetor de erros será $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Sendo a variância constante e $\mathbf{I}_n$ uma matriz identidade de ordem n.

Para o melhor entendimento da definição da não-linearidade de um modelo, consideremos três exemplos de funções esperança, dadas abaixo:

$$\text{(i) } f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

$$\text{(ii) } f(x) = \theta_0 + \theta_1 e^{-\theta_2 x}$$

$$\text{(iii) } f(x) = \theta_0 [1 - e^{-\theta_1 (x - \theta_2)}]$$

O modelo (i) é linear nos parâmetros θ_0, θ_1 e θ_2. Trata-se de um modelo polinomial de segunda ordem (ou segundo grau). No caso de (ii) temos um modelo linear em θ_0 e θ_1, porém não linear em θ_2, tratando-se, portanto de um modelo não linear. E o exemplo do modelo de Mitscherlich (iii) é não linear nos três parâmetros a serem estimados (θ_0, θ_1 e θ_2).

Para determinar a não-linearidade de um modelo, usamos como referência a derivada parcial da função esperança em relação a cada parâmetro do vetor $\boldsymbol{\theta}$. O modelo é classificado como não-linear se pelo menos uma das derivadas parciais da função esperança em relação ao parâmetro é função de parâmetros desconhecidos (PRUDENTE, 2009). Como desconhecemos o verdadeiro valor de $\boldsymbol{\theta}$ e seu estimador não depende exclusivamente da amostra, necessita-se de métodos computacionais para obter valores aproximados como estimativas dos parâmetros.

2.3. Método dos mínimos quadrados

Dadas as especificações acima, o método dos mínimos quadrados (MMQ) consiste em obter o vetor de estimadores $\widehat{\boldsymbol{\theta}}$ que minimize a soma de quadrados dos erros (**SQE**), ou seja, que minimize a quantidade abaixo:

$$\mathbf{SQE} = \sum_{i=1}^{n}[y_i - f(x_i; \boldsymbol{\theta})]^2.$$

Na prática, a equação acima representa a distância entre o valor estimado por meio de $f(x_i; \boldsymbol{\theta})$ e o valor real da i-ésima observação da variável Y, denotado por y_i. A mesma medida pode ser representada em notação vetorial:

$$\mathbf{SQE} = [\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})]'[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})] = ||\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})||^2$$

Como usualmente trabalha-se com a forma matricial, temos funções com diferentes valores de cada parâmetro e de cada observação x_i para acessar os valores estimados do vetor $\mathbf{y}$. Assim, tanto $\mathbf{y}$ quanto $f(\mathbf{x}; \boldsymbol{\theta})$ possuem dimensão nx1, onde $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ e $f(\mathbf{x};\ \boldsymbol{\theta}) = [f(x_1; \boldsymbol{\theta}),\ f(x_2; \boldsymbol{\theta}),\ \dots, f(x_n; \boldsymbol{\theta})\ ]'$; e o vetor de parâmetros (denotado por $\boldsymbol{\theta}$) tem dimensão px1, sendo p a quantidade de parâmetros considerado no modelo.

As resoluções para cada parâmetro com referência a cada observação podem ser atingida por meio das derivadas parciais de primeira ordem em relação ao vetor de parâmetros $\boldsymbol{\theta}$. Para descrição dessas derivadas, temos a matriz $\mathbf{F}(\boldsymbol{\theta})$, de dimensão nxp conhecida também como matriz Jacobiana do vetor $f(\mathbf{x}; \boldsymbol{\theta})$, descrita abaixo:

$$\mathbf{F}(\boldsymbol{\theta}) = \frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial f(x_1; \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(x_1; \boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(x_1; \boldsymbol{\theta})}{\partial \theta_p} \\ \frac{\partial f(x_2; \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(x_2; \boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(x_2; \boldsymbol{\theta})}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x_n; \boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(x_n; \boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(x_n; \boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}.$$

Para encontrar o vetor $\widehat{\boldsymbol{\theta}} = (\theta_1,\ \theta_2, \dots, \theta_p)'$ que minimiza a soma de quadrados dos erros, devemos derivar **SQE** e igualar a derivada a zero, como segue:

$$\left[\frac{\partial \mathbf{SQE}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}' = \widehat{\boldsymbol{\theta}}'} = \frac{\partial}{\partial \boldsymbol{\theta}^{\mathbf{t}}}[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})]'[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})] = \mathbf{0}$$

$$\left[\frac{\partial \mathbf{SQE}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathbf{t}}}\right]_{\boldsymbol{\theta}' = \widehat{\boldsymbol{\theta}}'} = -2[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})]' \frac{\partial \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathbf{t}}} = -2[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})]'\mathbf{F}(\boldsymbol{\theta})$$

$$\left[\frac{\partial \mathbf{SQE}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathbf{t}}}\right]_{\boldsymbol{\theta}' = \widehat{\boldsymbol{\theta}}'} = -2[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})]'\mathbf{F}(\boldsymbol{\theta}) = -2\mathbf{F}(\boldsymbol{\theta})'[\mathbf{y} - f(\mathbf{x}; \boldsymbol{\theta})] = \mathbf{0}$$

Logo,

$$\mathbf{F}(\widehat{\boldsymbol{\theta}})'[\mathbf{y} - \mathrm{f}(\mathbf{x}; \widehat{\boldsymbol{\theta}})] = \mathbf{0}$$

Substituindo os valores individuais dentro nas matrizes e vetores, teremos:

$$\begin{bmatrix} \frac{\partial \mathrm{f}(\mathrm{x}_1; \widehat{\boldsymbol{\theta}})}{\partial \theta_1} & \frac{\partial \mathrm{f}(\mathrm{x}_2; \widehat{\boldsymbol{\theta}})}{\partial \theta_1} & \cdots & \frac{\partial \mathrm{f}(\mathrm{x}_n; \widehat{\boldsymbol{\theta}})}{\partial \theta_1} \\ \frac{\partial \mathrm{f}(\mathrm{x}_1; \widehat{\boldsymbol{\theta}})}{\partial \theta_2} & \frac{\partial \mathrm{f}(\mathrm{x}_2; \widehat{\boldsymbol{\theta}})}{\partial \theta_2} & \cdots & \frac{\partial \mathrm{f}(\mathrm{x}_n; \widehat{\boldsymbol{\theta}})}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathrm{f}(\mathrm{x}_1; \widehat{\boldsymbol{\theta}})}{\partial \theta_p} & \frac{\partial \mathrm{f}(\mathrm{x}_2; \widehat{\boldsymbol{\theta}})}{\partial \theta_p} & \cdots & \frac{\partial \mathrm{f}(\mathrm{x}_n; \widehat{\boldsymbol{\theta}})}{\partial \theta_p} \end{bmatrix} \begin{bmatrix} \mathrm{y}_1 - \mathrm{f}(\mathrm{x}_1; \widehat{\boldsymbol{\theta}}) \\ \mathrm{y}_2 - \mathrm{f}(\mathrm{x}_2; \widehat{\boldsymbol{\theta}}) \\ \vdots \\ \mathrm{y}_n - \mathrm{f}(\mathrm{x}_n; \widehat{\boldsymbol{\theta}}) \end{bmatrix} = \mathbf{0}$$

Após o produto, teremos como resultado um vetor formado por p equações, referentes a cada parâmetro θ_j $(j = 1, \ldots, p)$, conhecidas como equações normais ou homogêneas, dadas por:

$$\sum_{j=1}^{p} \sum_{i=1}^{n} [\mathrm{y}_i - \mathrm{f}(\mathrm{x}_i; \widehat{\boldsymbol{\theta}})] \left[\frac{\partial \mathrm{f}(\mathrm{x}_i; \widehat{\boldsymbol{\theta}})}{\partial \theta_j} \right] = 0$$

Para o caso dos modelos não lineares, $\boldsymbol{\theta}$ estará presente em pelo menos uma das derivadas, o que torna um parâmetro dependente do outro. Isso leva a concluir que em modelos multiparamétricos, as soluções das equações normais podem ser extremamente difíceis de serem obtidas e algum método iterativo de resolução de equações normais não-lineares deve ser utilizado na maioria dos casos (Bates e Watts, 1988; Ratkowsky, 1990).

2.4. Método iterativo de Gauss-Newton

O método iterativo de Gauss-Newton é uma alternativa para solucionar a minimização da soma de quadrados quando esta depende de outros parâmetros. Neste procedimento, o objetivo é, a partir de um chute inicial para os parâmetros de interesse, buscar valores cada vez mais próximos de $\boldsymbol{\theta}$ a cada iteração por meio de uma aproximação linear.

Para cada indivíduo i, dispõe-se de pares de valores, geralmente denotados por $(\mathrm{x}_i, \mathrm{y}_i), i = 1, \ldots, n$. A função esperança $\mathrm{f}(\mathbf{x}; \boldsymbol{\theta})$ é especificada geralmente de acordo com o comportamento do fenômeno em estudo. Como visto anteriormente, o principal objetivo do algoritmo consiste na minimização da soma de quadrados dos erros, ou seja, da quantidade abaixo:

$$\mathbf{SQE}(\boldsymbol{\theta}) = \sum_{i=1}^{n} [\mathrm{y}_i - \mathrm{f}(\mathbf{x}_i; \boldsymbol{\theta})]^2.$$

O primeiro passo é estabelecer um chute inicial para o vetor de parâmetros ($\boldsymbol{\theta}^*$), supostamente próximo de $\boldsymbol{\theta}$, segundo algum critério. A ideia geral é aproximar, a cada iteração, o chute inicial do verdadeiro valor de $\boldsymbol{\theta}$**,** consequentemente minimizando cada vez mais a equação acima.

A função esperança, de maneira geral, pode ser descrita por uma aproximação de Taylor em volta do chute inicial $\boldsymbol{\theta}^*$, descrita na forma matricial por (WEISBERG, 2005):

$$f(\mathbf{x_i}; \boldsymbol{\theta}) \approx f(\mathbf{x_i}; \boldsymbol{\theta}^*) + \mathbf{u}_i(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

Em que o vetor $\mathbf{u} = \frac{\partial f(\mathbf{x_i};\boldsymbol{\theta})}{\partial \theta_j}$ caracteriza o vetor de derivadas (seção anterior) e varia de acordo com o modelo considerado.

Substituindo a aproximação por séries de Taylor na soma de quadrados dos erros, teremos:

$$\mathbf{SQE}(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - f(\mathbf{x_i}; \boldsymbol{\theta})]^2 = \sum_{i=1}^{n} [y_i - f(\mathbf{x_i}; \boldsymbol{\theta}^*) - \mathbf{u}_i(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2$$

Como $y_i - f(\mathbf{x_i}; \boldsymbol{\theta}^*) = \varepsilon_i^*$, podemos reescrever a quantidade acima da seguinte forma:

$$\mathbf{SQE}(\boldsymbol{\theta}) = \sum_{i=1}^{n} [\varepsilon_i^* - \mathbf{u}_i(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)]^2$$

Que na forma vetorial será:

$$\mathbf{SQE}(\boldsymbol{\theta}) = [\boldsymbol{\varepsilon}_\mathbf{i}^* - \mathbf{u_i}(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)]'[\boldsymbol{\varepsilon}_\mathbf{i}^* - \mathbf{u_i}(\boldsymbol{\theta}^*)'(\boldsymbol{\theta} - \boldsymbol{\theta}^*)]$$

Como trabalhamos com aproximações, devemos utilizar a derivada em função de $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. Após o produto e a derivada, igualando o resultado a zero chegamos a:

$$\widehat{\boldsymbol{\theta} - \boldsymbol{\theta}}^* = [\mathbf{U}(\boldsymbol{\theta}^*)'\mathbf{U}(\boldsymbol{\theta}^*)]^{-1}\mathbf{U}(\boldsymbol{\theta}^*)'\mathbf{W}\hat{\mathbf{e}}^*$$

Com isso, o valor atualizado do estimador $\widehat{\boldsymbol{\theta}}$ será:

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + [\mathbf{U}(\boldsymbol{\theta}^*)'\mathbf{U}(\boldsymbol{\theta}^*)]^{-1}\mathbf{U}(\boldsymbol{\theta}^*)'\hat{\mathbf{e}}^*$$

A atualização dos valores assumidos no vetor de estimadores é feita com o objetivo de nos aproximar cada vez mais do valor real de $\boldsymbol{\theta}$.

De forma resumida, Weisberg (2005) descreve o algoritmo de Gauss-Newton nos 4 passos abaixo:

i. Selecionar um chute inicial $\boldsymbol{\theta}^{(0)}$ para o vetor de parâmetros $\boldsymbol{\theta}$, computando $\mathrm{SQE}(\boldsymbol{\theta}^{(0)})$;
ii. Definir o contador de iterações em $j = 1$;

iii. Computar $\mathbf{U}(\boldsymbol{\theta}^{(j)})$ e $\hat{\mathbf{e}}^{*(j)}$ com o i-ésimo elemento de $y_i - f(\mathbf{x_i}, \boldsymbol{\theta}^{(j)})$. Assim, chegaremos ao cálculo da equação AAA, determinando o novo estimador $\boldsymbol{\theta}^{(j+1)}$. Em seguida, é calculado $SQE(\boldsymbol{\theta}^{(1)})$;

iv. Parar o algoritmo se $SQE(\boldsymbol{\theta}^{(j)}) - SQE(\boldsymbol{\theta}^{(j+1)})$ é suficiente pequeno. Nesse caso, dizemos que houve convergência. Caso contrário, definir $j = j + 1$ e utilizar $\boldsymbol{\theta}^{(1)}$ como valor inicial, repetindo o procedimento.

Ainda segundo Weisberg (2005), para que a estimação pelo método de Gauss-Newton seja possível são necessários: dos chutes iniciais e da existência das derivadas das funções $f(\mathbf{x_i}; \boldsymbol{\theta})$. Para os chutes iniciais, é de grande importância que sejam corretamente especificados, pois com uma boa especificação, é provável que ocorra a convergência, desejável pelo pesquisador.

Segundo Ritz (2008), adivinhar os valores iniciais é o método mais utilizado para determinação dos chutes. Porém, isso exige certa experiência do pesquisador, que certamente utilizará informações semelhantes a análises anteriores. A grande aplicação da "adivinhação" dos chutes se dá pelo fato de os parâmetros contidos nos modelos não lineares possuírem uma interpretação prática. Isso torna a estimativa dos parâmetros mais previsível, facilitando os chutes iniciais e permitindo que estes sejam estabelecidos de acordo com a experiência do pesquisador. Porém, quando não se tem o conhecimento prévio nem experiência em relação ao fenômeno em estudo, torna-se mais difícil a adivinhação de chutes iniciais. Para estes casos, temos algumas alternativas, como por exemplo: i) a exploração do gráfico de dispersão entre as variáveis em estudo; e ii) por meio de funções específicas que identificam bons valores iniciais (*Self-starter functions*). Quando os chutes iniciais se distanciam do valor real do parâmetro, o processo de convergência poderá falhar (WEISBERG, 2005).

3. Modelos não-lineares mistos

3.1. Introdução

A abordagem tradicional de ajuste de modelos não-lineares consiste na realização de um ajuste para cada indivíduo na análise, onde têm-se várias observações referentes à produtividade leiteira (em quilos por dia, semana ou mês) no decorrer de um determinado período de lactação de cada animal. Quando temos animais sob diferentes

condições que podem afetar a produtividade, como por exemplo, grupos contemporâneos, idade, ordem de parto, etc. devemos realizar diferentes ajustes de acordo com cada condição para que possamos identificar possíveis diferenças entre estas.

Uma abordagem que permite a modelagem dessas diferentes condições juntamente com os efeitos individuais de cada animal é o ajuste de Modelos Não-Lineares Mistos (MNLM). Neste caso, como utilizamos modelos mistos, variáveis físicas ou ambientais podem ser inseridas e mensuradas como efeito fixo e os efeitos individuais dos animais são analisadas como efeitos aleatórios centrados em torno da sua média, permitindo assim a extração das informações necessárias em apenas um ajuste, contando também com a modelagem conjunta dos resíduos (Silva *et al*., 2016).

Segundo Lindstrom e Bates (1990), modelos não lineares mistos têm se tornado populares porque permitem a modelagem de uma estrutura de covariâncias de maneira flexível, o que permite a correlação entre as observações, o que também é útil para o caso de dados desbalanceados, visto que os dados são modelados conjuntamente. Assim, indivíduos que possuem poucas observações, ou em situação em que não temos a mesma quantidade de observações para todos os indivíduos em estudo, não precisam ser excluídos da análise por este motivo, o que constitui uma das grandes vantagens desse método.

As subseções seguintes descrevem o modelo, a estimação dos parâmetros e algumas considerações importantes dos modelos não lineares mistos, além da representação dos parâmetros para o caso de diversos modelos de curvas de lactação.

3.2. Modelo

A representação do modelo não linear misto, de acordo com Pinheiro e Bates (2000) pode ser compreendida como um modelo hierárquico. A j-ésima observação para o i-ésimo animal pode ser denotada por:

$$y_{ij} = f(\boldsymbol{\phi}_i, \mathbf{x}_{ij}) + \mathbf{e}_{ij}, \quad i = 1, \dots, M; j = 1, \dots, n_i$$

em que M é o número de indivíduos, n_i é o número de observações do i-ésimo indivíduo, f é uma função não negativa e diferenciável do vetor de parâmetros $\boldsymbol{\phi}_i$ e do vetor de preditores $\mathbf{x}_{ij}$; e $\mathbf{e}_{ij}$ é o termo que representa os erros, sendo $\mathbf{e}_{ij} \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Lambda})$. A função f é não linear em pelo menos um dos componentes do vetor de parâmetros $\boldsymbol{\phi}_i$, o qual pode ser escrito da seguinte maneira:

$$\boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D} = \boldsymbol{\Psi}),$$

em que $\boldsymbol{\beta}$ é um vetor dimensional de dimensão p associado aos efeitos fixos inseridos no modelo; $\mathbf{b}_i$ é um vetor de dimensão q associado ao efeito aleatório do i-ésimo indivíduo; e $\mathbf{A}_i$ e $\mathbf{B}_i$ são as matrizes de incidência de dimensões r x p e r x q associadas, respectivamente, aos efeitos fixos e aleatórios; $\sigma^2\mathbf{D}$ é a matriz de covariâncias de $\mathbf{b}_i$.

O vetor de observações do i-esimo indivíduo pode ser escrito na forma matricial, como segue:

$$\mathbf{y}_i = \begin{bmatrix} y_{1n1} \\ y_{1n2} \\ \vdots \\ y_{1ni} \end{bmatrix} ; \ \mathbf{e}_i = \begin{bmatrix} e_{1n1} \\ e_{1n2} \\ \vdots \\ e_{1ni} \end{bmatrix} ; \text{ e } \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) = \begin{bmatrix} f(\boldsymbol{\phi}_i, \mathbf{x}_{i1}) \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{i2}) \\ \vdots \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{ini}) \end{bmatrix},$$

podendo ser representado por:

$$\mathbf{y}_i = \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) + \mathbf{e}_i,$$

em que $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i)$. Em muitos casos, $\boldsymbol{\Lambda}_i = \mathbf{I}$, porém pode se especificar estruturas específicas para a matriz de covariâncias residual, como AR(1), simetria composta, dentre outras (Pinheiro e Bates, 2000).

O modelo descrito anteriormente pode ser escrito para o conjunto de dados completo, formado por M indivíduos. Nesse caso, temos:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} ; \ \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_M \end{bmatrix} ; \text{ e } \boldsymbol{\eta}(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{\eta}_1(\boldsymbol{\phi}_1) \\ \boldsymbol{\eta}_2(\boldsymbol{\phi}_2) \\ \vdots \\ \boldsymbol{\eta}_M(\boldsymbol{\phi}_M) \end{bmatrix},$$

E o modelo completo é dado por:

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\phi}) + \mathbf{e}$$

Com as seguintes distribuições:

$$\mathbf{y}|\mathbf{b} \sim N(\boldsymbol{\eta}(\boldsymbol{\phi}), \sigma^2\boldsymbol{\Lambda}), \ \boldsymbol{\phi} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}_i,$$

$$\mathbf{b} \sim N(\mathbf{0}, \sigma^2\widetilde{\mathbf{D}}),$$

em que: $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_M)$, $\widetilde{\mathbf{D}} = \text{diag}(\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D})$, $\mathbf{B} = \text{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M)$,

$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_M \end{bmatrix}$ e $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}$.

O modelo acima permite a inclusão de efeitos aleatórios para qualquer parâmetro em análise. A utilização do modelo completo e a flexibilidade das matrizes de incidência permitem a inclusão de indivíduos com poucas observações na análise, além de possibilitar o ajuste de dados desbalanceados sem perda de eficiência do modelo.

Segundo Pinheiro e Bates (2000), como os efeitos aleatórios são quantidades inobserváveis, sua estimação por máxima verossimilhança nos modelos com efeitos mistos é baseada na densidade marginal de $\mathbf{y}$ considerando Q níveis de agrupamento (indivíduos), dada por:

$$p\left(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi_1}, \ldots, \boldsymbol{\Psi_Q}\right) = \int p(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}, \sigma^2)\, p\left(\mathbf{b}|\boldsymbol{\Psi_1}, \ldots, \boldsymbol{\Psi_Q}\right) d\mathbf{b}$$

Sendo $p\left(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi}, \ldots, \boldsymbol{\Psi_Q}\right)$ a densidade marginal de $\mathbf{y}$, $p(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}, \sigma^2)$ é a densidade condicional de $\mathbf{y}$ dados os efeitos aleatórios $\mathbf{b}$, e $p\left(\mathbf{b}|\boldsymbol{\Psi_1}, \ldots, \boldsymbol{\Psi_Q}\right)$ é a distribuição marginal de $\mathbf{b}$.

Como a densidade marginal de $\mathbf{y}$ pode ser não linear, sua integral não possui uma forma fechada. Diversas metodologias têm sido propostas para estimação dos parâmetros de efeito fixo e aleatório do modelo, bem como os componentes de variância. Como exemplo, podemos citar aproximações da equação acima tomando uma série de Taylor de primeira ordem em torno dos valores esperados dos efeitos aleatórios (Sheiner e Beal, 1980), do fator de precisão para representar a matriz de covariância de $\mathbf{b}$ (Lindstrom e Bates, 1990), ou por regras de quadratura gaussiana (Davidian e Gallant, 1992). No presente trabalho, será apresentado apenas a aproximação proposta por Lindstrom e Bates (1990), juntamente com seu algoritmo de estimação.

3.3. Estimação dos parâmetros (Algoritmo de Lindstrom e Bates)

O algoritmo proposto por Lindstrom e Bates (1990) consiste na alternância entre dois passos: o primeiro consiste na minimização da soma de quadrados não linear, conhecido como *Penalized Nonlinear Least Squares (PNLS) Step*, o qual faz uso do algoritmo de Gauss-Newton. O segundo é semelhante à estimação dos componentes de variância dos modelos lineares mistos, e por isso é chamado de *Linear Mixed Effects (LME) Step*. Neste procedimento, a matriz de covariâncias dos efeitos aleatórios ($\boldsymbol{\Psi}$) é escrita em função de uma medida chamada fator de precisão relativo ($\boldsymbol{\Delta}$), de modo que $\boldsymbol{\Psi}^{-1} = \sigma^2 \boldsymbol{\Delta}' \boldsymbol{\Delta}$. Abaixo segue uma descrição detalhada de cada um dos passos, descritos por Lindstrom e Bates (1990); e Pinheiro e Bates (2000).

1) *PNLS Step*: Neste passo, com base nos chutes iniciais para os parâmetros e consequentemente na atual estimativa de $\boldsymbol{\Delta}$, são obtidos os valores de $\boldsymbol{\beta}$ e $\mathbf{b_i}$ que minimizam a soma de quadrados penalizada (considerando M indivíduos):

$$\sum_{i=1}^{M} [\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)\|^2 + \|\boldsymbol{\Delta}\mathbf{b}_i\|^2]$$

Em que $\boldsymbol{\Delta}$ é o fator de precisão relativo. Adicionando "pseudo" observações ao conjunto de dados, podemos converter a equação anterior em uma soma de quadrados não linear simples. O vetor resposta aumentado e o vetor dos modelos são representados por:

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{0} \end{bmatrix}, \qquad \tilde{\mathbf{f}}_i(\boldsymbol{\beta}, \mathbf{b}_i) = \begin{bmatrix} \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i) \\ \boldsymbol{\Delta}\mathbf{b}_i \end{bmatrix}.$$

Após a nova notação, teremos uma nova soma de quadrados, que pode ser expressada da seguinte maneira:

$$\sum_{i=1}^{M} \left\| \tilde{\mathbf{y}}_i - \tilde{\mathbf{f}}_i(\boldsymbol{\beta}, \mathbf{b}_i) \right\|^2,$$

Que também é chamada de soma de quadrados aumentada (*Agumented Sum of Squares*). Portanto, a estimação de $\boldsymbol{\beta}$ e $\mathbf{b_i}$ consiste em uma resolução de mínimos quadrados não linear, sendo necessárias as derivadas em relação aos parâmetros. Como $\tilde{\mathbf{f}}_i(\boldsymbol{\beta}, \mathbf{b}_i)$ é não linear em pelo menos um dos seus parâmetros, métodos computacionais de otimização são necessários. O algoritmo de Gauss-Newton (Bates e Watts, 1988; Pinheiro e Bates, 2000) é apresentado em seguida, adaptado para os parâmetros de efeitos fixos ($\boldsymbol{\phi}$). Como visto anteriormente, neste procedimento $\mathbf{f}(\boldsymbol{\phi})$ é substituída por uma aproximação via séries de Taylor em torno das estimativas atuais $\widehat{\boldsymbol{\phi}}^{(w)}$, como segue:

$$\mathbf{f}(\boldsymbol{\phi}) \approx \mathbf{f}(\widehat{\boldsymbol{\phi}}^{(w)}) + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\phi}'}\Big|_{\widehat{\boldsymbol{\phi}}^{(w)}} (\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}^{(w)}).$$

Na w-ésima iteração, o parâmetro de incremento $\widehat{\boldsymbol{\delta}}^{(w+1)} = \widehat{\boldsymbol{\phi}}^{(w+1)} - \widehat{\boldsymbol{\phi}}^{(w)}$ deve ser calculado, como solução do seguinte problema de mínimos quadrados:

$$\left\| \left[\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\phi}}^{(w)})\right] - \frac{\partial \mathbf{f}}{\partial \boldsymbol{\phi}'}\Big|_{\widehat{\boldsymbol{\phi}}^{(w)}} (\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}^{(w)}) \right\|^2.$$

A cada iteração neste passo é calculada uma nova estimativa para $\widehat{\boldsymbol{\phi}}^{(w+1)} = \widehat{\boldsymbol{\delta}}^{(w+1)} + \widehat{\boldsymbol{\phi}}^{(w)}$ e com base neste, é realizado um novo cálculo para a soma de quadrados. Se esta for menor que a soma obtida em $\widehat{\boldsymbol{\phi}}^{(w)}$, esse valor é utilizado novamente e o algoritmo continua para o próximo passo ou a convergência é declarada. Caso contrário, a nova estimativa será $\widehat{\boldsymbol{\phi}}^{(w)} + \widehat{\boldsymbol{\delta}}^{(w+1)}/2$ e o procedimento é repetido.

Em termos de $\boldsymbol{\beta}$ e $\mathbf{b_i}$, a matriz de derivadas nas somas de quadrados destacadas anteriormente serão $(i = 1, \ldots, M)$:

$$\left.\frac{\partial \tilde{\mathbf{f}}_i(\boldsymbol{\beta}, \mathbf{b}_i|\boldsymbol{\Delta})}{\partial \boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i{}^{(w)}} = \widetilde{\mathbf{X}}_i^{(w)} = \begin{bmatrix} \widehat{\mathbf{X}}_i^{(w)} \\ \mathbf{0} \end{bmatrix},$$

$$\left.\frac{\partial \tilde{\mathbf{f}}_i(\boldsymbol{\beta}, \mathbf{b}_i|\boldsymbol{\Delta})}{\partial \mathbf{b}_i'}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i{}^{(w)}} = \tilde{\mathbf{Z}}_i^{(w)} = \begin{bmatrix} \widehat{\mathbf{Z}}_i^{(w)} \\ \boldsymbol{\Delta} \end{bmatrix},$$

Portanto, a solução por mínimos quadrados a ser resolvida em cada iteração do algoritmo de Gauss-Newton será:

$$\sum_{i=1}^{M} \left\| \left[\tilde{\mathbf{y}}_i - \tilde{\mathbf{f}}_i(\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i^{(w)}) \right] - \widetilde{\mathbf{X}}_i^{(w)} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(w)} \right) - \tilde{\mathbf{Z}}_i^{(w)} (\mathbf{b}_i - \hat{\mathbf{b}}_i^{(w)}) \right\|^2$$

São necessários chutes iniciais para $\boldsymbol{\beta}$ e $\mathbf{b_i}$, além da definição da estrutura de covariâncias entre as observações do mesmo indivíduo.

2) *LME Step*: O segundo passo consiste em atualizar a estimativa do fator de precisão relativo ($\boldsymbol{\Delta}$). Utiliza-se uma aproximação via séries de Taylor de primeira ordem em torno das estimativas de $\boldsymbol{\beta}$ e $\mathbf{b}_i$. Seja:

$$\widehat{\mathbf{X}}_i^{(w)} = \left.\frac{\partial \tilde{\mathbf{f}}_i}{\partial \boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i{}^{(w)}},$$

$$\widehat{\mathbf{Z}}_i^{(w)} = \left.\frac{\partial \tilde{\mathbf{f}}_i}{\partial \mathbf{b}_i{}'}\right|_{\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i{}^{(w)}},$$

$$\widehat{\mathrm{w}}_i^{(w)} = \mathbf{y}_i - \mathbf{f}_i \left(\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i{}^{(w)} \right) + \widehat{\mathbf{X}}_i^{(w)} \widehat{\boldsymbol{\beta}}^{(w)} + \widehat{\mathbf{Z}}_i^{(w)} \hat{\mathbf{b}}_i{}^{(w)},$$

O logaritmo da função de verossimilhança utilizada para estimativa de $\boldsymbol{\Delta}$ será:

$$\begin{aligned} & \mathrm{l}_{\mathrm{MV}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}|\mathbf{y}) \\ & = -\frac{\mathrm{N}}{2} \log(2\pi\sigma^2) \\ & -\frac{1}{2} \sum_{i=1}^{M} \left\{ \log|\boldsymbol{\Sigma}_i(\boldsymbol{\Delta})| + \sigma^{-2} [\widehat{\mathbf{w}}_i^{(w)} - \widehat{\mathbf{X}}_i^{(w)} \boldsymbol{\beta}]' \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}) [\widehat{\mathbf{w}}_i^{(w)} - \widehat{\mathbf{X}}_i^{(w)} \boldsymbol{\beta}] \right\} \end{aligned}$$

Em que $\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}) = \mathbf{I} + \widehat{\mathbf{Z}}_i^{(w)} \boldsymbol{\Delta}^{-1} \boldsymbol{\Delta}^{-\prime} \widehat{\mathbf{Z}}_i^{(w)\prime}$. Essa verossimilhança é baseada na distribuição marginal de Y. A estimação também pode ser feita por meio da função de máxima verossimilhança restrita, invariante aos parâmetros de posição. O logaritmo da função de máxima verossimilhança restrita é dado por:

$$\mathrm{l}_{\mathrm{MVR}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}|\mathbf{y}) = \mathrm{l}_{\mathrm{MV}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}|\mathbf{y}) - \frac{1}{2} \sum_{i=1}^{M} \log \left| \sigma^{-2} \widehat{\mathbf{X}}_i^{(w)\prime} \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}) \widehat{\mathbf{X}}_i^{(w)} \right|$$

O algoritmo consiste na alternância entre esses dois algoritmos até que algum critério de convergência seja obedecido.

3.4. Curvas de Lactação Sob Abordagem dos Modelos Não-Lineares Mistos

O primeiro modelo para ajuste de curvas de lactação foi proposto por Brody *et al*. (1923), que teve como ideia a utilização da função exponencial para descrever a etapa de declínio da lactação de vacas leiteiras após o parto. Desde então, diversos modelos vêm sendo propostos para caracterizar a produção leiteira de cabras, vacas leiteiras, búfalo, dentre outros.

O modelo sugerido por Sikka (1950) descreve a curva de lactação no formato de sino, onde o pico (produção máxima da lactação) ocorre na metade da curva; Nelder (1966) propôs um modelo conhecido também como polinomial inverso no intuito de identificar um efeito de saturação, induzindo ao decréscimo da curva; uma curva tipo Gama para descrever o comportamento da lactação foi proposta por Wood (1967) e tem sido amplamente utilizada em estudos de curvas de lactação (Ghavi Hossein-Zadeh, 2015). Essa equação é composta por três parâmetros, frequentemente denominados por a_i, b_i e c_i, os quais representam a produtividade no início da lactação e as taxas de ascensão e declínio após o pico de lactação, respectivamente.

Após o surgimento da equação de Wood, muitos dos modelos propostos posteriormente foram baseados em modificações dessa função. A equação proposta por Cobby e Le Du (1978), por exemplo, foi uma modificação feita no modelo de Wood onde após a produção máxima (pico) a produção de leite decai no formato de linha reta. Dhanoa (1981) sugeriu uma forma alternativa para o modelo de Wood, em que o parâmetro b é inserido como sendo o tempo para chegar ao pico de lactação, visando também reduzir a correlação entre os parâmetros que caracterizam a produção de leite. Papajcsik e Bodero (1988) propuseram um modelo semelhante ao de Wood, porém eliminando o efeito do parâmetro de ascensão b, substituído por 1 na equação. O modelo de Rook *et al.* (1993) faz parte da classe de modelos mecanicistas, em que sua formulação é baseada em fenômenos físicos ou biológicos, baseado em um conjunto de equações que representam a proliferação e morte celulares na glândula mamária, resultando em uma equação com 4 parâmetros (Ghavi Hossein-Zadeh, 2015). A equação de Cappio-Borlino *et al*. (1995) é ideal quando a queda de produção ocorre de maneira acentuada após o pico de lactação.

Na abordagem tradicional, estima-se os parâmetros de cada modelo individualmente, onde os indivíduos são separados por classes de acordo com as características físicas, ambientais ou genéticas. Já na abordagem de efeitos mistos, essas características são inseridas como efeitos fixos no modelo, que também dispõe dos efeitos

aleatórios, representando o desempenho individual. As equações para as duas abordagens seguem na Tabela 1.

Tabela 1. Equações não lineares para curvas de lactação, com respectivos autores.

Autor (es)	Equação (efeitos fixos)
Brody *et al.* (1923)	$y_{ij} = a_i e^{-c_i t_{ij}}$
Sikka (1950)	$y_{ij} = a_i e^{(b_i t_{ij} - c_i {t_{ij}}^2)}$
Nelder (1966)	$y_{ij} = \frac{t_{ij}}{a_i + b_i t_{ij} + c_i {t_{ij}}^2}$
Wood (1967)	$y_t = a_i {t_{ij}}^{b_i} e^{-c_i t_{ij}}$
Cobby e Le Du (1978)	$y_{ij} = a_i - b_i t_{ij} - a_i e^{-c_i t_{ij}}$
Dhanoa (1981)	$y_{ij} = a_i {t_{ij}}^{b_i c_i} e^{-c_i t_{ij}}$
Papajcsik e Bodero (1988)	$y_{ij} = a_i t_{ij} e^{-c_i t_{ij}}$
Rook (1993)	$y_{ij} = a_i \left(\frac{1}{1 + \frac{b_i}{c_i + t_{ij}}} \right) e^{-d_i t_{ij}}$
Cappio-Borlino *et al.* (1995)	$y_{ij} = a_i {t_{ij}}^{b_i \exp(-c_i t_{ij})}$

Nota: y_{ij} é a produtividade de leite do i-ésimo animal no j-ésimo dia de controle, medido a partir do início da lactação; a_i, b_i, c_i e d_i são os parâmetros que definem a escala e o formato da curva de lactação para o i-ésimo animal, variando de acordo com as equações e t_{ij} é o tempo correspondente ao i-ésimo animal no j-ésimo dia de controle, medido a partir do início da lactação.

As equações acima estão escritas de acordo com o método tradicional, de efeitos fixos. Tomando o modelo de Wood (1967) como referência, e levando em conta o nível de produção dos indivíduos como efeito fixo, podemos escrevê-lo na forma de modelo não linear misto da seguinte maneira:

$$y_{ij} = (\mu_a + NP_i + \xi_{ai}).\, {t_{ij}}^{(\mu_b + NP_i + \xi_{bi})} e^{-(\mu_c + NP_i + \xi_{ci}).t_{ij}} + \varepsilon_{ij}.$$

Esse modelo assume que $a_i = \mu_a + NP_i + \xi_{ai}$, $b_i = \mu_b + NP_i + \xi_{bi}$ e $c_i = \mu_c + NP_i + \xi_{ci}$, sendo μ_a, μ_b e μ_c as médias dos efeitos fixos de cada parâmetro; NP_i é o efeito do nível de produção do i-ésimo indivíduo; e ξ_{ai}, ξ_{bi} e ξ_{ci} são os efeitos aleatórios individuais para cada parâmetro do modelo; e ε_{ij} é o efeito residual associado à observação y_{ij}.

3.5. Cálculo de medidas da lactação com base nas equações propostas

Através de cada uma das equações acima, medidas importantes que caracterizam a produção leiteira podem ser estimadas, a exemplo da produção total, pico de lactação, produção no pico e persistência. A produção total é a quantidade de leite produzida (em kg) por cada animal no decorrer de toda a lactação, normalmente padronizada para 305 dias. Seja $f(\boldsymbol{\phi}, t)$ uma das equações descritas na Tabela 2 no tempo t a produção total pode ser calculada como abaixo (Ferreira *et al.*, 2015):

$$PT = \int_0^{305} f(\boldsymbol{\phi}, t)dt.$$

Ainda segundo Ferreira, a produção no pico de lactação e o tempo até o pico são calculados pelo ponto de máximo da curva de lactação, juntamente com o ponto no eixo t correspondente a esse ponto. O tempo até o pico ($t_{y.máx}$) é o ponto de t que maximiza $f(\boldsymbol{\phi}, t)$, ou seja, o ponto de t no qual $f'(\boldsymbol{\phi}, t) = 0$. A produção no pico ($Y_{t.máx}$), ponto correspondente a $t_{y.máx}$ no eixo Y, pode ser obtido pela substituição do ponto de máximo $t_{y.máx}$ em $f(\boldsymbol{\phi}, t)$. Quanto à persistência (PS) da lactação, esta pode ser calculada por diferentes métodos (Cobuci *et al.*, 2003). Porém, um dos métodos utilizados é o proposto por Wood (1967), o qual se baseia na obtenção dos seus parâmetros, e pode ser calculada por: $PS = -(b_i + 1).\ln(c_i)$ para o i-ésimo animal.

3.6. Medidas de qualidade de ajuste e seleção de modelos

Quando temos como objetivo descrever o comportamento de uma variável em função de outra por meio dos modelos não-lineares, usualmente são testadas mais de uma função, pois estamos buscando o modelo que melhor descreva essa relação. Segundo Silveira *et al*. (2011), as medidas de qualidades de ajuste são ferramentas estatísticas que permitem comparar diferentes modelos e encontrar o mais indicado na situação em que trabalhamos. O cálculo de cada uma dessas medidas varia de acordo com suas conjecturas.

Podemos citar como importantes medidas de adequabilidade e seleção de modelos os critérios de Akaike e Bayesiano (AIC e BIC), o desvio padrão residual, a raíz do erro quadrático médio (*Root Means Square Error - RMSE*), dentre outras. Usualmente utiliza-se também a estatística e o teste de Durbin-Watson para avaliação da autocorrelação residual nos modelos. Essas medidas são descritas abaixo:

i) AIC e BIC: O critério de informação de Akaike (1973) mede a adequabilidade de um modelo por meio da distância de Kullback-Leibler (K-L), de acordo com a equação abaixo:

$$\mathrm{AIC} = -2\mathrm{logL}\left(\hat{\boldsymbol{\phi}}\right) + 2\mathrm{n_{par}},$$

em que $\mathrm{L}\left(\hat{\boldsymbol{\phi}}\right)$ é o máximo da função de verossimilhança e $\mathrm{n_{par}}$ é a quantidade de parâmetros considerada no modelo. Já o critério de informação bayesiano (BIC), proposto por Schwarz (1978), é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O cálculo do critério BIC é definido pela medida abaixo:

$$\mathrm{BIC} = -2\mathrm{logL}\left(\hat{\boldsymbol{\phi}}\right) + \mathrm{n_{par}}\log(\mathrm{M}),$$

em que $\mathrm{L}\left(\hat{\boldsymbol{\phi}}\right)$ é o máximo da função de verossimilhança e $\mathrm{n_{par}}$ é a quantidade de parâmetros do modelo. Segundo Pinheiro e Bates (2000), sob essas definições, melhores modelos serão os que apresentam menores valores de AIC e BIC.

ii) RMSE: A raiz do erro quadrático médio (*Root Mean Squared Error - RMSE*) funciona como um desvio padrão generalizado (Ghavi Hossein-Zadeh, 2015) e pode ser calculado pela seguinte expressão:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})}{\mathrm{n}}},$$

em que $\mathrm{y_i}$ é a i-esima observação e $\hat{\mathrm{y}}_\mathrm{i}$ é a estimação de $\mathrm{y_i}$ pelo respectivo modelo e n é o número de observações. Como essa quantidade está associada à soma de quadrados dos erros, o melhor modelo será o que apresentar o menor RMSE.

iii) Estatística de Durbin-Watson: A estatística de Durbin-Watson (Durbin, 1970), também avaliada por teste estatístico, tem por finalidade identificar a presença de autocorrelação entre os resíduos do modelo especificado. Essa estatística é dada por:

$$\mathrm{DW} = \frac{\sum_{\mathrm{k}=1}^{\mathrm{M}}(\mathrm{e_k} - \mathrm{e_{k-1}})^2}{\sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{e_k}^2}.$$

Sendo $\mathrm{e_k}$ o resíduo da k-ésima observação. Essa medida pode variar entre 0 e 4, e podemos concluir que o modelo não possui correlação com valores próximos de 2. A suposição do teste é a de que os erros do modelo de regressão são gerados por um processo auto regressivo de primeira ordem, como dado no modelo abaixo:

$$\mathrm{e_k} = \rho \mathrm{e_k} + \mathrm{a_k}$$

Sendo $\mathrm{a_k} \sim \mathrm{N}(0, \sigma^2)$ e ρ é o parâmetro de correlação, obedecendo $-1 < \rho < 1$, no qual o teste é baseado. As hipóteses são dadas abaixo:

$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

A hipótese nula do teste de Durbin-Watson assume que os erros são independentes, ou seja, se for concluído que $\rho = 0$, o erro e_k será distribuído normalmente com média 0 e variância constante. Já considerando a hipótese alternativa, a conclusão é de que $\rho \neq 0$, ou seja, de que existe dependência entre os resíduos de acordo com uma estrutura auto regressiva de primeira ordem (Piccardi *et al.*, 2017).

3.7. Exemplo de aplicação

Suponha que nosso banco de dados seja composto por 6 indivíduos, correspondentes a dois diferentes grupos, digamos A e B. O conjunto de dados é composto pela observação da produtividade de cada indivíduo em um determinado dia (sem restrições) e do grupo correspondente. Os detalhes são mostrados na Tabela 2.

Tabela 2. Dados referentes a cada animal: o grupo do qual faz parte, o dia de coleta e a produtividade no respectivo dia.

Animal	Grupo	Produtividade (kg)	Dia	Animal	Grupo	Produtividade (kg)	Dia
1	A	13,3	29	4	B	40,4	33
1	A	18,3	60	4	B	43,4	59
1	A	14,7	101	4	B	31,0	94
1	A	7,6	138	4	B	34,4	120
1	A	12,8	178	4	B	32,7	156
1	A	12,5	211	4	B	29,4	180
1	A	12,1	241	4	B	30,4	207
1	A	12,6	272	4	B	23,4	240
1	A	14,4	302	4	B	23,0	269
2	A	10,8	22	4	B	16,6	304
2	A	10,2	51	5	B	18,0	15
2	A	8,8	79	5	B	22,5	56
2	A	7,2	120	5	B	19,0	99
2	A	8,8	163	5	B	16,2	144
2	A	11,0	207	5	B	13,5	189

Continua

Conclusão

Animal	Grupo	Produtividade (kg)	Dia	Animal	Grupo	Produtividade (kg)	Dia
2	A	9,4	248	5	B	11,5	234
2	A	6,6	292	5	B	9,5	268
3	A	21,6	20	6	B	12,3	9
3	A	20,0	58	6	B	11,4	42
3	A	21,4	79	6	B	12,5	72
3	A	22,0	117	6	B	10,1	104
3	A	22,0	146	6	B	11,1	135
3	A	19,4	176	6	B	9,2	163
3	A	21,2	203	6	B	7,6	195
3	A	19,6	239	6	B	6,4	224
3	A	11,6	266	6	B	6,1	254

Observa-se também que a quantidade de observações por indivíduo (variando de 7 a 10 por animal) é diferente. Quando observamos graficamente, podemos ver que existe uma diferença entre a produtividade dos grupos e indivíduos, como segue na Figura 1.

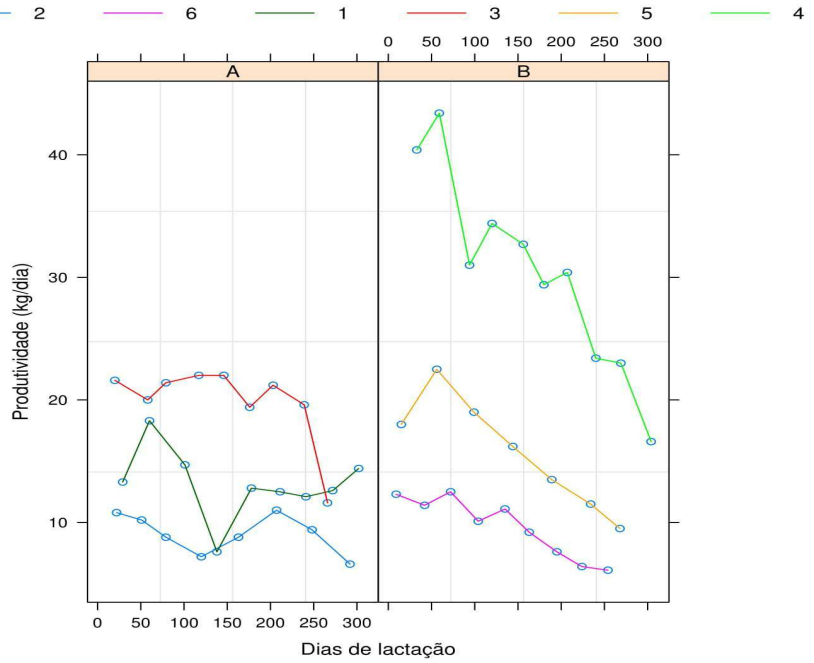


Figura 1. Observações das produtividades de acordo com o dia de lactação dos indivíduos pertencentes aos grupos A e B.

Inserindo o efeito de grupos no ajuste, essa diferença pode ser captada, ao mesmo tempo em que os efeitos aleatórios individuais também são mensurados. Para o exemplo, podemos considerar a equação de Wood (1967), a observação no tempo j do i-ésimo indivíduo é dada abaixo:

$$y_{ij} = a_i t_{ij}^{b_i} e^{-c_i t_{ij}} + \varepsilon_{it}$$

Sendo a_i o parâmetro que indica a produção inicial, b_i representa a taxa de ascensão e c_i a taxa de declínio após o pico de lactação. Porém, essa representação do modelo de Wood se refere à abordagem tradicional, a qual consiste em um ajuste por indivíduo.

Se considerarmos a abordagem de efeitos mistos, o efeito dos respectivos grupos de níveis de produtividade (A e B) pode ser inserido como efeito fixo, e os efeitos aleatórios (correspondentes aos indivíduos 1 a 6), facilitando a escolha do melhor indivíduo. Na abordagem dos modelos não lineares mistos com os efeitos de níveis de produção, o modelo de Wood (1967) para o i-esimo indivíduo na equação anterior pode ser representado por:

$$y_{ij} = (\mu_a + NP_i + \xi_{ai}).\, t_{ij}^{(\mu_b + NP_i + \xi_{bi})} e^{-(\mu_c + NP_i + \xi_{ci}).t_{ij}} + \varepsilon_{ij}.$$

Ou seja, a estimação de cada parâmetro do modelo pode ser dividida entre o efeito do respectivo grupo (NP_i), e o efeito individual (ξ_{ai}, ξ_{bi} e ξ_{ci} como efeitos aleatórios de a_i, b_i e c_i do i-esimo indivíduo).

Podemos observar inicialmente estes dados na Figura 1. O eixo X representa o tempo de lactação, usualmente padronizado para 305 dias (contabilizados após o parto), e no eixo Y podemos observar as produtividades nos respectivos dias em que se foi feita a coleta de dados, como descrito na Tabela 2.

Inserindo o efeito de grupos, para cada um dos três parâmetros temos 2 diferentes efeitos. Portanto, no nosso modelo será necessário a estimativa de 6 efeitos fixos (três efeitos para A, B). Logo, são necessários chutes iniciais para cada um desses parâmetros, além dos chutes para os efeitos aleatórios, que segundo Lindstrom e Bates (1990) geralmente são atribuídos zeros. Consideremos o nosso chute inicial para o vetor de efeitos fixos como sendo $\boldsymbol{\beta}^0 = [10 \;\; 10\; 0{,}05 \;\; 0{,}05 \;\; 0{,}001 \;\; 0{,}001]'$ e o chute para efeitos aleatórios o i-ésimo indivíduo (para todo i) dado por $\mathbf{b}_i = [0 \quad 0 \quad 0]'$, como sugerido por Lindstrom e Bates (1990). O processo iterativo, alternando entre os passos descritos anteriormente segue na Tabela 3.

Tabela 3. Somas de quadrados e logaritmo da função de verossimilhança de acordo com os dois passos descritos acima.

Passo (nº de iterações)	Soma de Quadrados	Loglik[1]	Critério de convergência
Iteração 1			
PNLS[2] (7)	261,3958		0,7915
LME[3] (26)		-131,5174	0,6423
Iteração 2			
PNLS (2)	261,3329		0,0027
LME (2)		-129,0340	0,0014
Iteração 3			
PNLS (1)	261,3329		<0,0001
LME (1)		-129,0256	<0,0001

1: Logaritmo da função de verossimilhança; 2: Minimização da soma de quadrados; 3: Maximização da verossimilhança via modelos lineares mistos.

Após a iteração 3, o ajuste apresentou convergência, tendo os seguintes resultados para o modelo:

Tabela 4. Estimativas dos parâmetros considerando os grupos A e B como efeitos fixos.

Parâmetros / Critérios	Estimativa	Erro padrão	Graus de Liberdade	p
a (grupo A)	10,7745	4,9625	41	0,0358
a (grupo B)	12,3056	4,1688	41	0,0052
b (grupo A)	0,1165	0,1059	41	0,2777
b (grupo B)	0,2353	0,0685	41	0,0014
c (grupo A)	0,0020	0,0010	41	0,0663
c (grupo B)	0,0048	0,0007	41	<0,0001
AIC	278,0513			
BIC	297,5637			
Loglik[1]	-129,0256			

1: Logaritmo da função de verossimilhança.

A Figura 2 contém as curvas na média geral e nas médias dos efeitos fixos dos grupos A e B. Podemos observar que a média geral de produtividade (área abaixo da curva) do grupo B é maior que a do grupo A, o que era esperado observando a Figura 1, vista anteriormente.

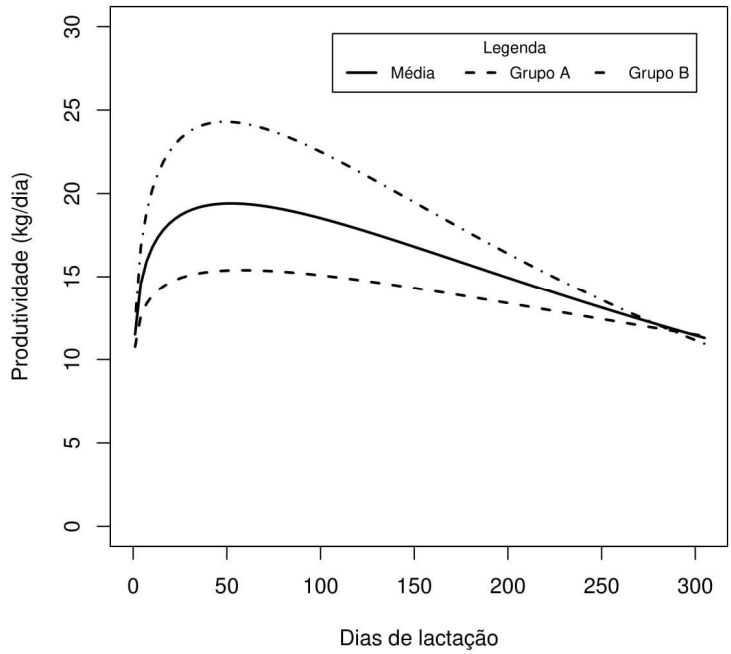


Figura 2. Curva média dos 6 indivíduos (linha contínua), em relação à curva média dos grupos A e B.

Podemos observar que existe uma pequena diferença entre os valores de a, b e c dos grupos A e B, a qual pôde ser obtida por meio da inserção de efeitos fixos. Os valores de 10,7745 e 12,3056, por exemplo, nos indicam as produções iniciais dos grupos A e B, respectivamente. Ou seja, podemos concluir que, em média, os indivíduos do grupo B produzem cerca de 1,5 kg de leite a mais que os do grupo A no início da lactação. A mesma interpretação é válida para os demais parâmetros. A Figura 2 mostra a diferença entre as curvas dos dois grupos.

Observando os grupos separadamente, podemos observar as curvas individuais em torno do efeito médio de cada grupo. Nota-se que o indivíduo 3 é o que mais se destaca no grupo A com produção inicial e pico de lactação superiores à média do grupo e aos demais. Já no grupo B o indivíduo 4 é o mais produtivo, com sua curva muito acima da curva média desse grupo.

Portanto, foram estimados os parâmetros para efeitos fixos, os quais usualmente interferem na análise e na seleção de indivíduos, e para os efeitos aleatórios simultaneamente. As medidas importantes para lactação (pico, persistência, etc.) podem ser calculadas individualmente a partir desses resultados.

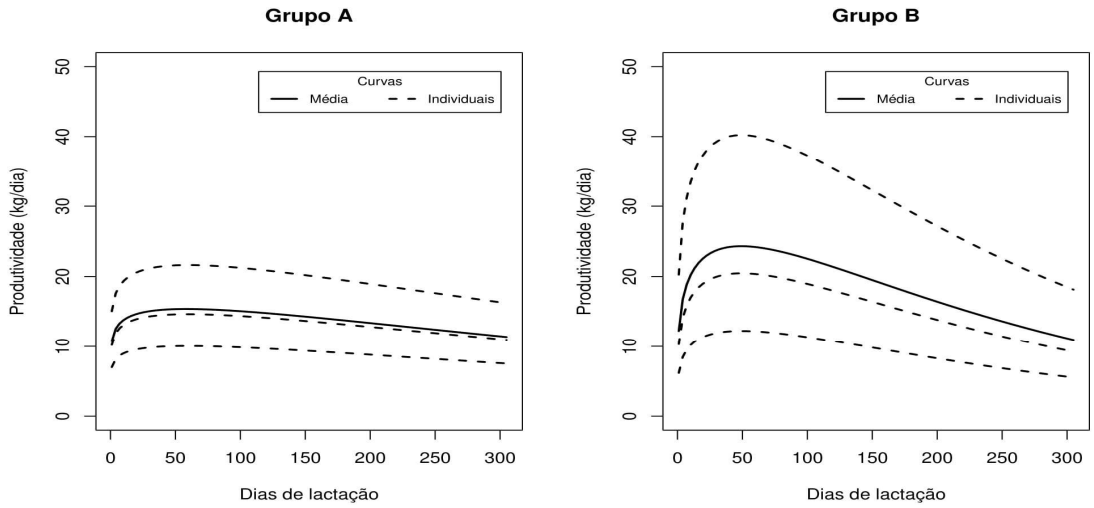


Figura 3. Curvas de lactação de cada um dos seis indivíduos em análise comparados com a média dos seus respectivos grupos.

4. Análise de Agrupamento

4.1. Introdução

Em algumas situações, temos uma grande quantidade de variáveis a serem inseridas como efeito fixo, comparado com o tamanho da amostra. No caso do presente trabalho, como efeito fixo tínhamos 4 variáveis: idade (dividida em 4 grupos), o número de ordenhas (3 grupos), a composição racial (3 grupos) e o grupo contemporâneo (139 grupos). A combinação destas resulta em 5.004 combinações de características, o que torna inviável o ajuste. Portanto, para resumir essa informação, os efeitos fixos foram considerados os grupos formados com base na análise de agrupamento dessas 4 variáveis.

De acordo com Mingoti (2017), a análise de agrupamentos, também conhecida como análise de *clusters*, tem como objetivo dividir os elementos de uma amostra, ou população, em grupos de forma que os elementos do mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características.

Portanto, se temos à disposição um conjunto de variáveis de interesse, essa técnica permite que identifiquemos a semelhança entre um conjunto de indivíduos de acordo com estas características, permitindo a seleção ou classificação dos mesmos segundo algum critério previamente especificado. Sua grande aplicabilidade, juntamente com sua simplicidade a tornam atrativa em diversas áreas de pesquisa.

A análise de *clusters* é baseada na construção da matriz de distâncias entre os indivíduos de acordo com suas respectivas características. Na literatura existem diversas maneiras para a construção dessas matrizes, as quais variam principalmente de acordo com o tipo de variável que se considera. Além da matriz de distâncias, é necessário que um método de agrupamento seja especificado, para que o procedimento de aglomeração dos indivíduos seja feito. As seções e subseções seguintes destacam alguns métodos para formação da matriz de distâncias e de agrupamento das variáveis.

4.2. Medidas de Proximidade

Quando o objetivo da pesquisa consiste no agrupamento de variáveis, o primeiro passo é a determinação de uma matriz proximidade entre as variáveis. Segundo Ferreira (2011), a proximidade é um termo utilizado para indicar similaridade ou dissimilaridade, que são medidas pelas distâncias. Seja n o tamanho da amostra (quantidade de indivíduos), a matriz de proximidade é uma matriz de dimensão n x n composta pelas respectivas medidas de proximidade entre os indivíduos. Portanto, a medida da i-ésima linha e j-ésima coluna representa a distância (similaridade ou dissimilaridade) entre o i-esimo e o j-ésimo indivíduos.

Considerando as observações de p variáveis associadas a n indivíduos (indexados por i e j, respectivamente), a matriz de observações pode ser representada da seguinte forma (FERREIRA, 2011):

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}'_1 \\ \boldsymbol{y}'_2 \\ \vdots \\ \boldsymbol{y}'_i \\ \vdots \\ \boldsymbol{y}'_n \end{bmatrix} = [\boldsymbol{y}_{(1)} \quad \cdots \quad \boldsymbol{y}_{(j)} \quad \cdots \quad \boldsymbol{y}_{(p)}].$$

Portanto, para cada indivíduo teremos observações associadas a p variáveis, podendo estas serem de natureza quantitativa (discreta ou contínua) ou qualitativa (nominal ou ordinal). Na maioria dos casos, a escolha da medida de proximidade a ser utilizada depende do tipo de variáveis disponíveis no estudo, ou seja, depende da matriz de observações. No presente trabalho serão apresentadas a distância Euclidiana e a distância de Gower (1971).

4.2.1. Distância Euclidiana

A medida de proximidades mais comum é a distância Euclidiana. Esta é utilizada quando se dispõe apenas de variáveis quantitativas. Sejam os indivíduos i e j ($i \neq j$) e seja X_k a k-ésima variável ($k = 1, \dots, p$) em estudo, a distância Euclidiana entre dois indivíduos pode ser calculada da seguinte maneira (MINGOTI, 2017):

$$d(i,j) = \left[\sum_{k=1}^{p} (X_{ki} - X_{kj})^2\right]^{1/2}$$

Vale ressaltar que a Distância Euclidiana é uma medida de dissimilaridade, portanto, quanto maior $d(l,k)$, maior será a distância entre os indivíduos l e k. Segundo CRUZ (2011), como em estudos de melhoramento é praticamente impossível avaliar um conjunto de características não relacionadas, o uso da distância euclidiana tem sido indiscriminado e se mostrado de grande utilidade mesmo nas situações em que a independência entre as características mensuradas não é constatada.

4.2.2. Distância de Gower

O algoritmo idealizado por Gower (1971) pode ser utilizado quando se dispõe de variáveis quantitativas e qualitativas (nominais, ordinais, ou dos dois tipos) e é utilizado principalmente quando temos os dois casos. Este algoritmo considera que a distância entre os indivíduos i e j é calculada da seguinte maneira Gower (1971):

$$D_{ijk} = \frac{\sum_{k=1}^{p} W_{ijk}.S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}$$

Em que k é o número de variáveis ($k = 1, \dots, p$); i e j são os dois indivíduos avaliados; W_{ijk} é uma variável indicadora na qual se atribui 1 para comparações válidas e 0 para comparações inválidas (quando temos observações ausentes); e S_{ijk} é a contribuição da k-ésima variável na similaridade entre os indivíduos i e j. Caso a variável k seja quantitativa (contínua), atribui-se a distância de Manhattan a S_{ijk}, como abaixo:

$$S_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

Em que x_{ik} e x_{jk} são as observações dos indivíduos i e j relacionadas à variável k; e R_k é a amplitude da variável k. No caso em que a variável k é qualitativa nominal, se o valor de k é o mesmo para ambos os indivíduos, atribuímos $S_{ijk} = 1$, caso contrário, $S_{ijk} = 0$.

4.3. Métodos de Agrupamento

Após a determinação e cálculo da matriz de proximidades, o próximo passo consiste na escolha do método de agrupamento das variáveis, o qual é baseado na matriz de proximidades. Basicamente, os métodos de agrupamento são divididos em duas classes: os hierárquicos (ou aglomerativos) e os não hierárquicos (divisivos). Segundo Mingoti (2017), as técnicas hierárquicas, na maioria das vezes, são utilizadas em análises exploratórias dos dados com o intuito de identificar possíveis agrupamentos e o valor provável do número de grupos. Neste caso, a quantidade inicial de *clusters* é igual ao tamanho da amostra, ou seja, a quantidade de indivíduos inseridos na análise (n). No decorrer do processo de agrupamento, de acordo com a matriz de proximidades, esses indivíduos se agrupam até, segundo algum critério, serem determinadas a quantidade de grupos e de indivíduos pertencentes aos mesmos. Já para o uso de técnicas não hierárquicas, é necessário que o valor do número de grupos já esteja pré-especificado pelo pesquisador, e os grupos são formados por separação. A seguir são apresentados dois métodos de agrupamento frequentemente utilizados em análises de agrupamento de indivíduos:

4.3.1. Método da Ligação Simples

Também conhecido como *Single Linkage* ou Método do Vizinho Mais Próximo, este método consiste no agrupamento dos indivíduos mais semelhantes entre si, sendo essa semelhança avaliada por meio das medidas de proximidade.

Dada a matriz de proximidades (de dimensão n x n), sendo n a quantidade de indivíduos, a cada passo do algoritmo os dois mais próximos são agrupados. Suponha que tenhamos um grupo formado por 5 indivíduos, numerados de 1 a 5, e seja $d(i,j) = d_{ij}$ a distância euclidiana entre os indivíduos i e j, descrita anteriormente. Portanto, teremos a seguinte matriz de distâncias:

$$D_{4x4} = \begin{bmatrix} & 1 & 2 & 3 & 4 \\ 1 & 0 & d_{12} & d_{13} & d_{14} \\ 2 & d_{21} & 0 & d_{23} & d_{24} \\ 3 & d_{31} & d_{32} & 0 & d_{34} \\ 4 & d_{41} & d_{42} & d_{43} & 0 \end{bmatrix}$$

Vale ressaltar também que $d_{ij} = d_{ji}$. Como a distância euclidiana avalia a dissimilaridade entre os indivíduos, quanto menor o valor de d_{ij}, mais similares serão os indivíduos i e j. Portanto, no primeiro estágio os indivíduos a serem agrupados serão aqueles que possuírem menor d_{ij}. Após o primeiro agrupamento, as matrizes de distância são recalculadas com a inclusão do primeiro grupo (agora com 2 indivíduos) e identifica-se novamente a menor distância, formando-se um novo grupo. O procedimento é repetido até que se tenha apenas um grupo, contendo todos os indivíduos.

Após este procedimento, existem técnicas na literatura para determinação da quantidade de grupos a serem considerados. Uma dessas metodologias é o critério de Mojena (1977), que será posteriormente descrito.

4.3.2. Método UPGMA

Abreviação de *Unweighted Pair-Group Method Using Arithmetic Averages*, e também conhecido como método da ligação média não ponderada, o método UPGMA utiliza as médias aritméticas (não ponderadas) das medidas de dissimilaridade, evitando assim caracterizar a dissimilaridade por valores extremos (máximo e mínimo) entre os genótipos considerados. O procedimento é descrito por CRUZ (2011). Neste método, a distância entre o indivíduo k e um grupo formado pelos indivíduos i e j é dada por:

$$d_{ij(k)} = \frac{d_{ik} + d_{jk}}{2}$$

Ou seja, $d_{ij(k)}$ é dada pelo conjunto das distâncias dos pares de indivíduos (i e k) e (j e k). Considerando a distância entre dois grupos de indivíduos, sendo um formado pelos indivíduos i e j e o outro formado pelos elementos k, l e m, temos a seguinte distância:

$$d_{ij(klm)} = \frac{d_{ik} + d_{il} + d_{im} + d_{jk} + d_{jl} + d_{jm}}{2}$$

Ou seja, é a distância entre os pares de indivíduos pertencentes a cada grupo. Assim como o método do vizinho mais próximo, o procedimento termina quando se tem apenas um grupo. Para o caso de medidas de dissimilaridade como a distância euclidiana, os

indivíduos ou grupos que deverão se agrupar serão aqueles que apresentarem menor quantidade para esta medida.

4.4. Critério de Mojena

Mojena (1977) estabeleceu um critério onde o ponto de divisão dos grupos se baseia na maior amplitude dos pontos de fusão onde as variáveis são agrupadas, visando otimizar o agrupamento dos dados. O número de grupos deve ser definido a partir da estatística abaixo:

$$\alpha_j > \bar{\alpha} + \phi S_\alpha,$$

Em que α_j é a distância para o estágio de agrupamento correspondente a n-j+1 grupos ($j = 1, \dots, n$); $\bar{\alpha}$ e S_α são a média e o desvio padrão dos níveis de fusão, e ϕ é uma constante, que segundo Mojena (1977), deve variar de 2,75 a 3,50. Já Milligan & Cooper (1985) sugerem que esta constante deve ser igual a 1,25 baseado em estudo de simulação.

5. Seleção Genômica Ampla (SGA)

5.1. Introdução e modelo estatístico

Com o objetivo de utilizar informações diretas do DNA na seleção de indivíduos e predição dos seus méritos genéticos, Meuwissen *et al*. (2001) propôs a Seleção Genômica Ampla utilizando os métodos BayesA e BayesB aplicados a dados simulados. Segundo Resende (2012), essa metodologia permite alta eficiência seletiva, grande obtenção de ganhos genéticos com a seleção de baixo custo comparado com a seleção tradicional baseada em dados fenotípicos.

A Seleção Genômica é baseada em marcadores moleculares conhecidos como SNP's (polimorfismo de um único nucleotídeo – *Single Nucleotide Polymorphisms*), os quais possuem baixa taxa de mutação e facilidade de genotipagem, além de sua ampla distribuição dentro do genoma. O procedimento consiste em identificar, analisando resultados obtidos pelos SNP's, quais marcadores e a região do DNA influenciam na alteração em determinada característica fenotípica de interesse para com base nessa informação, estimar valores genéticos genômicos e selecionar animais geneticamente superiores. Dessa forma, os efeitos de marcadores podem ser aplicados para predição de

valores genéticos para indivíduos da mesma raça mesmo na fase inicial da vida, acarretando na economia de tempo e dinheiro. O modelo geral da seleção genômica considerando p marcadores SNP's e n indivíduos, considerando o i-ésimo indivíduo ($i = 1, \dots, n$) é dado por (Meuwissen *et al.*, 2001):

$$y_i = \mu + \sum_{k=1}^{p} x_{ik}\beta_k + e_i,$$

Em que y_i é a i-ésima observação da variável de interesse Y, μ é a média geral; x_{ik} é o valor absoluto correspondente ao k-ésimo marcador do i-ésimo indivíduo; e e_i é o erro associado a y_i. Na forma matricial, o modelo pode ser escrito da seguinte forma:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

Sendo $\mathbf{y}$ o vetor (nx1) que contém as observações da variável Y; $\boldsymbol{\mu}$ é o vetor de médias para todos os indivíduos (nx1); $\mathbf{X}$ é a matriz de incidência dos marcadores de cada indivíduo (nxp); $\boldsymbol{\beta}$ é o vetor dos efeitos dos marcadores (px1); e $\mathbf{e}$ é o vetor de erros (nx1).

O principal objetivo das metodologias de Seleção Genômica consiste na estimação do vetor dos efeitos dos marcadores ($\widehat{\boldsymbol{\beta}}$), seja para predição ou para seleção. No caso de predição, os efeitos dos marcadores são estimados por alguma metodologia (RR-BLUP, BLASSO, etc.), e posteriormente utilizados com o objetivo de predizer o mérito genético de quaisquer indivíduos da mesma espécie. Já na seleção, o mérito genético é utilizado para identificação dos melhores indivíduos de acordo com a característica de interesse, de modo que os animais superiores sejam selecionados.

5.2. Métodos Bayesianos de Seleção Genômica

Diferentes metodologias de Seleção Genômica têm sido propostas para estimação de efeitos de marcadores, e muitas delas envolvem a utilização de inferência bayesiana. Meuwissen *et al*. (2001) propôs a utilização dos métodos BayesA e BayesB, os quais foram comparados com o BLUP (*Best Linear Unbiased Predictor*), proposto por Henderson (1974), e com o método de mínimos quadrados utilizado por Lande e Thompson (1990), obtendo melhores resultados em termos de acurácia. Os métodos bayesianos na SGA se diferem pelas distribuições assumidas à priori para os efeitos de marcadores **β.** No presente trabalho, foi utilizado o método BLASSO (*Bayesian LASSO*).

O método BLASSO, ou Lasso Bayesiano, diferentemente dos métodos Bayes A e Bayes B, conta com o parâmetro de suavização (λ), o qual tem como objetivo aproximar os efeitos de marcadores de zero e pode ser estimado por métodos MCMC (Markov Chain Monte Carlo). Esse parâmetro de suavização é responsável por controlar a distribuição das variâncias dos marcadores. Essa metodologia também inclui um termo de variância comum para modelar ambos os termos, os resíduos e os efeitos genéticos dos marcadores (Park e Casella, 2008; De Los Campos *et al*., 2009). Segundo Silva *et al*. (2013), o BLASSO é um método de regressão Bayesiana penalizada em que o estimador dos efeitos de marcadores é dado por:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\{ (\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'(\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} |\beta_k| \right\},$$

em que λ é o parâmetro de regularização. Quando $\lambda = 0$, não existe regularização e quando $\lambda > 0$ existe o encolhimento (*shrinkage*) dos efeitos de marcadores em torno de zero, com a possibilidade de alguns serem identicamente iguais a zero, resultando em um procedimento simultâneo de estimação e seleção de variáveis.

Neste procedimento, ainda segundo Silva *et al*. (2013) a distribuição conjunta dos efeitos e marcadores $(\beta_1, \beta_2, \ldots, \beta_k)$ é $\prod_{k=1}^{p} N(0, \sigma^2_{\beta_k})$, onde $\sigma^2_{\beta_k} = \sigma^2\tau^2_k$, sendo σ^2 a variância residual, com distribuição Qui-quadrado inversa (χ^{-2}) e τ^2_k é o parâmetro de escala relacionado a cada marcador. Essa metodologia também assume que a distribuição desses parâmetros de escala $(\tau^2_1, \tau^2_2, \ldots, \tau^2_k)$ é um produto de distribuições exponenciais, $\prod_{k=1}^{p} \exp(\lambda)$, e que a distribuição à priori de λ é uma Gama(v_1, v_2). A variância genética aditiva para o cálculo da herdabilidade, mostrada posteriormente, é dada por $\sigma^2_a = 2\sum_{k=1}^{p} \sigma^2_{\beta_k} p_k(1 - p_k)$.

A utilização de uma distribuição exponencial dupla para o efeito dos marcadores concentra a maior parte da frequência (densidade) em valores próximos de zero, o que usualmente ocorre na distribuição dos efeitos dos marcadores. Adicionalmente, o Lasso Bayesiano proporciona um melhor aprendizado com os dados do que o Bayes A e o Bayes B (Gianola, 2013; Gianola *et al.*, 2009), as quais foram as metodologias inicialmente aplicadas em estudos de seleção genômica ampla.

5.3. Associação entre características de lactação

Após estimadas as variáveis genômicas de lactação é possível mensurar a associação entre elas. Para avaliar essa relação, foram utilizadas a correlação entre os valores genômicos e o coeficiente Cohen's Kappa (Cohen, 1960). A correlação entre valores genômicos é dada por:

$$\mathrm{Cor}(\hat{y}_i, \hat{y}_j) = \frac{\mathrm{Cov}(\hat{y}_i, \hat{y}_j)}{\sqrt{\mathrm{Var}(\hat{y}_i).\mathrm{Var}(\hat{y}_j)}}$$

Em que $\hat{y}_i$ e $\hat{y}_j$ são os valores genômicos dos animais para as características y_i e y_j.

O coeficiente Cohen's Kappa avalia, considerando a probabilidade de a seleção ter acontecido ao acaso, a concordância entre os indivíduos selecionados de acordo com diferentes critérios. O cálculo deste coeficiente pode ser feito da seguinte maneira:

$$\hat{k} = \frac{\mathrm{Pr(a)} - \mathrm{Pr(e)}}{1 - \mathrm{Pr(e)}},$$

Em que o numerador $\mathrm{Pr}(a) - \mathrm{Pr}(e)$ representa a proporção de observações em que a concordância ocorreu além do esperado aleatoriamente, e o denominador $1 - \mathrm{Pr}(e)$ denota a proporção de observações em que não se houve concordância, também considerando a informação ao acaso. Este coeficiente pode apresentar até 1 como valor máximo, e assim como o coeficiente de concordância simples, a concordância aumenta à medida que se aproxima de 1.

REFERÊNCIAS

BATES, D. M.; WATTS, D. G. Nonlinear regression analysis and its applications. 2 ed. New York: John Wiley and Sons, 1988.

BEAL, S.; SHEINER, L. The NONMEM system. **American Statistician**, v. 34, p. 118-119, 1980.

BOUJENANE, I.; HILAL, B. Genetic and non genetic effects for lactation curve traits in Holstein-Friesian cows. **Archiv Tierzucht**. V. 5, p. 450-457, 2012.

BRODY, S. **Bioenergetics and Growth**. New York: Rheinhold Publishing, 1945.

BRODY, S.; RAGSDALE, A. C.; TURNER, C. W. The rate of decline of milk secretion with the advance of the period of lactation. **The Journal of Animal Science**, v. 5, p. 441-444, 1923.

CANAZA-CAYO, A. W.; LOPES, P. S.; SILVA, M. V. G. B.; TORRES, R. A.; MARTINS, M. F.; ARBEX, W. A.; COBUCI, J. A. Genetic Parameters for Milk Yield Lactation Persistency Using Random Regression Models in Girolando Cattle. **Asian Australasian Journal of Animal Science**, v. 28, n. 10, p. 1407-1418, 2015.

CAPPIO-BORLINO, A.; PULINA, G.; ROSSI, G. A non-linear modification of Wood's equation fitted to lactation curves of Sardinian dairy ewes. **Small Ruminant Research**, v. 18, p. 75-79, 1995.

CARDONA, S. J. C.; ÁLVAREZ, J. D. C.; SARMENTO, J. L. R.; HERRERA, L. G. G.; CADAVID, H. C. Associação entre SNPs nos genes para k-caseína e β-lactoglobulina com curvas de lactação em cabras leiteiras. **Pesquisa Agrupecuária Brasileira**. V. 50, n. 3, p. 224-232, 2015.

COBBY, J. M.; LE DU, Y. L. P. On fitting curves to lactation data. **Animal Production**, v. 26, p. 127-133, 1978.

COBUCI, J. A.; EUCLYDES, R. F.; COSTA, C. N.; LOPES, P. S.; TORRES, R. A.; PEREIRA, C. S. Análises da Persistência na Lactação de Vacas da Raça Holandesa,

Usando Produção do Dia do Controle e Modelo de Regressão Aleatória. **Revista Brasileira de Zootecnia**, v. 33, n. 3, p. 546-554, 2004.

COBUCI, J. A.; EUCLYDES, R. F.; PEREIRA, C. S.; TORRES, R. A.; COSTA, C. N.; LOPES, P. S. Persistência na lactação – uma revisão. **Archivos latino-americanos de producción animal**, v. 11, n. 3, p. 163-173, 2003.

COBUCI, J. A.; EUCLYDES, R. F.; TEODORO, R. L.; VERNEQUE, R. S.; LOPES, P. S.; SILVA, M. A. Aspectos Genéticos e Ambientais da Curva de Lactação de Vacas da Raça Guzerá. **Revista Brasileira de Zootecnia**. v. 30, n. 4, p. 1204-1211, 2001.

COHEN, J. A coeficient of agreement for nominal scales. **Educational and Psychologial Measurement.**, v. 20, p. 37-46, 1960.

CRUZ, C. D.; SALGADO, C. C.; BHERING, L. L. **Genômica Aplicada**. Visconde do Rio Branco, MG: Ed. Suprema, 2013.

DAVIDIAN, M.; GALLANT, A. R. Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. **Journal of Pharmacokinetics and Biopharmaceutics**, v. 20, p. 529-556, 1992.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, n. 1, p. 375-385, 2009.

DHANOA, M. S. A note on an alternative form of the lactation curve model of Wood. **Animal Production**, v. 32, p. 349-351, 1981.

EL-AWADY, H. G. Genetic aspects of lactation curve traits and persistency índices in Friesian cows. **Archiva Zootechnica**. V. 16, n. 1, p. 15-29, 2013.

ESTEVES, A. M. C.; BERGMAN, J. A. G.; DURÃES, M. C.; COSTA, C. N.; SILVA, H. M. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**. v. 56, n. 4, p. 529-535, 2004.

FARHANGFAR, H.; ROWLINSON, P. Genetic Analysis of Wood's Lactation Curve for Iranian Holstein Heifers. **Journal of Biological Sciences**. V. 7, n. 1, p. 127-135, 2007.

FERREIRA, D. F. **Estatística Multivariada**. 2.Ed. Lavras: Ed. UFLA. 675p., 2011.

FERREIRA, A. G. T.; HENRIQUE, D. S.; VIEIRA, R. A. M.; MAEDA, E. M.; VALOTTO, A. A. Fitting mathematical models to lactation curves from Holstein cows in the southwestern region of the state of Parana, Brazil. **Anais da Academia Brasileira de Ciências**, v. 87, n. 1, p. 503-517, 2015.

GENGLER, N. Persistency of lactation yields: A review. Proc. Int. Workshop on Genetic Improvement of functional Traits in Cattle. **Interbull Bolletin**, v. 12, p. 97-102, 1996.

GHAVI HOSSEIN-ZADEH, N. Comparison of non-linear models to describe the lactation curves for milk yield and composition in buffaloes (*Bubalus bubalis*). **Animal**, Cambridge, v. 10, n. 2, p. 248-261, 2015.

GIANOLA, D. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. **Genetics**, v. 194, n. 3, p. 573-596, 2013.

GIANOLA, G.; DE LOS CAMPOS, G.; HILL, W. G.; MANFREDI, E.; FERNANDO, R. Additive Variability and the Bayesian Alphabet. **Genetics**, v. 183, n. 1, p. 347-363, 2009.

GONÇALVES, T. M.; OLIVEIRA, A. I. G.; FREITAS, R. T. F.; PEREIRA, I. G. Curvas de Lactação em Rebanhos da Raça Holandesa no Estado de Minas Gerais. Escolha do Modelo de Melhor ajuste. **Revista Brasileira de Zootecnia**, v, 31, n. 4, p. 1689-1694, 2002.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, v. 27, n. 4, p. 857-871, 1971.

GROENEWALD, P.C.N.; VILJOEN, C.S. A Bayesian model for analysis of lactation curves of dairy goats. **Journal of Agricultural, Biological and Environmental Statistics**, v.8, p.75-83, 2003.

HENDERSON, C. R. Aplications of linear models in animal breeding. University of Guelph, Guelph, 1984.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Produção da Pecuária Municipal**, v. 44, p. 1-51, 2016.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p. 743-756, 1990.

LEDIC, I. L.; TONHATI, H.; VERNEQUE, R. S.; EL FARO, L.; MARTINEZ, M. L.; COSTA, C. N.; PEREIRA, J. C. C.; FERNANDES, L. O.; ALBUQUERQUE, L. G. Estimativa de Parâmetros Genéticos, Fenotípicos e Ambientes para as Produções de Leite no Dia do Controle em 305 Dias de Lactação de Vacas da Raça Gir. **Revista Brasileira de Zootecnia**, v. 31, n. 5, p. 1953-1963, 2002.

LINDSTOM, M. J.; BATES, D. M. Nonlinear Mixed Effects Models for Repeated Measures Data. **Biometrics**, v. 46, n. 3, p. 673-687, 1990.

MACCIOTTA, N. P. P.; VICARIO, D.; CAPPIO-BORLINO, A. Detection of diferente shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. **Journal of Dairy Science**, v. 88, p. 1178-1191, 2005.

MAZUCHELI, J.; ARCHAR, J. A. Algumas considerações em regressão não linear. **Acta Scientiarum**, Maringá, v. 24, n. 6, p. 1791-1770, 2002.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction os total genetic value using Genome-Wide dense marker maps. **Genetics Society of America**, v. 157, p. 1819-1829, 2001.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, v. 50, n. 2, p. 159-179, 1985.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada**. 3ª reimpressão. Belo Horizonte – MG: Editora UFMG, 2005.

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

MOLENTO, C. F. M.; MONARDES, H.; RIBAS, N. P.; BLOCK, E. Curvas de lactação de vacas holandesas do Estado do Paraná, Brasil. **Ciência Rural**. v. 34, n. 5, p. 1585-1591, 2004.

MUIR, B. L.; FATEHI, J.; SCHAEFFER, L. R. Genetic Relationships Between Persistency and Reproductive Performance in First-Lactation Canadian Holsteins. **Journal of Dairy Science**. v. 87, p. 3029-3037, 2004.

NELDER, J. A. Inverse polynomials, a useful group of multi-factor response functions. **Biometrics**, v. 22, p. 128-141, 1966.

OLIVEIRA, H. T. V.; REIS, R. B.; GLÓRIA, J. R.; QUIRINO, C. R.; PEREIRA, J. C. C. Curvas de lactação de vacas F1 Holandês-Gir ajustadas pela função gama incompleta. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 59, n. 1, p. 233-238, 2007.

PAPAJSCIK, I.; BODERO, J. Modeling lactation curves of friesian cows in subtropical climate. **Animal Production**, v. 47, p. 201-207, 1988.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, v. 103, n. 482, p. 681-686, 2008.

PICCARDI, M.; MACCHIAVELLI, E.; FUNES, A. C.; BÓ, G. A.; BALZARINI, M. Fitting milk production curves through nonlinear mixed models. **Journal of Dairy Research**, v. 84, n. 2, p. 146-153, 2017.

PINHEIRO, J. C.; BATES, D. M. **Mixed-Effect Models in S and S-PLUS**. New York: Springer, 2000. 528p.

PRUDENTE, A. A. **Modelos não-lineares de regressão: alguns aspectors de teoria assintótica**. Dissertação (Mestrado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco. Recife, 2009. 108p.

RATKOWSKY, D. A. Handbook of nonlinear regression models. New York: Marcel Dekker, 1990.

REBOUÇAS, G. F.; GONÇALVES, T. M.; MARTINEZ, M. L.; AZEVEDO JUNIOR, J.; KOOPS, W. Novas funções para estimar a produção de leite, em 305 dias de lactação, de vacas da raça Gir. **Revista Brasileira de Zootecnia**, v. 37, n. 7, p. 1222-1229, 2008.

REKAYA, R.; CARABAÑO, M. J.; TORO, M. A. Bayesian Analysis of Lactation Curves in Holstein-Friesian Cattle. **Journal of Dairy Science**, v. 83, p. 2691-2701, 2000.

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa: Universidade Federal de Viçosa/Departamento de Estatística. 2012. 291p. Disponível em: <http://www.det.ufv.br/ppestbio/corpo_docente.php>. Acesso em Set. 2018.

RITZ, C.; STREIBIG. J. C. **Nonlinear Regression With R**. 3 ed. New York: Springer, 2008.

ROOK, A. J.; FRANCE, J; DHANOA, M. S. On mathematical description of lactation curves. **The Journal of Agricultural Science**, v. 121, p. 97-102, 1993.

SAGHANEZHAD, F.; ATASHI, H.; DADPASAND, M.; ZAMIRI, M. J.; SHOKRI-SANGARI, F. Estimation of Genetic Parameters for Lactation Curve Traits in Holstein Dairy Cows in Iran. **Iranian Journal of Applied Animal Science**. V. 7, n. 4, p. 559-566, 2017.

SIKKA, L. C. A. A study of lactation as affected by heredity and environment. **Journal of Dairy Research**, n. 17, p. 231-252.

SILVA, F. F.; RESENDE, M. D. V.; ROCHA, G. S.; DUARTE, D. A. S.; LOPES, P. S.; BRUSTOLINI, O. J. B.; THUS, S.; VIANA, J. M. S.; GUIMARÃES, S. E. F. Genomic growth curves of na outbred pig population. **Genetics and Molecular Biology**, v. 36, n. 4, p. 520-527, 2013.

SILVA, F. F.; ZAMBRANO, M. F. B.; VARONA, L.; GLÓRIA, L. S.; LOPES, P. S.; SILVA, M. V. G. B.; ARBEX, W.; LÁZARO, S. F.; RESENDE, M. D. V.; GUIMARÃES, S. E. F. Genome association study through nonlinear mixed models revealed new candidate for pig growth curves. **Scientia Agricola**, v. 74, n. 1, p. 1-7, 2016.

TEKERLI, M.; AKINCI, Z.; DOGAN, I.; AKCAN, A. Factor Affecting the Shape of Lactation Curves of Holstein Cows from the Balikesir Province of Turkey. **Journal of Dairy Science**. V. 83, p. 1381-1386, 2000.

TEODORO, R. L.; VERNEQUE, R. S. Orientações para o controle leiteiro. Instrução Técnica para o Produtor de Leite. Embrapa Gado de Leite, Juiz de Fora - MG. Dez, 2000.

VILELA, D.; RESENDE, J. C.; LEITE, J. B.; ALVES, E. A evolução do leite no Brasil em cinco décadas. **Revista de Política Agrícola**. V. 26, n. 1, p. 5-24, 2017.

WEISBERG, S. **Applied Linear Regression**. 3 ed. New Jersey: John Wiley & Sons, Inc, 2005.

WOOD, P. D. P. Algebraic model of the lactation curve in cattle. **Nature**, v. 216, p. 164-165, 1967.

CAPÍTULO 1

Modelos Não-Lineares Mistos na Descrição da Lactação de Bovinos da Raça Girolando

Resumo: Visando identificar a equação que melhor descreve a lactação de bovinos Girolando, o presente trabalho teve como objetivo comparar nove modelos não lineares (Brody, Cappio-Borlino, Cobby & Le Du, Dhanoa, Nelder, Papajscik e Bodero, Rook, Sikka e Wood) utilizando modelos mistos. Além disso, utilizou-se o melhor modelo para caracterização da produção leiteira do rebanho e identificação dos melhores indivíduos, por meio de análise de agrupamento. Os dados utilizados foram fornecidos pela Embrapa Gado de Leite (Juiz de Fora – MG), e consistiram em 1.822 registros de produção de leite de 226 indivíduos na primeira lactação em diferentes dias, juntamente com informações sobre a idade, agrupamento genético, grupo contemporâneo e número de ordenhas, as quais foram agrupadas e inseridas como efeito fixo. Os modelos foram comparados de acordo com os critérios de Akaike (AIC), Bayesiano (BIC), o desvio padrão residual (s), raiz do erro quadrático médio (*Root Means Square Error - RMSE*) e estatística de Durbin-Watson (DW). O modelo de Wood foi o que melhor ajustou aos dados, apresentando menores medidas de AIC e BIC, além de ausência de autocorrelação residual pela estatística DW. A partir dos efeitos individuais foi possível identificar um pequeno grupo de animais com média de produtividade acima dos demais, apresentando maior destaque dentro do rebanho.

Palavras-chave: Equação de Wood, Curvas de lactação, Análise de agrupamento.

Abstract: In order to identify the equation that describes Girolando cattle lactation better, the present work aimed to compare nine nonlinear models (Brody, Cappio-Borlino, Cobby & Le Du, Dhanoa, Nelder, Papajscik and Bodero, Rook, Sikka and Wood) using mixed models. In addition, the best model was used to characterize dairy production of the herd and identification of the best individuals, through cluster analysis. The data used were provided by Embrapa Gado de Leite (Juiz de Fora - MG), and consisted of 1,822 records of milk production of 226 individuals in the first lactation on different days, together with information on age, genetic grouping, contemporary group and number of

milks, which were grouped and inserted as fixed effect. The models were compared according to Akaike (AIC), Bayesian (BIC) criteria, residual standard deviation (s), Root Mean Square Error (RMSE) and Durbin-Watson (DW). The Wood model was the one that best fit the data, presenting smaller measures of AIC and BIC, besides the absence of residual autocorrelation by the DW statistic. From the individual effects it was possible to identify a small group of animals with an average of productivity above the others, showing greater prominence within the herd.

Keywords: Wood equation, lactation curves, cluster analysis.

1. Introdução

A curva de lactação fornece informações sobre o padrão de produção de leite durante a lactação (Jingar *et al.*, 2014). Segundo Ghavi Hossein-Zadeh (2014), a informação adquirida nas curvas de lactação é de grande importância para programas de melhoramento genético, manejo de rebanhos, monitoramento de saúde e avaliação de lucros. Em termos do melhoramento genético, observar e analisar o comportamento da curva permite a identificação e a seleção dos melhores indivíduos para novos acasalamentos, resultando no aumento da produção leiteira.

Usualmente, não se recomenda a construção de curvas de lactação por meio de modelos lineares ou polinomiais, visto que biologicamente a curva não apresenta comportamento linear ao longo do tempo. Portanto, nessa situação, recomenda-se a utilização de modelos não lineares. Nessa abordagem, além da relação não linear entre a produção de leite e o tempo de lactação ser considerada, os modelos são usualmente parcimoniosos e os parâmetros estimados possuem interpretação prática ou biológica, o que é indispensável na caracterização da produção leiteira do rebanho.

Equações propostas por diferentes autores são capazes de descrever o comportamento da lactação de várias espécies. As curvas de lactação têm sido utilizadas para caracterização da produção leiteira em caprinos (Guimarães *et al.*, 2006), búfalos (Ghavi Hossein-Zadeh, 2015) e principalmente para bovinos (Macciotta *et al.*, 2005; Jingar *et al.*, 2014; Bangar & Verma 2017).

Em geral, a estimação dos parâmetros do modelo é realizada para cada animal em cada situação (física ou ambiental) diferente, como por exemplo, Jingar *et al.* (2014) e Ghavi Hossein-Zadeh (2014) ajustaram diferentes modelos de acordo com a ordem do

parto. Oliveira *et al.* (2007) também fizeram diferentes ajustes da equação de Wood (1967) de acordo com os efeitos de ordem de parto, juntamente com a época de parição.

Para contornar a necessidade de diversos ajustes, uma alternativa que simplifica o processo de estimação é a utilização modelos não lineares mistos (MNLM). Essa metodologia permite a inserção de efeitos fixos e aleatórios no modelo, correspondendo a efeitos de diferentes fatores e dos indivíduos, respectivamente, em um único ajuste. Segundo Lindstrom e Bates (1990), modelos não lineares mistos tornaram-se populares porque sua estrutura de covariâncias flexível permite a modelagem da correlação entre as observações e/ou dados desbalanceados. Nessa abordagem, indivíduos com poucas observações geralmente não precisam ser excluídos da análise, visto que a modelagem é feita utilizando o conjunto de dados completo. Além disso, os modelos não lineares mistos permitem a modelagem conjunta dos resíduos (Silva *et al.*, 2016).

No contexto de curvas de lactação, essa técnica foi utilizada por Albertini *et al.* (2015) para avaliar a secreção de leite, energia líquida, e a exigência de proteínas de vacas de corte, concluindo que o uso dos modelos não lineares mistos é uma ótima ferramenta para descrever as variáveis em estudo; e por Piccardi *et al.* (2017), onde concluiu-se que o melhor ajuste foi obtido pelo uso dos MNLM em termos de AIC quando comparado com o método tradicional.

Em 1989, o Ministério da Agricultura, Pecuária e Abastecimento (MAPA), juntamente com as Associações representativas da raça traçaram as normas para formação do Girolando - Gado Leiteiro Tropical (5/8 Holandês + 3/8 Gir), transformando-o em prioridade nacional. Em 1996, com a oficialização da raça Girolando pelo MAPA, a Associação Brasileira dos Criadores de Girolando (ABCG) foi encarregada de executar, desenvolver e acompanhar o Registro Genealógico e as Provas Zootécnicas da raça, sendo coletados e avaliados dados de características de produção, reprodução, conformação e manejo (Silva *et al*., 2010). Segundo Canaza-Cayo *et al.* (2016), as vacas Girolando são responsáveis por aproximadamente 80% do leite produzido no Brasil.

Apesar da grande quantidade de trabalhos na literatura que fazem o uso dos modelos não-lineares mistos em diferentes situações, essa metodologia ainda não foi aplicada para caracterizar a produção leiteira da raça Girolando, o que seria de grande importância, dada a relevância e a alta produtividade de leite dessa raça no contexto nacional.

Diante do exposto, o presente trabalho tem por objetivo: i) comparar nove modelos não lineares mistos para bovinos leiteiros da raça Girolando, visando identificar

o melhor ajuste das curvas de lactação para esta raça; ii) utilizar o melhor modelo para caracterização da produtividade leiteira do rebanho e iii) identificar grupos de indivíduos superiores por meio de análise de agrupamento.

2. Material e Métodos

2.1. Banco de dados

Os dados de fenótipos utilizados no presente trabalho foram provenientes do Arquivo Zootécnico Nacional de Gado de Leite fornecidos pelo Programa de Melhoramento Genético da Raça Girolando (PMGG) sob gerenciamento da Empresa Brasileira de Pesquisa Agropecuária, Embrapa Gado de Leite em parceira com a Associação Brasileira dos Criadores de Girolando. Os dados foram compostos por 226 animais da raça Girolando, totalizando 1.822 registros de controle leiteiro referentes às mesmas. Além dos registros de produção de leite, quatro fatores referentes a cada indivíduo foram mensurados: idade (dividida em 4 grupos), o número de ordenhas (3 grupos), a composição racial (3 grupos) e o grupo contemporâneo (139 grupos).

2.2. Modelos não-lineares para curvas de lactação

Modelos não lineares são aqueles em que a função esperança é não linear em pelo menos um dos parâmetros. Isso acontece quando a relação entre as variáveis preditora e resposta acontece de maneira não linear. No contexto de bovinos leiteiros, a produção de leite segue um formato curvilíneo durante a lactação (Dongre *et al*., 2012), o que caracteriza a relação não linear entre a produção de leite e o período de lactação. Portanto, diversos modelos não lineares foram propostos para modelagem de curvas de lactação.

O primeiro modelo para ajuste de curvas de lactação foi proposto por Brody *et al*. (1923), que teve como ideia a utilização da função exponencial para descrever a etapa de declínio da lactação de vacas leiteiras após o parto. Desde então, diversos modelos vêm sendo propostos para caracterizar a produção leiteira de cabras, vacas leiteiras, búfalo, dentre outros. A curva tipo Gama proposta por Wood (1967) para descrever curvas de lactação tem sido amplamente utilizada em estudos de curvas de lactação (Ghavi Hossein-Zadeh, 2015). Essa equação é composta por três parâmetros: a_i, b_i e c_i, os quais representam a produtividade no início da lactação e as taxas de ascensão e declínio após

o pico de lactação, respectivamente. Após o surgimento da equação de Wood (1967), muitos dos modelos propostos posteriormente foram baseados em modificações dessa função. As equações não-lineares utilizadas são descritas na Tabela 1.

Tabela 1. Equações não lineares para curvas de lactação, com respectivos autores.

Autor (es)	Equação	Número de parâmetros
Brody *et al.* (1923)	$y_{ij} = a_i e^{-c_i t_{ij}}$	2
Sikka (1950)	$y_{ij} = a_i e^{(b_i t_{ij} - c_i {t_{ij}}^2)}$	3
Nelder (1966)	$y_{ij} = \frac{t_{ij}}{a_i + b_i t_{ij} + c_i {t_{ij}}^2}$	3
Wood (1967)	$y_t = a_i {t_{ij}}^{b_i} e^{-c_i t_{ij}}$	3
Cobby & Le Du (1978)	$y_{ij} = a_i - b_i t_{ij} - a_i e^{-c_i t_{ij}}$	3
Dhanoa (1981)	$y_{ij} = a_i {t_{ij}}^{b_i c_i} e^{-c_i t_{ij}}$	3
Papajcsik & Bodero (1988)	$y_{ij} = a_i t_{ij} e^{-c_i t_{ij}}$	2
Rook (1993)	$y_{ij} = a_i \left(\frac{1}{1 + \frac{b_i}{c_i + t_{ij}}} \right) e^{-d_i t_{ij}}$	4
Cappio-Borlino *et al.* (1995)	$y_{ij} = a_i {t_{ij}}^{b_i \exp(-c_i t_{ij})}$	3

Nota: y_{ij} é a produtividade de leite do i-ésimo animal no j-ésimo dia de controle, medido a partir do início da lactação; a_i, b_i, c_i e d_i são os parâmetros que definem a escala e o formato da curva de lactação, variando de acordo com as equações.

Dada a estimação de parâmetros e a determinação do comportamento das curvas de lactação, medidas importantes que caracterizam a produção leiteira podem ser estimadas. Podem ser citadas como exemplo a produção total, pico de lactação, produção no pico e persistência. A produção total (PT) é a quantidade de leite produzida (em kg) por cada animal no decorrer de toda a lactação, considerando 305 dias. Seja $f(\boldsymbol{\phi}, t)$ uma das equações descritas na Tabela 1 no tempo t a produção total pode ser estimada como abaixo (Ferreira *et al.*, 2015):

$$PT = \int_0^{305} f(\boldsymbol{\phi}, t) dt.$$

A produção no pico da curva ($Y_{máx}$) é a produção de cada indivíduo quando a curva atinge o seu ponto de máximo, conhecida também como pico de lactação. O tempo até o pico ($t_{y.máx}$) é o período que os animais levam para atingir o pico de lactação, ou seja, é o ponto do eixo do tempo correspondente ao ponto de máximo do eixo da produção ($Y_{máx}$). Já a persistência de lactação, segundo Cobuci *et al*. (2003), é a capacidade da vaca em manter sua produção de leite após atingir a produção máxima na lactação.

O tempo até o pico ($t_{Y_{t.máx}}$) é o ponto de t que maximiza $f(\boldsymbol{\phi}, t)$, ou seja, o ponto de t no qual $f'(\boldsymbol{\phi}, t) = 0$. Pode-se verificar se $t_{y.máx}$ realmente é ponto de máximo por meio da segunda derivada $f''(\boldsymbol{\phi}, t)$. Caso esta seja negativa, podemos afirmar que realmente o ponto encontrado na derivada anterior é o ponto de máximo. A produção no pico ($Y_{t.máx}$), ponto correspondente a $t_{y.máx}$ no eixo Y, pode ser obtido pela substituição do ponto de máximo $t_{y.máx}$ em $f(\boldsymbol{\phi}, t)$. Quanto à persistência (PS) da lactação, esta pode ser calculada por diferentes métodos (Cobuci *et al.*, 2003). Porém, um dos métodos utilizados é o proposto por Wood (1967), o qual se baseia na obtenção dos seus parâmetros, e pode ser calculada por: $PS = -(b_i + 1).\ln(c_i)$.

2.3. Modelos não-lineares mistos

Segundo Bates & Watts (1988) os modelos não lineares mistos podem ser considerados como uma extensão de modelos lineares mistos onde a relação entre as variáveis explicativa e resposta é não linear. Para esta abordagem, os efeitos fixos são inseridos por meio de fatores que também influenciam na variável de interesse (ordem de lactação, grupos contemporâneos, etc.), enquanto os efeitos aleatórios são estimados para representar o efeito genético de cada indivíduo, livre de outros tipos de influência.

Para Lindstrom & Bates (1990), o modelo não-linear com efeitos mistos pode ser formulado como um modelo hierárquico. Seja y_{ij} a j-ésima observação do i-ésimo indivíduo o modelo não linear misto pode ser descrito como:

$$y_{ij} = f(\boldsymbol{\phi}_i, \mathbf{x}_{ij}) + \mathbf{e}_{ij}, \quad i = 1, \dots, M; j = 1, \dots, n_i$$

em que M é o número de indivíduos, n_i é o número de observações do i-ésimo indivíduo, f é uma função não negativa e diferenciável do vetor de parâmetros $\boldsymbol{\phi}_i$ e do vetor de preditores $\mathbf{x}_{ij}$; e $\mathbf{e}_{ij}$ é o termo que representa os erros, sendo $\mathbf{e}_{ij} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}_i)$. A função f é não linear em pelo menos um dos componentes do vetor de parâmetros $\boldsymbol{\phi}_i$, o qual pode ser escrito da seguinte maneira:

$$\boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D}),$$

em que $\boldsymbol{\beta}$ é um vetor dimensional de dimensão p associado aos efeitos fixos inseridos no modelo. Esse grupo foi formado de acordo com a análise de agrupamento dos fatores descritos anteriormente: idade, número de ordenhas, composição racial e o grupo contemporâneo. Nesse agrupamento, foi utilizada a distância de Gower (1971), o método de agrupamento foi o UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) e o número de grupos foi definido de acordo com o critério de Mojena (1977); $\mathbf{b}_i$ é um vetor de dimensão q associado ao efeito aleatório do i-ésimo indivíduo; e $\mathbf{A}_i$ e $\mathbf{B}_i$ são as matrizes de incidência de dimensões $r_x p$ e $r_x q$ associadas, respectivamente, aos efeitos fixos e aleatórios; $\sigma^2\mathbf{D}$ é a matriz de covariâncias de $\mathbf{b}_i$.

Escrevendo o modelo em função do vetor resposta do i-ésimo indivíduo, como dado abaixo:

$$\mathbf{y}_i = \begin{bmatrix} y_{1n1} \\ y_{1n2} \\ \vdots \\ y_{1ni} \end{bmatrix} ; \ \mathbf{e}_i = \begin{bmatrix} e_{1n1} \\ e_{1n2} \\ \vdots \\ e_{1ni} \end{bmatrix} ; \text{ e } \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) = \begin{bmatrix} f(\boldsymbol{\phi}_i, \mathbf{x}_{i1}) \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{i2}) \\ \vdots \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{ini}) \end{bmatrix}$$

Podendo ser representado por:

$$\mathbf{y}_i = \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) + \mathbf{e}_i$$

Em que $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i)$, sendo $\boldsymbol{\Lambda}_i$ a matriz de variâncias e covariâncias das observações do i-ésimo indivíduo. O conjunto de dados completo (incluindo todos os M indivíduos) pode ser representado por:

$$\boldsymbol{y} = \boldsymbol{\eta}(\boldsymbol{\phi}) + \mathbf{e}$$

Em que

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} ; \ \boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_M \end{bmatrix} ; \text{ e } \boldsymbol{\eta}(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{\eta}_1(\boldsymbol{\phi}_1) \\ \boldsymbol{\eta}_2(\boldsymbol{\phi}_2) \\ \vdots \\ \boldsymbol{\eta}_M(\boldsymbol{\phi}_M) \end{bmatrix}$$

Assim, a distribuição do vetor de observações, dados os efeitos aleatórios ($\mathbf{y}|\mathbf{b}$) será:

$$\mathbf{y}|\mathbf{b} \sim N(\boldsymbol{\eta}(\boldsymbol{\phi}), \sigma^2\boldsymbol{\Lambda}), \ \boldsymbol{\phi} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}_i,$$

$$\mathbf{b} \sim N(\mathbf{0}, \sigma^2\widetilde{\mathbf{D}}),$$

Em que: $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_M)$, $\widetilde{\mathbf{D}} = \text{diag}(\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D})$, $\mathbf{B} = \text{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M)$,

$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_M \end{bmatrix}$ e $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}$.

Simplificando a notação, o modelo de Wood (1967) para o i-ésimo indivíduo, por exemplo, poderia ser representado da seguinte maneira:

$$y_{ij} = (\mu_a + NP_i + \xi_{ai})t_{ij}^{(\mu_b + NP_i + \xi_{bi})}e^{-(\mu_c + NP_i + \xi_{ci})t_{ij}} + e_{ij}$$

Ou seja, cada parâmetro estimado para o i-ésimo indivíduo possui influência do respectivo nível (NP_i) de produção e do seu efeito individual para cada parâmetro (ξ_{ai}, ξ_{bi} e ξ_{ci}), livre de outras influências.

Para modelos não-lineares mistos, a estimação dos parâmetros pode ser realizada por meio do algoritmo computacional de Lindstrom & Bates (1990). Esse procedimento consiste na alternância entre dois passos. No primeiro, são encontradas estimativas para os vetores $\boldsymbol{\beta}$ e $\mathbf{b}$ que minimizam a soma de quadrados aumentada, dada abaixo:

$$\sum_{i=1}^{M}[\|\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)\|^2 + \|\boldsymbol{\Delta}\mathbf{b}_i\|^2],$$

em que $\boldsymbol{\Delta} = \widetilde{\mathbf{D}}^{-1/2}$ chamado de fator de precisão relativo. Como $\mathbf{f}$ é não linear em pelo menos um dos parâmetros, utiliza-se o algoritmo de Gauss-Newton para minimização da equação (Pinheiro e Bates, 2000). O segundo passo objetiva a estimação dos componentes de variância presentes no modelo, sendo também utilizado para obter novas estimativas de $\boldsymbol{\beta}$ e $\mathbf{b}$. Obtém-se também uma estimativa atualizada para $\boldsymbol{\Delta}$, a qual é utilizada posteriormente no passo 1. Para isso, é necessária a maximização do logaritmo da função de verossimilhança baseada na distribuição marginal de $\mathbf{y}$, na w-ésima iteração, dado por:

$$l_{MV}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}|\mathbf{y}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{M}\left\{\log|\boldsymbol{\Sigma}_i(\boldsymbol{\Delta})| + \sigma^{-2}\left[\widehat{\mathbf{w}}_i^{(w)} - \widehat{\mathbf{X}}_i^{(w)}\boldsymbol{\beta}\right]'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta})\left[\widehat{\mathbf{w}}_i^{(w)} - \widehat{\mathbf{X}}_i^{(w)}\boldsymbol{\beta}\right]\right\},$$

em que $\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Delta}) = \mathbf{I} + \widehat{\mathbf{Z}}_i^{(w)}\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^{-\prime}\widehat{\mathbf{Z}}_i^{(w)'}$, $\widehat{\mathbf{w}}_i^{(w)}$ é uma aproximação de $\mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i)$ por meio de uma expansão de série de Taylor de primeira ordem e $\widehat{\mathbf{X}}_i^{(w)}$ e $\widehat{\mathbf{Z}}_i^{(w)}$ são as matrizes de derivadas associadas aos efeitos fixos e aleatórios, respectivamente. O algoritmo consiste na alternância entre os dois passos até que se obtenha convergência.

2.4. Comparação entre modelos

Os modelos foram avaliados e comparados por meio das seguintes medidas de qualidade de ajuste: critérios de informação de Akaike (AIC) e Bayesiano (BIC), desvio

padrão residual amostral (s), raiz do erro quadrático médio (*Root Means Square Error – RMSE*) e a estatística de Durbin-Watson para verificação de autocorrelação dos resíduos.

O critério de informação de Akaike (1973) mede a adequabilidade de um modelo por meio da distância de Kullback-Leibler (K-L), de acordo com a equação abaixo:

$$\mathrm{AIC} = -2\mathrm{logL}\left(\widehat{\boldsymbol{\phi}}\right) + 2\mathrm{n_{par}},$$

em que $L\left(\widehat{\boldsymbol{\phi}}\right)$ é o máximo da função de verossimilhança e $\mathrm{n_{par}}$ é a quantidade de parâmetros considerada no modelo. Já o critério de informação bayesiano (BIC), proposto por Schwarz (1978), é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O cálculo do critério BIC é definido pela medida abaixo:

$$\mathrm{BIC} = -2\mathrm{logL}\left(\widehat{\boldsymbol{\phi}}\right) + \mathrm{n_{par}}\log(\mathrm{M}),$$

em que $L\left(\widehat{\boldsymbol{\phi}}\right)$ é o máximo da função de verossimilhança e $\mathrm{n_{par}}$ é a quantidade de parâmetros do modelo. Segundo Pinheiro e Bates (2000), sob essas definições, quanto menor, melhor o modelo.

A raiz do erro quadrático médio (*Root Mean Squared Error - RMSE*) funciona como um desvio padrão generalizado (Ghavi Hossein-Zadeh, 2015) e pode ser calculado pela seguinte expressão:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})}{\mathrm{n}}},$$

em que $\mathrm{y_i}$ é a i-esima observação e $\hat{\mathrm{y}}_\mathrm{i}$ é a estimação de $\mathrm{y_i}$ pelo respectivo modelo e n é o número de observações. Como essa quantidade está associada à soma de quadrados dos erros, o melhor modelo será o que apresentar o menor RMSE.

A estatística de Durbin-Watson (Durbin, 1970), também avaliada por teste estatístico, tem por finalidade identificar a presença de autocorrelação entre os resíduos do modelo especificado. Essa estatística é dada por:

$$\mathrm{DW} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{e_i} - \mathrm{e_{i-1}})^2}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{e_i}^2}.$$

Essa medida pode variar entre 0 e 4, e podemos concluir que o modelo não possui correlação com valores próximos de 2. A hipótese nula do teste de Durbin-Watson assume que os erros são independentes, contra a hipótese alternativa de que os erros são correlacionados com uma estrutura auto-regressiva de primeira ordem (Piccardi *et al.*, 2017).

2.5. Identificação dos melhores animais

Após a identificação do melhor entre os modelos de acordo com os critérios de comparação, com base neste formou-se um novo conjunto de variáveis, composto pelos efeitos aleatórios individuais de a, b e c (somados à média geral) e as características da lactação baseada nessas medidas (produção total, tempo até o pico, persistência e produção no pico). Uma análise nova análise de agrupamento foi realizada para identificação dos melhores grupos de indivíduos, dessa vez com a utilização da distância euclidiana. Os grupos formados foram avaliados por meio de estatísticas descritivas e da construção de curvas de lactação no ponto médio de cada grupo.

2.6. Aspectos computacionais

As análises estatísticas foram realizadas no software R (R Core Team, 2018). O ajuste dos modelos não lineares mistos foi feito pelo algoritmo de Linstrom e Bates (1990), implementado no pacote *nlme* (Pinheiro *et al.*, 2017). As análises de agrupamento para formação dos grupos antes e depois dos ajustes do modelo foram realizadas com o auxílio dos pacotes *cluster* (Maechler *et al.*, 2017) e *ggdendro* (De Vries e Ripley, 2016). Medidas de qualidade de ajuste foram extraídas dos modelos utilizando o pacote *Metrics* (Hammer, 2017). A extração de algumas medidas descritivas foi feita com a utilização do pacote *dplyr* (Wickham *et al*., 2017).

3. Resultados

3.1. Agrupamento de efeitos fixos

Para estimação de efeitos fixos, os indivíduos foram separados em quatro grupos de níveis de produção por meio do agrupamento entre as variáveis descritas anteriormente. Os grupos formados, juntamente com algumas medidas descritivas são mostrados na Tabela 2. A maior entre as médias de produção dentre os quatro grupos formados foi a do quarto grupo, formado por 248 observações referentes a 30 animais.

Tabela 2. Médias, desvios padrão e amplitude dos controles registrados de acordo com os grupos formados.

Níveis de Produção	N[1]	Produção média (desvio-padrão)	Mínimo (kg)	Máximo (kg)
1 (n = 143)	1.115	14,18 (7,00)	1,80	47,20
2 (n = 43)	378	18,33 (8,87)	2,50	53,00
3 (n = 10)	81	19,11 (9,55)	3,80	48,40
4 (n = 30)	248	23,06 (9,04)	3,90	48,40

1: quantidade total de coletas em cada grupo, retiradas dos n animais de cada linha.

O maior dentre os grupos de níveis de produção foi composto por 143 animais (aproximadamente 63% da amostra) e teve a menor média de produção nos respectivos dias.

3.2. Comparação entre modelos

Os valores dos critérios AIC e BIC variaram, respectivamente, de 10.013,79 a 12.625,04 e de 10.101,92 a 12.713,16 (Tabela 3). De maneira geral, três modelos se destacaram dos demais em relação aos critérios analisados, são eles: Wood (1967), Nelder (1966) e o de Rook (1993).

O modelo de Wood foi o mais relevante no tocante aos critérios de AIC e BIC (10.013,79 e 10.101,92, respectivamente), e também apresentou baixos valores para o desvio padrão (s = 2,71) e RMSE (2,42). Em relação aos critérios do desvio-padrão (s) e a raiz do erro quadrático médio (RMSE), o modelo de maior destaque foi o de Cobby & Le Du (s = 2,70; RMSE=2,36), seguido pela equação de Wood (s = 2,71; RMSE = 2,42), mostrando boa precisão e proximidade entre os valores reais e ajustados para ambos os modelos. Outras equações, como as de Nelder (1966) e Rook (1993) também ajustaram bem os dados (Tabela 2). Os maiores valores de AIC e BIC correspondem aos modelos de Brody *et al.* (1923), Papajcsik e Bodero (1988) e Dhanoa (1981). Os demais resultados para as medidas de adequabilidade são detalhados na Tabela 3.

Tabela 3. Medidas de adequabilidade de acordo com cada modelo.

Autor (es)	AIC[1]	BIC[2]	s[3]	RMSE[4]	DW[5]
Brody *et al.* (1923)	11.390,06	11.450,64	4,66	4,40	1,05*

Continua

Conclusão

Autor (es)	AIC[1]	BIC[2]	s[3]	RMSE[4]	DW[5]
Sikka (1950)	10.887,88	10.976,00	3,40	3,06	1,61*
Nelder (1966)	10.127,15	10.215,27	2,76	2,46	1,92^{ns}
Wood (1967)	10.013,79	10.101,92	2,71	2,42	2,04^{ns}
Cobby & Le Du (1978)	10.198,85	10.286,98	2,70	2,36	2,05
Dhanoa (1981)	12.625,04	12.713,16	7,67	7,66	0,35*
Papajcsik & Bodero (1988)	11.782,34	11.842,92	5,17	4,86	1,36*
Rook (1993)	10.100,85	10.216,51	2,81	2,52	2,00*
Cappio-Borlino *et al.* (1995)	10.637,92	10.726,05	3,66	3,37	1,45*

1: Critério de Informação de Akaike; 2: Critério de Informação Bayesiano; 3: Desvio padrão residual; 4: *Root Means Square Error* (Raíz do erro quadrático médio); 5: Estatística de Durbin-Watson; *: autocorrelação significativa ao nível de 1%; ns: autocorrelação não significativa ao nível de 1%.

Quanto a presença de autocorrelação residual, o teste estatístico de Durbin-Watson apontou os ajustes de Nelder, Wood e Cobby e Le Du como os que não apresentaram correlação residual significativa ($p < 0,05$).

3.3. Caracterização da produção leiteira

Por apresentar melhor ajuste de acordo com os critérios AIC e BIC, as variáveis de lactação, estimadas pela equação de Wood foram utilizadas para estimação das médias de produção total, produção no pico, tempo até o pico de lactação e a persistência (Tabela 4). Foram excluídos três indivíduos da análise, visto que foram encontrados valores fora do espaço paramétrico das variáveis nas suas respectivas estimativas. Portanto, os resultados posteriores são referentes a 223 indivíduos.

O parâmetro *a* teve aproximadamente 8,56 como média, variando de 3,409 a 16,084. Isso indica que, em média, os indivíduos desse rebanho produziram aproximadamente 8,5 kg/dia no início da lactação, com desvio padrão correspondente a cerca de um quarto da média (2,21 kg/dia). A taxa de ascensão, representada pelo parâmetro *b*, variou de 0,1467 a 0,3776 com média de 0,2358 e uma variabilidade pequena (coeficiente de variação em torno de 14%), mostrando que o crescimento de produção após o início da lactação apresenta taxa semelhante entre os indivíduos. Já a taxa de declínio, denotada pelo parâmetro *c*, apresentou grande amplitude em relação à

média, variando de 0,0009 a 0,0080 (CV = 28,21%), porém seu desvio padrão foi baixo (0,0011).

Pode-se observar as que o rebanho produziu, em média, 4.679,1 kg de leite em até 305 dias de lactação (Tabela 4), com estimativas variando de 926,4 kg até 14.664,5 kg, e desvio padrão de 2.045,5 kg, indicando grande variabilidade de produção entre os indivíduos do rebanho. O tempo médio até o pico de lactação foi de 67,54 dias, produzindo, em média, 18,78 kg nesse ponto.

Tabela 4. Estimativas das estatísticas descritivas das variáveis de lactação após a estimação dos parâmetros *a*, *b* e *c*, além das medidas de lactação do rebanho (produção total, tempo até o pico, persistência e produção no pico).

Parâmetros	Medidas			
	Mínimo	Máximo	Média (desvio padrão)	Coeficiente de Variação (%)
a[1]	3,409	16,084	8,558 (2,214)	25,87
b[2]	0,1467	0,3776	0,2358 (0,0335)	14,21
c[3]	0,0009	0,0080	0,0039 (0,0011)	28,21
Produção total (kg)	926,4	14.664,5	4.679,1 (2045,5)	43,72
Tempo até o pico (dias)	26,18	267,76	67,54 (30,15)	44,64
Persistência	5,95	8,70	6,91 (0,46)	6,66
Produção no pico (kg/dia)	4,75	53,50	18,78 (7,50)	39,94

1: produção inicial (kg/dia); 2: taxa de ascensão; e 3: taxa de decréscimo após o pico, estimados a partir do modelo de Wood (1967).

A persistência apresentou baixa variabilidade (CV = 6,66%), indicando que após atingir o pico, o comportamento da curva dos indivíduos do rebanho não apresenta grandes diferenças entre si, condizendo também com o baixo desvio padrão de *c* (0,0011).

3.4. Identificação dos grupos

Após os parâmetros individuais (*a*, *b*, *c*) e características de lactação do rebanho (produção total, tempo até o pico, persistência e produção no pico) terem sido estimados pelo modelo de Wood, essas medidas foram submetidas à análise de agrupamento. Foram formados cinco grupos, compostos por 108, 76, 6, 25 e 8 indivíduos. As estatísticas descritivas de cada parâmetro por grupo, juntamente com as características de lactação

(produção total, tempo até o pico, persistência e produção no pico) dos mesmos são apresentadas na Tabela 5.

O primeiro grupo, composto por quase metade (aproximadamente 48%) dos indivíduos (n = 108), foi o que apresentou menores médias da produção inicial e taxa de ascensão, representados pelo parâmetro *a* e *b*, respectivamente (*a* = 7,36 e *b* = 0,2132). A menor taxa de declínio após o pico de produção, descrita pelo parâmetro *c*, foi a do grupo 3 (*c* = 0,0014), formado por 6 indivíduos. O quarto grupo foi constituído apenas por 25 animais, o qual apresentou uma alta produção inicial (*a* = 11,84), juntamente com boa taxa de ascensão (*b* = 0,2786). Porém, sua taxa de declínio após o pico foi a mais alta (*c* = 0,0048), indicando que esses indivíduos atingem o pico de produção mais rápido do que a maioria, porém essa produtividade decai de maneira mais acentuada do que os outros grupos até o fim da lactação. O quinto grupo, com oito indivíduos, apresentou a maior produção inicial entre os grupos (*a* = 12,47), atrelada a maior taxa de ascensão (*b* = 0,3059) e a segunda menor taxa de declínio (*c* = 0,0027), apresentando indícios de ser o grupo mais produtivo dentre os cinco formados.

Tabela 5. Estatísticas descritivas das variáveis de lactação estimadas pelo modelo de Wood de acordo com cada um dos 5 grupos formados pela análise de agrupamento.

Variáveis	Grupos - Média (Desvio padrão)				
	1 (n=108)	2 (n=76)	3 (n=6)	4 (n=25)	5 (n=8)
a – produção inicial (kg)	7,36 (1,62)	8,79 (1,53)	8,36 (1,51)	11,84 (1,40)	12,47 (1,81)
b – taxa de ascensão	0,2132 (0,0165)	0,2455 (0,0225)	0,2455 (0,0216)	0,2786 (0,0334)	0,3059 (0,0274)
c – taxa de declínio	0,0045 (0,0006)	0,0030 (0,0006)	0,0014 (0,0004)	0,0048 (0,0010)	0,0027 (0,0004)
Produção total (kg)	3.173,6 (806,8)	5.394,0 (1143,9)	6.513,0 (1058,3)	6.680,0 (1437,9)	10.584 (1812,2)
Tempo até o pico (dias)	48,49 (8,32)	83,58 (17,99)	180,30 (50,31)	59,93 (12,65)	111,54 (15,36)
Persistência	6,57 (0,21)	7,24 (0,25)	8,18 (0,32)	6,85 (0,27)	7,70 (0,19)
Produção no pico (kg/dia)	13,68 (3,37)	20,43 (4,36)	23,14 (3,66)	28,30 (6,01)	38,81 (6,61)

As estimativas das características de lactação indicaram o grupo 5 como o de maior produção total (10.584,0 kg/lactação), seguido pelos grupos 4, 3, 2 e 1 (que produziram, respectivamente, 6.680,0 kg; 6.513,0 kg; 5.394,0 kg e 3.173,6 kg). O grupo 1, maior dentre os que foram formados, possui uma baixa produtividade em relação aos demais (Tabela 5). As curvas de lactação das médias de cada grupo da análise de agrupamento são ilustradas na Figura 1.

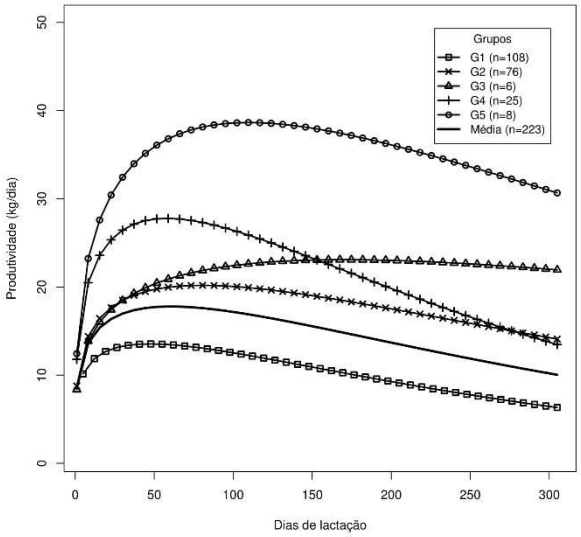


Figura 1. Curvas de lactação para cada um dos 5 grupos formados de acordo com as médias das suas características de lactação via análise de agrupamento, juntamente com a curva da média geral (linha contínua).

A curva representada pela linha contínua foi construída de acordo com a média geral do rebanho. Pode-se observar que os 108 animais integrantes do grupo 1 estão abaixo da média, e a curva do grupo 2 é ligeiramente superior. Observando-se a curva do grupo 4, podemos ver que esta atinge a produção máxima de 28,30 kg/dia em um período de tempo relativamente pequeno (aproximadamente 50 dias) quando comparado com os outros grupos. Porém, os indivíduos deste grupo têm uma baixa persistência, indicando o decaimento da produção maior que os demais, como sugerido também pelo parâmetro *c*. Podemos notar também que o grupo mais produtivo (grupo 5) possui apenas 8 indivíduos (aproximadamente 3,6% do total de animais no rebanho).

A curva de lactação média de bovinos puros da raça Holandesa é superior à média dos animais Girolando, como mostrado na Figura 2, evidenciando a superioridade da raça Holandesa em relação aos mestiços Girolando, menor produtora de leite. Porém, podemos observar que comparando essas médias com as dos indivíduos do grupo 5, essas são inferiores, mostrando que dentro do rebanho Girolando existem 8 indivíduos com as

médias superiores aos demais, superando até mesmo a média dos animais de raça Holandesa.

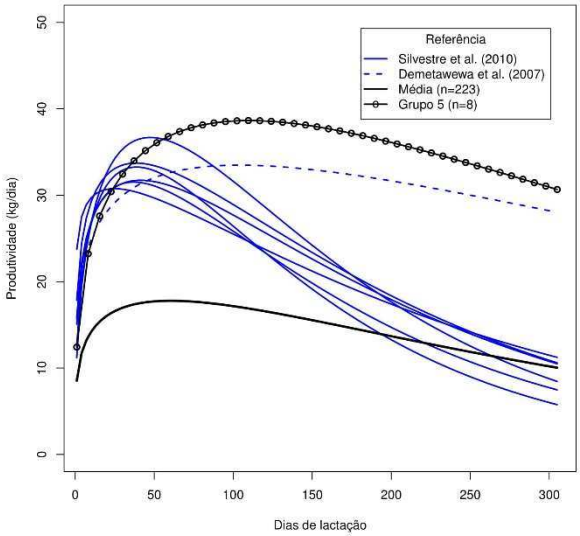


Figura 2. Curvas de lactação média do rebanho e do grupo 5 comparadas com as curvas médias de bovinos Holandeses encontradas nos estudos de Silvestre *et al*. (2010) e Demetawewa *et al*. (2007).

O mesmo ocorre quando a mesma curva média dos bovinos Girolando é comparada com animais mestiços (Figura 3). A curva média apresentou produção inferior quando comparado com outros trabalhos, porém o grupo 5 apresentou o pico de lactação mais elevado, além de uma persistência superior, visto que sua taxa de declínio após o pico é menor.

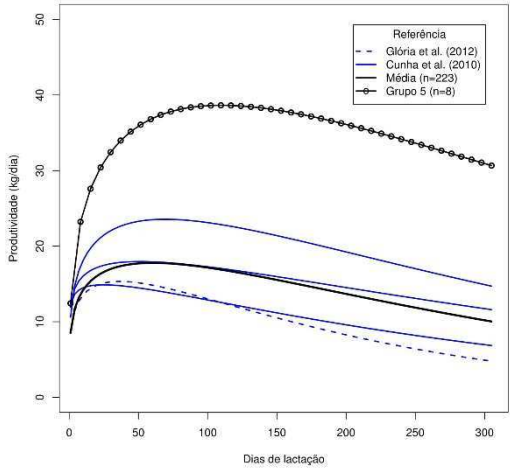


Figura 3. Curvas de lactação média do rebanho e do grupo 5 comparadas com as curvas médias de cruzamentos entre bovinos Holandeses e raças zebuínas dos estudos de Glória *et al*. (2012) e Cunha *et al*. (2010).

4. Discussão

No presente estudo, a utilização de modelos não lineares mistos permitiu, além da inserção de variáveis (4 níveis de produção descritos anteriormente) como efeitos fixos, a adição dos efeitos aleatórios (individuais) em um único ajuste. Como Lindstrom e Bates (1990) concluíram, a incorporação dos efeitos aleatórios extras dá maior flexibilidade ao modelo, eliminando a necessidade de ajustar diferentes formas funcionais para curvas individuais da mesma população. Essa característica se difere da abordagem tradicional, na qual diversos ajustes são realizados diante de diferentes situações, como nos estudos de Gupta *et al.* (2016), Ghavi Hossein-Zadeh (2015) e Bilgin *et al*. (2010), os quais ajustaram diferentes equações não lineares de acordo com a ordem de parto dos animais; Bangar e Verma (2017), consideraram a ordem da lactação. Já Gantner *et al.* (2010) dividiram os animais do rebanho em subgrupos considerando a idade, estação e produtividade, onde foi realizado um ajuste por subgrupo. Além disso, Piccardi *et al*. (2017), numa comparação entre modelos não lineares considerando apenas efeitos fixos e a abordagem de modelos mistos, concluíram que o acréscimo dos efeitos aleatórios individuais mostrou melhoramento de ajuste de acordo com os critérios de AIC e BIC. Ainda pode-se destacar que se trabalhou com dados desbalanceados. Os dados de lactação foram extraídos na maioria das vezes em diferentes dias para cada indivíduo, os quais possuíam muitas vezes a quantidade de observações diferentes um do outro.

De acordo com os critérios AIC e BIC, o modelo de Wood foi o mais eficiente para o ajuste dos dados (respectivamente, 10.013,79 e 10.101,92). Este resultado é corroborado com o estudo realizado por Bangar & Verma (2017), que também objetivou a comparação de modelos com base nas mesmas medidas, e apontou o modelo de Wood como a melhor para descrever as curvas de lactação de vacas leiteiras da raça Gir. A superioridade do ajuste no modelo de Wood também foi constatado nos estudos de Ferreira *et al*. (2015) para bovinos Holandeses criados no sudoeste do Paraná. Piccardi *et al*. (2017) também julgaram a equação de Wood como a mais eficiente para descrever a curva de lactação de vacas leiteiras na região de Santa Fé e Córdoba, na Argentina. O modelo de Wood apresentou o menor desvio padrão na comparação de modelos para ajuste da lactação de gados da raça Jersey por Cankaya *et al.* (2011), também considerado o melhor para ajustes dessa raça. Essa equação também apresentou ausência de autocorrelação residual, o que também foi constatado por Ghavi Hossein-Zadeh (2015) para dados de búfalos; Piccardi *et al.* (2017) trabalhando com vacas leiteiras de Santa Fé,

Argentina; e por Gonçalves *et al.* (2002), que ajustaram a equação de Wood para gados da raça Holandesa. A melhor adequabilidade da equação de Wood para diferentes raças e espécies justificam a grande utilização desta equação, destacada por Ghavi Hossein-Zadeh (2015) e Macciotta *et al*. (2011).

Macciotta *et al*. (2011) afirmaram que raças de alta produtividade apresentam maiores pico de lactação, que ocorre de maneira tardia, além de uma alta persistência. Dentre os grupos formados apenas o quinto se enquadra nesse perfil. Observa-se que a taxa de ascensão e o pico de lactação são superiores aos demais, juntamente com um baixo decréscimo da curva após o pico, caracterizando uma alta persistência e alta produtividade (Figura 1). Pode-se ressaltar também que o grupo 5 foi o mais produtivo, com média de mais de 10.000 kg/lactação entre seus indivíduos (Tabela 5). A média geral e a do grupo 5 foram comparadas com o que foi obtido por diferentes autores (Figuras 2 e 3), os quais estudaram raças semelhantes (Silvestre *et al.*, 2010; Dematawewa *et al.*, 2007; Glória *et al.*, 2012 e Cunha *et al.*, 2010).

Em comparação com seis ajustes médios para bovinos Holandeses realizados por Silvestre *et al*. (2010) e Demetawewa *et al.* (2007), ambos com a utilização do modelo de Wood, podemos observar que a curva média para a raça Girolando no presente estudo é inferior (Figura 2). Porém, se considerarmos apenas a curva média do grupo 5 (8 indivíduos), o pico de lactação é superior à média dos indivíduos analisados nos referidos estudos, e a persistência semelhante (Figura 2).

A Figura 3 destaca a curva média e a do grupo 5, dessa vez comparados com os resultados obtidos em outros estudos que envolveram cruzamentos de bovinos holandeses com raças zebuínas semelhantes aos bovinos Girolando realizados por Glória *et al.* (2012) e Cunha *et al.* (2010), respectivamente. A curva média dos indivíduos do presente estudo mostrou uma produtividade ligeiramente superior aos animais analisados por Glória *et al.* (2012), apresentando um pico de lactação superior (18,78 kg) e uma menor taxa de decréscimo da curva. O estudo realizado por Cunha *et al.* (2010) envolveu animais mestiços (Holandês x Zebu) e puros (Holandês), onde foram realizados ajustes de acordo com o nível de produção de leite (baixo, médio e alto). As curvas para os três níveis também estão na Figura 3. Observa-se que as curvas dos bovinos Girolando obtidas neste trabalho apresenta comportamento semelhante ao que foi constatado por Cunha *et al.* (2010) para o seu grupo de média produtividade. Isso era esperado devido não só à semelhança entre as raças estudadas, mas também a em ambos os casos serem analisados dados próximos à média dos indivíduos. O grupo de alta produtividade analisado por

Cunha *et al.* (2010) apresentou produtividade superior à média dos 223 indivíduos, porém, a curva média do grupo 5 também foi superior, apresentando maior pico de lactação e maior persistência, o que reforça a alta produtividade dos animais pertencentes a este grupo.

De maneira geral, o grupo 3 apresentou alta persistência, porém baixa taxa de ascensão e pico de lactação, diferente do grupo 4, que apresentou uma produção mais elevada no pico de lactação, porém uma baixa persistência. Considerando o que foi afirmado por Macciotta *et al.* (2011), a curva considerada ideal dentre os grupos formados é a do grupo 5, pois os animais que compõe esse grupo foram capazes de unir uma alta taxa de ascensão, alta produtividade no pico e alta persistência, resultando em uma produção total média superior tanto em relação aos demais grupos, quanto em relação à média de raças semelhantes constatado por outros autores.

5. Conclusões

A utilização de modelos não lineares mistos teve um resultado satisfatório, possibilitando a identificação da melhor equação segundo os critérios avaliados para ajustar dados de lactação de bovinos da raça Girolando. Essa abordagem também a caracterização leiteira das vacas dessa raça através da estimação de efeitos genéticos individuais, permitindo o conhecimento das variáveis livres de fatores externos que afetam a lactação. Isso também propiciou a identificação de um grupo seleto de animais (os 8 animais do grupo 5) que, além de uma maior produção total, possuem maior persistência que a maioria dos demais, podendo estes serem considerados os melhores indivíduos, independentes do efeito de outras variáveis.

6. Referências

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716-723, 1974.

ALBERTINI, T. Z.; MEDEIROS, S. R.; TORRES JÚNIOR, R. A. A.; ZOCCHI, S. S.; OLTJEN, J. W.; STRATHE, A. B.; LANNA, D. P. D. A methodological approach to estimate the lactation curve and net energy and protein requirements of beef cows using

nonlinear mixed-effect modeling. **Journal of Animal Science**, v. 90, n. 11, p. 3867-3878, 2015.

BANGAR, Y. C.; VERMA, M. R. Non-linear modelling to describe lactation curve in Gir crossbred cows. **Journal of Animal Science and Technology**, v. 59, n. 3, p. 1-7, 2017.

BATES, D. M.; WATTS, D. G. **Nonlinear Regression Analysis and Its Applications**. 2 ed. New York: Wiley, 1988. 90p.

BILGIN, O. C.; ESENBUGA, N.; DAVIS, M. E. Comparison of models for describing the lactation curve of Awassi, Morkaraman and Tushin sheep. **Archiv Tierzucht**, v. 53, p. 447-456, 2010.

BRODY, S.; RAGSDALE, A. C.; TURNER, C. W. The rate of decline of milk secretion with the advance of the period of lactation. **The Journal of Animal Science**, v. 5, p. 441-444, 1923.

CANKAYA, S.; UNALAN, A.; SOYDAN, E. Selection of a mathematical model to describe the lactation curves of Jersey cattle. **Archiv Tierzucht**, v. 54, n. 1, p. 27-35, 2011.

CAPPIO-BORLINO, A.; PULINA, G.; ROSSI, G. A non-linear modification of Wood's equation fitted to lactation curves of Sardinian dairy ewes. **Small Ruminant Research**, v. 18, p. 75-79, 1995

COBBY, J. M.; LE DU, Y. L. P. On fitting curves to lactation data. **Animal Production**, v. 26, p. 127-133, 1978.

COBUCI, J. A.; EUCLYDES, R. F.; PEREIRA, C. S.; TORRES, R. A.; COSTA, C. N.; LOPES, P. S. Persistência na lactação – uma revisão. **Archivos Latinoamericanos de Producción Animal**, v. 11, n. 3, p. 163-173, 2003.

CUNHA, D. N. F. V.; PEREIRA, J. C.; SILVA, F. F.; CAMPOS, O. F.; BRAGA, J. L.; MARTUSCELLO, J. A. Selection of models of lactation curves to use in milk production simulation systems. **Revista Brasileira de Zootecnia**, v. 39, p. 891-902, 2010.

DE VRIES, A.; RIPLEY, B. D. ggdendro: Create Dendograms and Tree Diagrams Using 'ggplot2'. R package version 0.1-20. URL: https://CRAN.R-project.org/package=ggdendro.

DHANOA, M. S. A note on an alternative form of the lactation curve model of Wood. **Animal Production**, v. 32, p. 349-351, 1981.

DEMATAWEWA, C. M. B.; PEARSON, R. E.; VANRADENT, P. M. Modeling extended lactations of Holsteins. **Journal of Dairy Science**, v. 90, p. 3924-3936, 2007.

DONGRE, V.; GHANDI, R. S.; SINGH, A. Comparison of different lactation curve models in Sahiwal cows. **Turkish Journal of Veterinary and Animal Sciences**, v. 36, n. 6, p. 723-726.]

FERREIRA, A. G. T.; HENRIQUE, D. S.; VIEIRA, R. A. M.; MAEDA, E. M.; VALOTTO, A. A. Fitting mathematical models to lactation curves from Holstein cows in the southwestern region of the state of Parana, Brazil. **Anais da Academia Brasileira de Ciências**, v. 87, n. 1, p. 503-517, 2015.

GANTNER, V.; JOVANOVAC, S.; RAGUZ, N.; SOLIC, D.; KUTEROVAC, K. Nonlinear Vs. linear regression models in lactation curve prediction. **Bulgarian Journal of Agricultural Science**, v. 16, n. 6, p. 794-800, 2010.

GHAVI HOSSEIN-ZADEH, N. Comparison of non-linear models to describe the lactation curves for milk yield and composition in buffaloes (*Bubalus bubalis*). **Animal**, Cambridge, v. 10, n. 2, p. 248-261, 2015.

GHAVI HOSSEIN-ZADEH, N. Comparison of non-linear models to describe the lactation curves of milk yield and composition in Iranian Holsteins. **The Journal of Agricultural Science**, v. 152, p. 309-324, 2014.

GONÇALVES, T. M.; OLIVEIRA, A. I. G.; FREITAS, R. T. F.; PEREIRA, I. G. Curvas de lactação em rebanhos da raça Holandesa no estado de Minas Gerais. Escolha do modelo de melhor ajuste. **Revista Brasileira de Zootecnia**, v. 31, n. 4, p. 1689-1694, 2002.

GUIMARÃES, V. P.; RODRIGUES, M. T.; SARMENTO, J. L. R.; ROCHA, D. T. Utilização de funções matemáticas no estudo da curva de lactação em caprinos. **Revista Brasileira de Zootecnia**, v. 35, n. 2, 2006.

GUPTA, A.; GHANDI, R. S.; SINGH, M.; SINGH, A.; PRAKASH, V.; DASH, S. K.; DASH, S. Comparison of different lactation curve models in Sahiwal cattle up to fourth parity using monthly test day milk yields. **Indian Journal of Dairy Science**, v. 69, n. 4, p. 460-466, 2016.

HAMMER, B. Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.2. URL: https://CRAN.R-project.org/package=Metrics.

JINGAR, S.; MEHLA, R. K.; SINGH, M.; ROY, A. K. Lactation curve pattern and prediction of milk production performance in crossbred cows. **Journal of Veterinary Medicine**, 2014. doi:10.1155/2014/814768.

LINDSTOM, M. J.; BATES, D. M. Nonlinear Mixed Effects Models for Repeated Measures Data. **Biometrics**, v. 46, n. 3, p. 673-687, 1990.

MACCIOTTA, N. P. P.; DIMAURO, C.; RASSU, S. P. G.; STERI, R.; PULINA, G. The mathematical description of lactation curves in dairy cattle. **Italian Journal of Animal Science**, v. 10, n. 51, p. 213-223, 2011.

MACCIOTTA, N. P. P.; VICARIO, D.; CAPPIO-BORLINO, A. Detection of diferente shapes of lactation curve for milk yield in dairy cattle by empirical mathematical models. **Journal of Dairy Science**, v. 88, p. 1178-1191, 2005.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.

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

NELDER, J. A. Inverse polynomials, a useful group of multi-factor response functions. **Biometrics**, v. 22, p. 128-141, 1966.

OLIVEIRA, H. T. V.; REIS, R. B.; GLÓRIA, J. R.; QUIRINO, C. R.; PEREIRA, J. C. C. Curvas de lactação de vacas F1 Holandês-Gir ajustadas pela função gama incompleta. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, Belo Horizonte-MG, v. 59, n. 1, p. 233-238, 2007.

PAPAJSCIK, I.; BODERO, J. Modeling lactation curves of friesian cows in subtropical climate. **Animal Production**, v. 47, p. 201-207, 1988.

PICCARDI, M.; MACCHIAVELLI, E.; FUNES, A. C.; BÓ, G. A.; BALZARINI, M. Fitting milk production curves through nonlinear mixed models. **Journal of Dairy Research**, v. 84, n. 2, p. 146-153, 2017.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D. and R Core Team (2017). nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-131, URL: https://CRAN.R-project.org/package=nlme.

PINHEIRO, J. C.; BATES, D. M. **Mixed-Effect Models in S and S-PLUS**. New York: Springer, 2000. 528p.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

ROOK, A. J.; FRANCE, J; DHANOA, M. S. On mathematical description of lactation curves. **The Journal of Agricultural Science**, v. 121, p. 97-102, 1993.

SCHWARZ, G. Estimating the dimension of a model. **Annals of Statistics**, v. 6, p. 461-464, 1978.

SIKKA, L. C. A. A study of lactation as affected by heredity and environment. **Journal of Dairy Research**, n. 17, p. 231-252.

SILVESTRE, A. M. D.; ALMEIDA, J. C. M.; SANTOS, V. A. C.; FONTES, P. J. P.; ALVES, V. C. Modelling lactation curves of "Barrosã" beef cattle with Wood's model. **Italian Journal of Animal Science**, v. 9, p. 243-247, 2010.

SILVA, F. F.; ZAMBRANO, M. F. B.; VARONA, L.; GLÓRIA, L. S.; LOPES, P. S.; SILVA, M. V. G. B.; ARBEX, W.; LÁZARO, S. F.; RESENDE, M. D. V.;

GUIMARÃES, S. E. F. Genome association study through nonlinear mixed models revealed new candidate for pig growth curves. **Scientia Agricola**, v. 74, n. 1, p. 1-7, 2016.

WOOD, P. D. P. Algebraic model of the lactation curve in cattle. **Nature**, v. 216, p. 164-165, 1967.

CAPÍTULO 2

Curvas de lactação genômicas de bovinos Girolando baseadas em modelos não-lineares mistos

Resumo: Objetivando a caracterização do comportamento genômico das características de lactação de bovinos da raça Girolando, o presente trabalho teve como proposta a construção de curvas de lactação genômicas por meio da Seleção Genômica Ampla. Objetivou-se também a descrição do valor genético das variáveis de lactação. Tais variáveis foram estimadas por meio de modelos não-lineares mistos com a utilização da equação de Wood. Os dados foram fornecidos pela Embrapa Gado de Leite (Juíz de Fora-MG), referentes a 1.822 registros de controle leiteiro correspondente a 226 bovinos Girolando, juntamente com a informação de 37.673 marcadores SNPs associados aos animais em estudo. As curvas de lactação genômicas mostraram maiores diferenças genéticas de produtividade quando observadas segundo às produções total e inicial, taxa de ascensão e o pico de lactação. As herdabilidades estimadas foram de 0,10, 0,10 e 0,09 para os parâmetros do modelo de Wood (*a*, *b*, e *c*, respectivamente) e de 0,27, 0,12, 0,29 e 0,22 para a produção total, pico de lactação, persistência e tempo até o pico. As correlações entre os valores genômicos apresentaram resultados entre -0,85 e 0,98. A concordância entre os animais superiores de acordo com diferentes variáveis foi maior quando a correlação genômica entre as mesmas se apresentou também elevada. A partir dessa abordagem, foi possível o conhecimento genético das curvas de lactação da raça Girolando, bem como o conhecimento entre as divergências individuais entre os animais e associação genômica entre as suas variáveis de lactação.

Palavras-chave: Herdabilidade, BLASSO, Seleção genômica ampla.

Abstract: Aiming to characterize the genomic behavior of the lactation traits of Girolando cattle, the present work had as proposal the construction of genomic lactation curves through the Genome Wide Selection. The objective was also to describe the genetic value of lactation variables. These variables were estimated by using nonlinear mixed models using the Wood equation. The data were provided by Embrapa Gado de Leite (Juíz de Fora-MG), referring to 1,822 records of dairy control corresponding to 226 Girolando cattle, together with the information of 37,673 markers SNPs associated to the

animals under study. The genomic lactation curves showed higher genetic differences in productivity when observed according to total and initial yield, ascending rate and peak lactation. The estimated heritabilities were 0.10, 0.10 and 0.09 for the parameters of the Wood model (a, b, and c, respectively) and of 0.27, 0.12, 0.29 and 0.22 for the total production, peak lactation, persistence and time to peak. Correlations between the genomic values presented results between -0.85 and 0.98. The concordance between the superior animals according to different variables was higher when the genomic correlation between them was also high. From this approach, it was possible the genetic knowledge of Girolando lactation curves, as well as the knowledge between the individual divergences between the animals and genomic association among their lactation variables.

Keywords: Heritability, BLASSO, genome-wide selection.

1. Introdução

Estudos que visam a construção de curvas de lactação fazem o uso de modelos não lineares visto que tais modelos são flexíveis e seus parâmetros possuem interpretação biológica (Barroso *et al*., 2016). Além disso, os modelos não-lineares são parcimoniosos, geralmente considerando uma equação que contém uma quantidade pequena e pré-fixada de parâmetros.

Segundo Macciotta *et al*. (2011) o principal interesse nesse tipo de modelagem não se trata da identificação do comportamento do fenômeno, mas no ajuste de desvios individuais de uma curva média. Portanto é conveniente a utilização dos modelos não lineares mistos, já que essa metodologia permite a inclusão de efeitos fixos (fatores ambientais que afetam a lactação) e aleatórios (efeito individual, representado por variações em torno da sua média) em um único ajuste, permitindo também a modelagem conjunta dos resíduos. Uma vez estimados os efeitos aleatórios individuais de cada animal, livres da influência de fatores ambientais, podemos ter acesso às suas características de lactação e assim estudar suas relações e identificar indivíduos superiores levando em conta essas variáveis.

Como as diferenças entre as medidas que compõe as curvas de lactação de gados de leite usualmente ocorrem devido a influências genéticas, metodologias que visam a associação entre variáveis fenotípicas e informações do DNA para identificar essas

diferenças vêm sendo utilizadas (Macciotta *et al*., 2015; Cardona *et al*., 2015), baseadas em marcadores moleculares. Dentre elas, podemos destacar a Seleção Genômica Ampla – SGA (Meuwissen *et al*., 2001), que visa a predição do mérito genético pela construção de modelos estatísticos utilizando informações diretamente do genoma por meio de marcadores SNPs.

Silva *et al*. (2013) aplicaram a Seleção Genômica por meio dos métodos RR-BLUP e Lasso Bayesiano (BLASSO) para construção de curvas genômicas de crescimento de suínos, baseadas em marcadores SNPs. Sua utilização permitiu a obtenção dos parâmetros genéticos das curvas, possibilitando a identificação de variações genéticas destes parâmetros, em uma abordagem definida como procedimento em dois passos (Varona *et al*., 1999; Pong-Wong e Hadjipavlou, 2010). Este princípio pode também ser aplicado para construção de curvas de lactação genômicas, visto que os os valores genéticos das variáveis de lactação por indivíduo podem ser estimados com base nas informações de marcadores SNPs, possibilitando o conhecimento do comportamento genômico da curva e dos demais parâmetros de lactação.

Além disso, espera-se que a construção de curvas de lactação genômicas baseada em informações obtidas por modelos não-lineares mistos apresente maior capacidade de diferenciação entre animais superiores, uma vez que os efeitos individuais são estimados sem a influência dos demais fatores que afetam a produtividade na lactação, estes por sua vez estimados como efeitos fixos.

Diante do exposto, o presente trabalho tem como objetivo i) propor a utilização da Seleção Genômica Ampla para construção das curvas de lactação genômicas de bovinos da raça Girolando usando como variáveis os parâmetros estimados por meio de modelos não lineares mistos; ii) estimar os parâmetros genéticos importantes para caracterização do rebanho; iii) identificar a herdabilidade e as correlações entre valores genômicos das características de lactação e iv) identificar os melhores animais utilizando cada característica genética de lactação como critério.

2. Material e métodos

2.1. Banco de dados

Os dados foram fornecidos pela Embrapa (Centro Nacional de Pesquisa em Gado de Leite), Juíz de Fora, Minas Gerais. Inicialmente, as informações disponíveis eram

referentes a 94.263 medidas de controle leiteiro referentes a 11.459 animais da raça Girolando. Como desejou-se trabalhar apenas com as informações em que estavam disponíveis dados moleculares, através de marcadores SNPs, utilizou-se os dados referentes aos 226 animais genotipados, compostos por 1.822 registros de controle leiteiro vacas Girolando. Os dados genotípicos foram provenientes de *Illumina BovineSNP50 BeadChip*, contendo informações de 37.673 marcadores SNPs associados aos 226 animais considerados no estudo.

2.2. Modelo não-lineares mistos para curvas de lactação

Modelos de regressão são utilizados para descrever a relação entre duas variáveis (explicativa e resposta). No contexto de curvas de lactação, a variável explicativa é o tempo de lactação e a variável resposta é a produtividade, a qual varia de acordo com cada dia. Seja y_{ij} a i-ésima observação da variável resposta, representada pela produção (kg/dia) no tempo j, a função geral que descreve a relação entre a produtividade e o período de lactação é dada por:

$$\mathbf{y}_{ij} = f(\boldsymbol{\phi}_i; t_{ij}) + \boldsymbol{\varepsilon}_{ij}$$

Em que $f(\boldsymbol{\phi}_i; t_{ij})$ é uma função não linear pré-especificada que relaciona o tempo à produção de leite; $\boldsymbol{\phi}_i$ é o vetor de parâmetros, que varia de acordo com a função especificada; e $\boldsymbol{\varepsilon}_{ij}$ é o erro associado à observação $\mathbf{y}_{ij}$. Dizemos que $f(\boldsymbol{\phi}_i; t_{ij})$ é não linear se esta for não linear em pelo menos um dos parâmetros que compõe $\boldsymbol{\phi}_i$.

Diferentes equações foram propostas na literatura para descrever curvas de lactação (Wood *et al*., 1967; Nelder, 1966; Dijkstra *et al*., 1997), bem como alguns estudos já foram propostos com o intuito de comparar diferentes equações por meio de medidas de adequabilidade (Ghavi Hossein-Zadeh, 2015; Bangar e Verma, 2017). Porém, segundo Piccardi *et al.* (2017), o modelo de Wood (1967) têm sido o mais utilizado por apresentar melhor qualidade de ajuste. Com base em análises preliminares (Capítulo 1), escolheu-se a equação de Wood (1967) para ajustar as curvas de lactação do presente estudo. Considerando a observação da produtividade (y) para o i-esimo indivíduo no tempo j (em dias), essa equação pode ser escrita da seguinte forma:

$$y_{ij} = a_i t_{ij}^{b_i} e^{-c_i t_{ij}},$$

em que a representa a produção inicial (kg/dia), b é a taxa de ascensão até o pico e c denota a taxa de decréscimo após o pico. Uma vez estabelecida a função que será utilizada

para ajuste das curvas, podemos extrair informações importantes a respeito da lactação. São elas o tempo até o pico de lactação: tempo no eixo X que leva até o animal atingir o ponto máximo (tp), o pico de produção: ponto máximo da curva de produtividade no eixo Y ($Y_{máx}$); a produção total: quantidade absoluta que cada indivíduo produziu no período de lactação (PT) e a persistência: capacidade de cada indivíduo de manter a produtividade após o pico (PS). Segundo Wood (1967), considerando o i-ésimo indivíduo no tempo t, essas medidas são obtidas por:

$$tp_i = \frac{b_i}{c_i},$$

$$Y_{máx.i} = a_i(b_i/c_i)^{b_i}e^{-b_i} \ ,$$

$$PT_i = \int_0^{305} a_i.t^{b_i}e^{-c_i.t}dt \ ,$$

$$PS_i = -(b_i + 1).\ln(c_i) \ ,$$

Em que a, b e c são os parâmetros da equação de Wood (1967), estimados por indivíduo.

No contexto de modelos não lineares mistos, cada um dos parâmetros do modelo de Wood (1967) é decomposto, sendo cada um deles atribuído a efeitos fixos e aleatórios. No presente estudo, quatro níveis de produção foram considerados como efeitos fixos. Esses níveis foram obtidos por meio de análise de agrupamento das variáveis: número de ordenhas (3 grupos), idade (4 grupos), agrupamento genético (3 grupos) e grupos contemporâneos (139 grupos). Nessa análise de agrupamento, a matriz de distância foi calculada com base no algoritmo de Gower (1971). O método utilizado para o agrupamento de variáveis foi o UPGMA e a quantidade de grupos foi estabelecida segundo o critério de Mojena (1977).

Sob essa abordagem, o modelo anteriormente pode ser escrito, de maneira simplificada, da seguinte forma:

$$y_{ij} = (\mu_a + NP_i + \xi^2_{ai}).t_{ij}^{(\mu_b + NP_i + \xi^2_{bi})}e^{-(\mu_c + NP_i + \xi^2_{ci}).t_{ij}} + \varepsilon_{ij}.$$

Esse modelo assume que $a_i = \mu_a + NP_i + \xi^2_{ai}$, $b_i = \mu_b + NP_i + \xi^2_{bi}$ e $c_i = \mu_c + NP_i + \xi^2_{ci}$, sendo μ_a, μ_b e μ_c as médias dos efeitos fixos de cada parâmetro; NP_i é o efeito do nível de produção do i-ésimo indivíduo; e ξ^2_{ai}, ξ^2_{bi} e ξ^2_{ci} são os efeitos aleatórios individuais para cada parâmetro do modelo; e ε_{ij} é o efeito residual associado à observação y_{ij}.

A estimação de parâmetros foi feita por meio do algoritmo de Lindstrom e Bates (1990), baseado em dois passos. Segundo Pinheiro e Bates (1995), este algoritmo consiste na alternância entre dois passos: 1) *PNLS (Penalized Nonlinear Least Squares) Step* e 2) *LME (Linear Mixed Effects) Step*. O primeiro passo consiste na minimização da soma de

quadrados penalizada com base obtendo-se a estimativa dos efeitos fixos e aleatórios. Seja $f(\boldsymbol{\phi}_i; t_{ij}) = f(\boldsymbol{\beta}; \mathbf{b}_i)$ a função esperança do i-ésimo indivíduo e seja $\boldsymbol{\beta}$ e $\mathbf{b}_i$ os vetores que denotam os efeitos fixos e aleatórios, o primeiro passo consiste na minimização da função abaixo:

$$\sum_{i=1}^{M} [||\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}; \mathbf{b}_i)||^2 + ||\boldsymbol{\Delta}\mathbf{b}_i||^2] ,$$

Já no segundo, atualiza-se a estimativa de $\boldsymbol{\Delta}$ por meio do estimador de máxima verossimilhança baseado na distribuição marginal de $\mathbf{y}$. Para a estimação de $\boldsymbol{\Delta}$, utiliza-se uma expansão de Séries de Taylor de primeira ordem em torno das estimativas atuais de $\boldsymbol{\beta}$ e $\mathbf{b}_i$, chamadas de $\widehat{\boldsymbol{\beta}}^{(w)}$ e $\mathbf{b}^{(w)}$. Sejam:

$$\widehat{\mathbf{Z}}_i = \left.\frac{\partial \mathbf{f}_i}{\partial \mathbf{b}_i'}\right|_{\widehat{\boldsymbol{\beta}}, \mathbf{b}^{(w)}}$$

$$\widehat{\mathbf{X}}_i = \left.\frac{\partial \mathbf{f}_i}{\partial \boldsymbol{\beta}'}\right|_{\widehat{\boldsymbol{\beta}}, \mathbf{b}^{(w)}}$$

A aproximação é feita pelo resíduo $\widehat{\mathbf{w}}_i^{(w)}$, dado por:

$$\widehat{\mathbf{w}}_i^{(w)} = \mathbf{y}_i - \mathbf{f}_i\left(\widehat{\boldsymbol{\beta}}^{(w)}, \hat{\mathbf{b}}_i^{(w)}\right) + \widehat{\mathbf{X}}_i^{(w)}\widehat{\boldsymbol{\beta}}^{(w)} + \widehat{\mathbf{Z}}_i^{(w)}\hat{\mathbf{b}}_i^{(w)}$$

Com a obtenção de $\widehat{\mathbf{w}}_i^{(w)}$ (resíduo no passo w), podemos usar a aproximação do log da função de verossimilhança para obtenção de $\boldsymbol{\Delta}$, dada abaixo:

$$\begin{aligned} & l_{MV}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Delta}|\mathbf{y}) \\ & = -\frac{N}{2}\log(2\pi\sigma^2) \\ & -\frac{1}{2}\sum_{i=1}^{M}\left\{\log\left|\sigma^2\left(\mathbf{I} + \widehat{\mathbf{Z}}_i^{(w)}\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}'\widehat{\mathbf{Z}}_i^{(w)'}\right)\right|\right. \\ & \left. + \sigma^{-2}\left[\widehat{\mathbf{w}}_i^{(w)} - \widehat{\mathbf{X}}_i^{(w)}\boldsymbol{\beta}\right]'\left(\mathbf{I} + \widehat{\mathbf{Z}}_i^{(w)}\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}'\widehat{\mathbf{Z}}_i^{(w)'}\right)^{-1}\left[\widehat{\mathbf{w}}_i^{(w)} - \widehat{\mathbf{X}}_i^{(w)}\boldsymbol{\beta}\right]\right\}. \end{aligned}$$

Além disso, vale ressaltar que neste passo também são fornecidos também os componentes de variância e estimativas para $\boldsymbol{\beta}$ e $\mathbf{b}_i$. Os passos se alternam até que se atinja o critério de convergência estabelecido. Para o presente estudo, vale lembrar que foram estimados os efeitos de níveis de produção como fixos ($\boldsymbol{\beta}$) e efeitos individuais como aleatórios ($\mathbf{b}_i$).

2.3. Seleção Genômica Ampla (SGA)

A estimação os parâmetros e funções paramétricas das curvas de lactação permite apenas o conhecimento das curvas de lactação fenotípicas. Para que seja possível a construção das curvas genômicas, devemos estimar os valores genômicos dos animais para essas características. Uma das maneiras para estimação dos méritos genéticos das variáveis de lactação é via Seleção Genômica Ampla (SGA). Essa metodologia consiste no uso de marcadores moleculares SNPs, amplamente distribuídos em todo o genoma, para predição do mérito genético dos animais para as características de interesse.

Portanto, as variáveis representadas pelos parâmetros da curva de lactação (*a*, *b* e *c*), juntamente com as características de lactação obtidas por meio destes (produção total, pico de lactação, persistência e tempo até o pico) foram submetidas à análise de SGA. Com o conhecimento da estimativa do mérito genético dos parâmetros da curva de lactação, é possível a construção das curvas genômicas.

Diferentes metodologias de Seleção Genômica Ampla têm sido propostas para estimação de efeitos de marcadores, e muitas delas envolvem a utilização de inferência Bayesiana. Meuwissen *et al*. (2001) propôs a utilização dos métodos Bayes A e Bayes B, os quais foram comparados com o BLUP (*Best Linear Unbiased Predictor*), este proposto por Henderson (1974), e com o método de mínimos quadrados utilizado por Lande e Thompson (1990), obtendo melhores resultados em termos de acurácia. Fan *et al*. (2011) e Teixeira *et al.* (2016) também obtiveram bons resultados com a utilização de métodos bayesianos na SGA.

Considerando as variáveis estimadas no passo anterior (produção inicial – parâmetro *a*, taxa de ascensão – parâmetro *b*, taxa de declínio – parâmetro *c*, e produção total), o modelo geral da seleção genômica considerando 37.673 marcadores SNP's presentes neste estudo, considerando o i-ésimo indivíduo ($i = 1, \dots, 223$) é dado por (Meuwissen *et al*., 2001):

$$y_i = \mu + \sum_{k=1}^{37.673} x_{ik}\beta_k + e_i,$$

Em que y_i é a i-ésima observação da variável de interesse Y, μ é a média geral; x_{ik} é o valor absoluto correspondente ao k-ésimo marcador do i-ésimo indivíduo; e e_i é o erro associado a y_i. Na forma matricial, o modelo pode ser escrito da seguinte forma:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

Sendo $\mathbf{y}$ o vetor (223 x 1) que contém as observações da variável Y; $\boldsymbol{\mu}$ é o vetor de médias para todos os indivíduos (223 x 1); $\mathbf{X}$ é a matriz de incidência dos marcadores de cada indivíduo (223 x 37.673); $\boldsymbol{\beta}$ é o vetor dos efeitos dos marcadores (37.673 x 1); e $\mathbf{e}$ é o vetor de erros (223 x 1).

As metodologias bayesianas se diferem no que diz respeito às pressuposições feitas a respeito da distribuição de probabilidade assumida pelos marcadores, ou seja, sobre suas distribuições à priori. No presente trabalho foi utilizado o método BLASSO (*Bayesian LASSO*) (DE LOS CAMPOS, 2009) para estimação de efeitos de marcadores e predição das características genômicas. Para a obtenção da distribuição combinada da distribuição a *priori* e da verossimilhança dos dados, ou seja, para a obtenção da distribuição a *posteriori* dos efeitos genéticos dos marcadores, adota-se o procedimento de simulação estocástica (método de Monte Carlo via Cadeias de Markov – MCMC) denominado amostragem de Gibbs.

O método BLASSO, diferentemente de algumas outras metodologias de seleção genômica, conta com o parâmetro de suavização (λ), o qual tem como objetivo aproximar os efeitos de marcadores de zero e pode ser estimado por métodos MCMC. Esse parâmetro de suavização é responsável por controlar a distribuição das variâncias dos marcadores. Essa metodologia também inclui um termo de variância comum para modelar ambos os termos, os resíduos e os efeitos genéticos dos marcadores (Park; Casella, 2008; De Los Campos *et al*., 2009). A utilização de uma distribuição exponencial dupla para o efeito dos marcadores concentra a maior parte da frequência (densidade) em valores próximos de zero, o que usualmente ocorre na distribuição dos efeitos dos marcadores. Segundo Silva *et al*. (2013), o BLASSO é um método de regressão Bayesiana penalizada em que o estimador dos efeitos de marcadores é dado por:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\{ (\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'(\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} |\beta_k| \right\},$$

em que λ é o parâmetro de regularização. Quando $\lambda = 0$, não existe regularização e quando $\lambda > 0$ existe o encolhimento (*shrinkage*) dos efeitos de marcadores em torno de zero, com a possibilidade de alguns serem identicamente iguais a zero, resultando em um procedimento simultâneo de estimação e seleção de variáveis.

Neste procedimento, ainda segundo Silva *et al*. (2013) a distribuição conjunta dos efeitos e marcadores $(\beta_1, \beta_2, \ldots, \beta_k)$ é $\prod_{k=1}^{p} N(0, \sigma^2_{\beta_k})$, onde $\sigma^2_{\beta_k} = \sigma^2\tau^2_k$, sendo σ^2 a variância residual, com distribuição Qui-quadrado inversa (χ^{-2}) e τ^2_k é o parâmetro de

escala relacionado a cada marcador. Essa metodologia também assume que a distribuição desses parâmetros de escala $(\tau_1^2, \tau_2^2, \ldots, \tau_k^2)$ é um produto de distribuições exponenciais, $\prod_{k=1}^{p} \exp(\lambda)$, e que a distribuição à priori de λ é uma Gama(v_1, v_2). A variância genética aditiva para o cálculo da herdabilidade, mostrada posteriormente, é dada por $\sigma_a^2 = 2\sum_{k=1}^{p} \sigma_{\beta_k}^2 p_k(1 - p_k)$.

Os procedimentos iterativos buscam aproximar os estimadores dos efeitos de marcadores dos verdadeiros parâmetros, e o algoritmo é encerrado quando existe a convergência para os parâmetros. Adicionalmente, o Lasso Bayesiano proporciona um melhor aprendizado com os dados do que o BayesA e o BayesB (Gianola, 2013; Gianola *et al.*, 2009), as quais foram as metodologias inicialmente aplicadas em estudos de seleção genômica ampla.

Com base nas variâncias genética e fenotípica, foram calculadas as herdabilidades de cada variável. Essa medida representa a parte da variância herdável entre os indivíduos em estudo, e é calculada como a proporção da variância genética em relação à variância total (fenotípica), da seguinte maneira para a i-ésima variável:

$$h_i^2 = \frac{V_{gen.i}}{V_{fen.i}},$$

Em que V_{gen} é a variância genética e V_{fen} é a variância fenotípica.

As sete características genômicas da curva de lactação, descritas anteriormente, foram obtidas através da estimação dos valores genéticos genômicos (*EGBV – Estimated Genomic Breeding Values*), calculado como dado abaixo:

$$\text{EGBV} = \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

em que $\mathbf{X}$ é a matriz de incidência, descrita anteriormente, e $\widehat{\boldsymbol{\beta}}$ é o vetor do efeito de marcadores. Através dos ajustes foram calculadas as herdabilidades e correlações entre os valores genômicos dos animais para cada característica, juntamente com os coeficientes de concordância baseados nos EGBVs.

2.4. Coeficiente Cohen's Kappa

Os indivíduos superiores de acordo com as características das curvas de lactação genômicas foram comparados com a utilização do coeficiente de concordância de Kappa (Cohen, 1960) dois a dois. Este método avalia, considerando a probabilidade de a seleção ter acontecido ao acaso, a concordância entre os indivíduos selecionados de acordo com diferentes critérios. O cálculo deste coeficiente pode ser feito da seguinte maneira:

$$\hat{\mathrm{k}} = \frac{\mathrm{Pr(a)} - \mathrm{Pr(e)}}{1 - \mathrm{Pr(e)}},$$

Em que o numerador $\mathrm{Pr}(a) - \mathrm{Pr}(e)$ representa a proporção de observações em que a concordância ocorreu além do esperado aleatoriamente, e o denominador $1 - \mathrm{Pr}(e)$ denota a proporção total, ponderada pela informação ao acaso. Este coeficiente pode variar de 0 a 1, e assim como o coeficiente de concordância simples, a concordância aumenta à medida que se aproxima de 1.

2.5. Aspectos computacionais

As análises estatísticas foram realizadas no software R (R Core Team, 2018). O ajuste dos modelos não lineares mistos foi feito pelo algoritmo de Linstrom e Bates (1990), implementado no pacote *nlme* (Pinheiro *et al.*, 2017). A análise de agrupamento para obtenção dos fatores de efeito fixo do modelo foi realizada com a utilização dos pacotes *cluster* (Maechler *et al.*, 2017). A análise de Seleção Genômica foi auxiliada pelos pacotes *BGLR* (DE LOS CAMPOS e RODRIGUEZ, 2016). A extração de algumas medidas descritivas foi feita com a utilização do pacote *dplyr* (Wickham *et al*., 2017).

3. Resultados

3.1. Descrição dos valores genéticos das variáveis de lactação

Os parâmetros da lactação dos 223 animais (a, b e c) foram estimados através da equação de Wood utilizando efeitos aleatórios individuais e efeitos fixos de níveis de produção, estes obtidos por análise de agrupamento. A partir desses parâmetros, as características fenotípicas (produção total, pico de lactação, persistência e tempo até o pico) foram estimadas de acordo com Wood (1967). As 7 variáveis foram submetidas à seleção genômica ampla para construção das curvas de lactação com base nos efeitos dos marcadores SNPs, onde foram calculados seus respectivos EGBVs. As medidas descritivas destes parâmetros com base na informação obtida pelo BLASSO se encontram na Tabela 1.

O coeficiente de variação para os valores genômicos estimados das variáveis apresentou-se baixo, variando de 1,39% para a taxa de ascensão até 11,02% para a produção total.

Tabela 1. Estatísticas descritivas dos valores genéticos das características de lactação, estimados pelo BLASSO.

Características	Média	Desvio padrão	CV (%)	Mínimo	Máximo
a – produção inicial (kg/dia)	8,56	0,22	2,58	8,09	9,21
b – taxa de ascensão	0,2358	0,0033	1,39	0,2283	0,2485
c – taxa de declínio	0,0054	0,0001	1,61	0,0052	0,0057
Produção total (kg/lactação)	4.680,0	515,8	11,02	3.837,0	7.306,0
Tempo até o pico (dias)	67,55	6,25	9,24	58,59	107,96
Pico de lactação (kg/dia)	18,77	0,91	4,84	17,21	23,29
Persistência	6,91	0,12	1,78	6,66	7,38

A produção inicial, representada pelo parâmetro *a*, apresentou média de 8,56 kg, com o coeficiente de variação de apenas 2,58%, variando de 8,09 a 9,21 kg. A produção total foi a característica que apresentou maior variação em relação à média, ainda que este valor possa ser considerado baixo (11,02%), indicando um baixo desvio em relação ao valor central. Podemos constatar também que existe uma amplitude de 3.469 kg/lactação entre os indivíduos mais e os menos produtivos. A persistência apresentou resultados semelhantes entre os indivíduos, com média e desvio padrão de 6,91 e 0,12, respectivamente, resultando em apenas 1,78% de coeficiente de variação. De acordo com as estimativas dos parâmetros genéticos, o pico de lactação ocorre em aproximadamente 67,55 dias, onde neste a produção média é de 18,77 kg como ponto máximo da curva de lactação. Ambas as medidas apresentaram baixos desvio padrão e coeficiente de variação. A taxa de declínio (parâmetro *c*) também apresentou baixo desvio padrão (0,0001) e baixo coeficiente de variação (1,61%), mostrando que não existe grandes diferenças entre as taxas de declínio dos indivíduos deste rebanho.

3.2. Herdabilidades e correlações entre valores genômicos

A partir das variâncias genéticas e fenotípicas, e dos EGBVs estimados pelo BLASSO, foram calculadas as herdabilidades e correlações entre os valores genômicos dos animais para as variáveis em estudo. A Tabela 2 contém essas informações, com as herdabilidades na diagonal e as correlações genômicas entre os parâmetros estudados. Podemos observar que as herdabilidades variaram de 0,09 para a taxa de ascensão até 0,29 para a persistência, sendo esta a característica mais herdável. A herdabilidade dos

parâmetros da curva de lactação obtidos pelo modelo de Wood apresentaram resultados semelhantes, sendo 0,10 para *a* e *b* e 0,09 para *c*. O pico de lactação apresentou baixa herdabilidade em relação a maioria das outras variáveis (0,12).

Tabela 2. Herdabilidades (diagonal principal) e correlações entre valores genômicos (triangular superior) entre os valores genômicos das características de lactação.

Variáveis	a[1]	b[2]	c[3]	PT[4]	Pico[5]	Pers.[6]	TP[7]
a	**0,10**	0,70	0,04	0,84	0,91	0,25	0,17
b		**0,10**	-0,20	0,87	0,88	0,54	0,40
c			**0,09**	-0,37	-0,19	-0,90	-0,85
PT				**0,27**	0,98	0,65	0,54
Pico					**0,12**	0,51	0,41
Pers.						**0,29**	0,95
TP							**0,22**

1: produção inicial (kg/dia); 2: taxa de ascensão; 3: taxa de declínio; 4: produção total (kg/lactação); 5: pico de lactação (kg/dia); 6: persistência; 7: tempo até o pico (dias).

O pico de lactação está relacionado positivamente com os parâmetros *a* (0,91) e *b* (0,88), assim como esses dois parâmetros também possuem uma correlação elevada (0,70). Isso indica que, dentre essas três variáveis, o aumento do valor genético de qualquer uma implica no aumento das demais. A correlação entre a persistência e o tempo até o pico de lactação também foi elevada (0,95), indicando que os animais que levam mais tempo para alcançar o pico de lactação apresentam também uma maior persistência. A taxa de declínio (*c*) apresentou uma alta correlação negativa com a persistência e o tempo até o pico, pois se a produção de determinado animal cai de maneira mais acentuada, este possui menor persistência, ou seja, são inversamente proporcionais.

3.3. Curvas de lactação genômicas

Podemos observar que a alta taxa de ascensão da curva leva a um pico prematuro, pois os melhores indivíduos segundo o parâmetro *b* atingem ao ápice de produtividade antes dos 50 dias de lactação, semelhante ao que ocorre com as curvas dos indivíduos superiores de acordo com a produção inicial (Figura 1). A variação da taxa de declínio nos indivíduos selecionados via produção inicial é maior do que nos outros casos, pois podemos observar que os indivíduos que mais produzem no início da lactação podem tanto apresentar uma taxa de declínio mais alta (decair rapidamente) como também

manter a produtividade alta após o pico de lactação (decair lentamente), conforme mostrado na Figura 1.

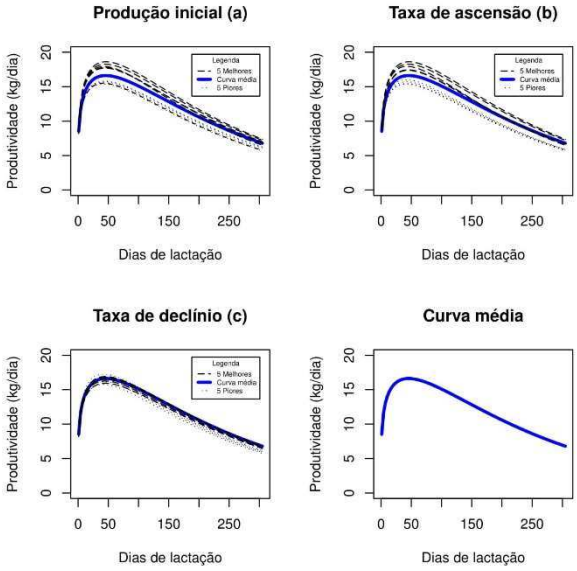


Figura 1. Curvas de lactação genômicas para os 5 indivíduos superiores e inferiores com base nos valores genômicos de *a*, *b* e *c*, juntamente com a curva média do rebanho (223 indivíduos).

As curvas de lactação dos melhores e piores indivíduos segundo a produção total, pico de lactação, persistência e tempo até o pico são mostradas na Figura 2.

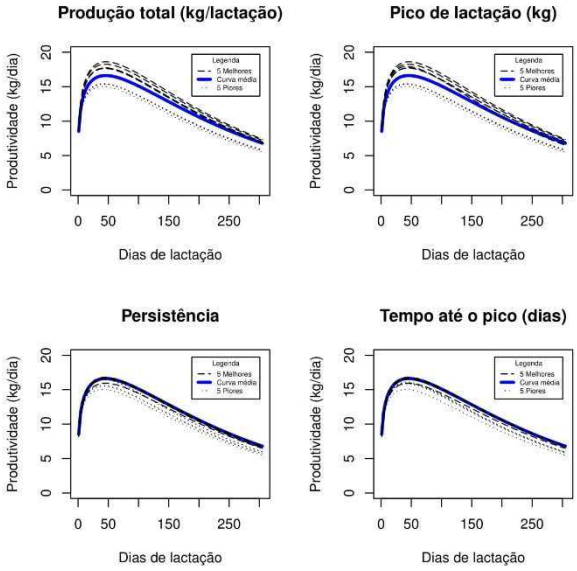


Figura 2. Curvas de lactação para os 5 indivíduos superiores e inferiores com base nos valores genômicos da produção total (kg/lactação), pico de lactação (kg/dia), persistência e tempo até o pico (dias) do rebanho (223 indivíduos).

O pico de produção dos maiores produtores de leite durante a lactação ocorre de maneira tardia em relação à média do rebanho (Figura 2). Podemos perceber também que a taxa de ascensão se mostra elevada, juntamente com uma baixa taxa de

declínio, resultando numa maior área abaixo da curva, e consequentemente no aumento da produção total destes indivíduos em relação à produtividade representada pela curva média.

3.4. Coeficientes de concordância na seleção (Kappa)

Após a estimação dos parâmetros de lactação, foram selecionados os 22 melhores indivíduos (correspondente a aproximadamente 10% da amostra total) com base em cada variável. Para a taxa de declínio (parâmetro *c*) foram escolhidos os que apresentaram menores valores, pois quanto menor a taxa de declínio, maior a capacidade do animal de continuar produzindo. O coeficiente Cohen's Kappa foi usado para verificar a concordância entre esses grupos, formados de acordo com o valor genético das variáveis de lactação estimadas pelo BLASSO, apresentando os resultados presentes na Tabela 3.

Tabela 3. Coeficiente de concordância de Kappa para os valores genômicos das variáveis de lactação.

	a[1]	b[2]	c[3]	PT[4]	Pico[5]	Pers.[6]	TP[7]
a	1,00	0,45	0,00	0,55	0,60	0,09	0,00
b		1,00	0,00	0,75	0,80	0,14	0,04
c			1,00	0,00	0,00	0,60	0,70
PT				1,00	0,95	0,29	0,19
Pico					1,00	0,24	0,14
Pers.						1,00	0,90
TP							1,00

1: produção inicial (kg/dia); 2: taxa de ascensão; 3: taxa de declínio; 4: produção total (kg/lactação); 5: pico de lactação (kg/dia); 6: persistência; 7: tempo até o pico (dias).

Podemos notar que temos concordâncias elevadas entre os parâmetros *b* e o pico de lactação (0,80) e entre a persistência e o tempo até o pico (0,90). Isso mostra que os melhores indivíduos selecionados de acordo com estes pares de variáveis são comuns entre si. Podemos destacar também o elevado coeficiente de concordância entre a produção inicial (*a*) e a produção no pico de lactação (0,60). A concordância nula (0,00) entre a produção inicial e a taxa de declínio indica que nenhum indivíduo selecionado com base em *a* está no mesmo grupo dos que foram selecionados de acordo com *c*. Os animais selecionados de acordo com a produção total são praticamente os mesmos

selecionados de acordo com o pico de lactação (Kappa de 0,95), apontando uma relação entre essas duas variáveis. Os demais resultados se encontram na Tabela 3.

4. Discussão

A equação de Wood com o uso da abordagem de modelos não lineares mistos permitiu a identificação genética dos indivíduos superiores de acordo com suas respectivas características de lactação sem o efeito dos níveis de produtividade. Assim como em Silva *et al*. (2016), os efeitos fixos (níveis de produção) e aleatórios (indivíduos) foram estimados simultaneamente através da abordagem dos modelos não lineares mistos. Como os efeitos aleatórios são centrados em torno de zero, como citado anteriormente, a média de cada parâmetro foi somada ao efeito de cada indivíduo para a obtenção das variáveis que representam as características da lactação (*a*, *b*, *c*, produção total, pico de lactação, persistência e tempo até o pico).

Nas Figuras 1 e 2, as quais apresentam as curvas genômicas dos melhores e piores animais do rebanho, podemos identificar o comportamento genômico da produção leiteira destes indivíduos. Levando em conta a produção total, por exemplo, podemos observar que os animais que mais produzem chegam ao pico de lactação pouco tempo depois (aproximadamente 50 dias) que os animais que menos produzem (aproximadamente 45 dias), e a taxa de declínio é semelhante entre os indivíduos superiores e inferiores. Essa variável foi a que mais apresentou diferenças graficamente, e também foi a que apresentou maior coeficiente de variação (11,02% - Tabela 1), o que pode explicar a maior diferença entre os indivíduos.

Silva *et al*. (2013) identificaram diferenças genéticas entre curvas genômicas de crescimento aplicando a metodologia em dois passos. Em Silva *et al*. (2013), para o comportamento de curvas de lactação genômicas, também foram identificadas diferenças genéticas entre os indivíduos considerados (Figuras 1 e 2), mesmo com o baixo coeficiente de variação de maneira geral para as variáveis. Levando em conta a produtividade individual, representada pela área abaixo da curva, maiores divergências puderam ser observadas nos indivíduos selecionados com base na produção inicial, na taxa de ascensão, no pico de lactação e na própria produção total.

As estatísticas descritivas mostradas na Tabela 1 são provenientes dos dados genéticos estimados por meio do BLASSO para cada uma das variáveis que caracterizam a lactação dos 223 animais em estudo. O baixo coeficiente pode ter ocorrido devido ao

fato de que as estimativas dos valores genômicos foram obtidas com base no valor esperado médio de cada animal, o que pode tornar a variabilidade dos valores genômicos menor do que a variabilidade fenotípica. A conclusão sobre a proporção entre as variabilidades genética e fenotípica foi avaliada pela Tabela 2, que contém as herdabilidades e correlações genômicas.

As variáveis *a*, *b* e *c*, correspondentes aos parâmetros do modelo de Wood, apresentaram baixa herdabilidade (0,10 para *a* e *b* e 0,09 para *c*). Tais herdabilidades foram inferiores às de Yilmaz *et al*. (2011) e Rekaya *et al*. (2000), estes tendo como resultados de 0,14 a 0,26 para *a*, de 0,18 a 0,32 para *b* e de 0,15 a 0,19 para *c*. Em ambos os trabalhos, foi utilizada a equação de Wood para estimação dos parâmetros. Porém, as herdabilidades do presente trabalho foram superiores às de Saghanezhad *et al*. (2017), com herdabilidades de 0,017, 0,022 e 0,06 para *a*, *b*, e *c*, respectivamente, para dados de vacas Holandeses usando o modelo de Wood. Shanks *et al*. (1981) encontraram resultados próximos para herdabilidade, sendo 0,10 para ln(*a*), utilizando uma padronização para a produção inicial, 0,06 para *b* e 0,14 para *c* na primeira lactação.

As variáveis de maior herdabilidade foram a persistência estimada segundo Wood (1967) e a produção total (kg/lactação). Canaza-Cayo *et al*. (2015) estimaram as herdabilidades para 9 diferentes estimativas de persistência, propostas por diferentes autores, encontrando resultados que variaram de 0,18 a 0,33, corroborando com a herdabilidade da persistência (0,29) encontrada no presente trabalho. Vale ressaltar que estes autores também analisaram bovinos Girolando, o que justifica a similaridade dos resultados. A herdabilidade de 0,29 para a persistência foi superior ao que foi constatado por Yilmaz *et al.* (2011), Muir *et al.* (2004) e Saghanezhad *et al*. (2017), que variaram de 0,05 até 0,23.

A herdabilidade de produção total (0,27) foi superior ao que foi constatado por Yilmaz *et al*. (2011), Pereira *et al*. (2012) e Dorneles *et al*. (2009), analisando Gados Suíços, Gir e Holstein, apresentando resultados de 0,18, 0,21 e 0,25, respectivamente. A parte herdável da produção total do presente estudo corroborou com Canaza-Cayo *et al*. (2015), os quais também analisaram gados Girolando, e constataram uma herdabilidade também de 0,27 utilizando a abordagem de modelos de regressão aleatória. O valor de 0,27 foi inferior ao que foi encontrado por Jakobsen *et al* (2002), Cobuci *et al*. (2006) e Biassus *et al*. (2010), em que suas herdabilidades variaram de 0,31 a 0,42 para gados Holandeses de diferentes localidades.

O pico de lactação apresentou baixa herdabilidade (0,12) quando comparado com López-Ordaz *et al*. (2009), Shanks *et al.* (1981), Saghanezhad *et al.* (2017), Yilmaz *et al.* (2011) e Muir *et al.* (2004), em que este resultado variou de 0,16 até 0,42. Observando a herdabilidade do tempo até o pico de lactação (0,22), temos um valor mais elevado em comparação com o que foi constatado pelos autores citados anteriormente, onde houveram herdabilidades de 0,013 até 0,13 para a mesma variável.

Tekerli *et al*. (2000) avaliaram as relações entre características de lactação por meio das correlações fenotípicas entre elas. Correlações genéticas entre características de lactação via modelos mistos foram estimadas por Boujenane e Hilal (2012), El-Awardy (2013), Farhangfar *et al*. (2007), dentre outros autores. Para o presente trabalho, foram utilizadas as correlações genômicas entre as variáveis, obtidas pela correlação simples entre os EGBVs de cada característica. Na Figura 3 constam as correlações entre os valores genômicos na triangular superior e os coeficientes Kappa para concordância na triangular inferior (valores em branco apresentaram concordância 0), e com as características de lactação na diagonal.

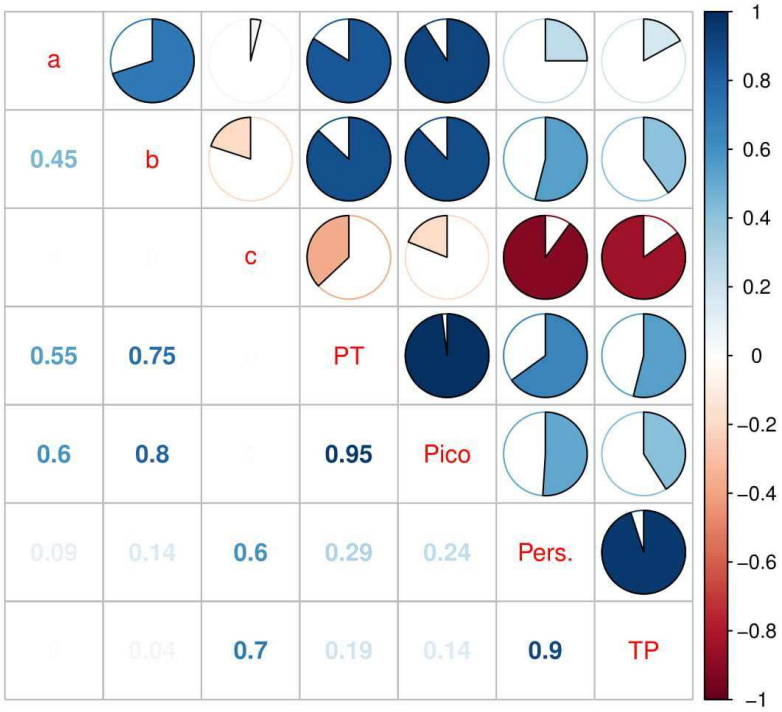


Figura 3. Correlações entre valores genômicos (triangular superior) e coeficientes de Kappa para concordância dos 10% melhores animais (triangular inferior).

Podemos observar que quando temos uma correlação genômica elevada entre duas características, o coeficiente Cohen's Kappa para seleção dos melhores indivíduos segundo as mesmas também é elevado. Isso era esperado, visto que uma alta correlação entre duas características indica o comportamento de uma em relação a outra é semelhante, acarretando também na semelhança entre os indivíduos superiores em relação às mesmas características. A correlação elevada entre a produção inicial (*a*) e a taxa de ascensão (*b*) com o pico de lactação (0,91 e 0,88, respectivamente) indica que quanto mais o animal produzir no início da lactação ou com uma alta taxa de ascensão, maior será o pico da sua curva de lactação. O pico de lactação também está fortemente correlacionado com a produção total, sendo dentre as variáveis analisadas, a característica mais associada à produtividade total dos animais em estudo, com correlação de 0,98.

Pela Figura 3, observa-se também que o coeficiente de concordância entre os indivíduos selecionados de acordo com o pico de lactação e a produção total é de 0,95, também o mais elevado dentre os coeficientes de concordância avaliados. A correlação entre valores genômicos elevada entre essas variáveis também foi encontrada por Boujenane e Hilal (2012), El-Awady (2013), Farhangfar e Rowlinson (2007), Muir *et al*. (2004), Saghanezhad *et al*. (2017) e Rekaya *et al*. (2000), com correlações genéticas variando de 0,87 a 0,98. Temos, portanto, um indício de que essas variáveis possuem uma correlação genética elevada para diferentes raças e espécies.

A produção inicial (*a*) e a taxa de ascensão (*b*) também estão estreitamente correlacionadas à produção total na lactação dos animais (0,84 e 0,87, respectivamente), juntamente com um alto coeficiente de Kappa (0,55 e 0,75, respectivamente). A correlação de 0,84 entre a produção inicial e a produção total foi superior ao que foi encontrado por Boujenane e Hilal (2012) e Rekaya *et al*. (2000), sendo 0,38 e 0,23 para bovinos Holandeses-Friesian em ambos.

A persistência de lactação está correlacionada moderadamente com a produção total e é próxima de 1 para o tempo até o pico, mostrando que animais que levam mais tempo para chegar à sua produção máxima de lactação são os que apresentam maior persistência de lactação, esta medida segundo Wood (1967). Utilizando essa medida para estimar a persistência de lactação, Boujenane e Hilal (2012) encontraram uma correlação genética de 0,77 e Saghanezhad *et al*. (2017) também constataram uma correlação genômica próxima de 1 (0,99) entre a persistência de lactação e o tempo até o pico.

5. Conclusões

A aplicação da Seleção Genômica Ampla por meio do BLASSO permitiu o conhecimento das curvas de lactação de bovinos da raça Girolando, maior produtora de leite no país, bem como a visualização da diferença genéticas entre os animais do rebanho. Tais diferenças evidenciaram a variação entre medidas genéticas avaliadas de maneira livre de fatores externos, assegurada pelo ajuste de modelos não-lineares mistos. Além desse conhecimento, essa metodologia permitiu o acesso às herdabilidades e correlações entre os valores genômicos das características estudadas. A obtenção de valores genômicos propiciou a identificação dos melhores indivíduos de acordo com diferentes características, e por meio destes grupos, analisando o coeficiente de concordância para os indivíduos selecionados de acordo com diferentes parâmetros de lactação, encontrou-se um coeficiente elevado nas comparações onde houve maior correlação entre valores genômicos.

6. Referências

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; SILVA, F. F.; SERÃO, N. V. L.; CRUZ, C. D.; RESENDE, M. D. V.; SILVA, F. L.; AZEVEDO, C. F.; LOPES, P. S.; GUIMARÃES, S. E. F. Regularized quantile regression for SNP marker estimation of pig growth curves. **Journal of Animal Science and Biotechnology**, v. 8, n. 59, p. 1-9, 2016.

BOUJENANE, I.; HILAL, B. Genetic and non genetic effects for lactation curve traits in Holstein-Friesian cows. **Archiv Tierzucht**, v. 5, p. 450-457, 2012.

BIASSUS, I. O.; COBUCI, J. A.; COSTA, C. N.; RORATO, P. R. N.; BRACCINI NETO, J.; CARDOSO, L. L. Persistence in milk, fat and protein production of primiparous Holstein cows by random regression models. **Revista Brasileira de Zootecnia**, v. 39, p. 2617-2624, 2010.

CARDONA, S. J. C.; ÁLVAREZ, J. D. C.; SARMENTO, J. L. R.; HERRERA, L. G. G.; CADAVID, H. C. Associação entre SNPs nos genes para k-caseína e β-lactoglobulina com curvas de lactação em cabras leiteiras. **Pesquisa Agrupecuária Brasileira**. V. 50, n. 3, p. 224-232, 2015.

CANAZA-CAYO, A. W.; LOPES, P. S.; SILVA, M. V. G. B.; TORRES, R. A.; MARTINS, M. F.; ARBEX, W. A.; COBUCI, J. A. Genetic Parameters for Milk Yield Lactation Persistency Using Random Regression Models in Girolando Cattle. **Asian Australasian Journal of Animal Science**, v. 28, n. 10, p. 1407-1418, 2015.

COBUCI, J. A.; COSTA, C. N.; TEIXEIRA, N. M.; FREITAS, A. F. Use of Lagendre polynomials and Wilmink functions in genetic evaluations for persistency of lactation in Holstein cows. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 58, p. 614-623, 2006.

COHEN, J. A coeficient of agreement for nominal scales. **Educational and Psychologial Measurement.**, v. 20, p. 37-46, 1960.

DE LOS CAMPOS, G.; RODRIGUES, P. P. Bayesian Generalized Linear Regression. R package version 1.0.5. URL: , 2016.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, n. 1, p. 375-385, 2009.

DIJKSTRA, J. FRANCE, J.; DHANOA, M. S.; MAAS, J. A.; HANIGAN, M. D.; ROOK, A. J.; BEEVER, D. E. A model to describe growth curve patterns of the mammary gland during pregnancy and lactation. **Journal of Dairy Science**, v. 80, p. 2340-2354, 1997.

DORNELES, C. K.; COBUCI, J. A.; RORATO, P. R. N.; WEBER, T.; LOPES, J. S.; OLIVEIRA, H. N. Estimation of genetic parameters for Holstein cows milk production by random regression models. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 61, p. 407-412, 2009.

EL-AWADY, H. G. Genetic aspects of lactation curve traits and persistency índices in Friesian cows. **Archiva Zootechnica**, v. 16, n. 1, p. 15-29, 2013.

FAN, B.; ONTERU, S. K.; DU, Z. Q.; GARRICK, D. J.; STALDER, K. J.; ROTHSCHILD, M. F. Genome-wide association study identifies Loci for composition and sctructural soundness traits in pigs. **PlosOne**. v. 6. p. 1-11. 2011.

FARHANGFAR, H.; ROWLINSON, P. Genetic Analysis of Wood's Lactation Curve for Iranian Holstein Heifers. **Journal of Biological Sciences**, v. 7, n. 1, p. 127-135, 2007.

GANTNER, V.; JOVANOVAC, S.; RAGUZ, N.; SOLIC, D.; KUTEROVAC, K. Nonlinear Vs. linear regression models in lactation curve prediction. **Bulgarian Journal of Agricultural Science**, v. 16, n. 6, p. 794-800, 2010.

GHAVI HOSSEIN-ZADEH, N. Comparison of non-linear models to describe the lactation curves for milk yield and composition in buffaloes (*Bubalus bubalis*). **Animal**, Cambridge, v. 10, n. 2, p. 248-261, 2015.

GHAVI HOSSEIN-ZADEH, N. Comparison of non-linear models to describe the lactation curves of milk yield and composition in Iranian Holsteins. **The Journal of Agricultural Science**, v. 152, p. 309-324, 2014.

GIANOLA, D. Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. **Genetics**, v. 194, n. 3, p. 573-596, 2013.

GIANOLA, G.; DE LOS CAMPOS, G.; HILL, W. G.; MANFREDI, E.; FERNANDO, R. Additive Variability and the Bayesian Alphabet. **Genetics**, v. 183, n. 1, p. 347-363, 2009.

GOWER, J. C. A general coeficiente of similarity and some of its properties. **Biometrics**, v. 27, p. 857-54, 1971.

HENDERSON, C. R. Aplications of linear models in animal breeding. University of Guelph, Guelph, 1984.

JAKOBSEN, J. H.; MADSEN, P.; JENSEN, J.; PEDERSEN, J.; CHRISTENSEN, L. G.; SORENSEN, D. A. Genetic parameters for milk production and persistency for Danish Holsteins estimated in random regression models using REML. **Journal of Dairy Science**, v. 85, p. 1607-1616, 2002.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, p. 743-756, 1990.

LINDSTOM, M. J.; BATES, D. M. Nonlinear Mixed Effects Models for Repeated Measures Data. **Biometrics**, v. 46, n. 3, p. 673-687, 1990.

LÓPEZ-ORDAZ, R.; CASTILLO-JUÁREZ, H.; MONTALDO, H. H. Covarianzas genéticas y fenotípicas para días abiertos y características de la curva de lactancia em vacas Holstein em el norte de México. **Revista Veterinaria México**, v. 40, n. 4, p. 343-356, 2009.

MACCIOTTA, N. P. P.; GASPA, G.; BOMBA, L.; VICARIO, D.; DIMAURO, C.; CELLESI, M.; AJMONE-MARSAN, P. Genome-wide association analysis in Italian Simmental cows for lactation curve traits using low-density (7K) SNP panel. **Journal of Dairy Science**, v. 98, p. 8175-8185, 2015.

MACCIOTTA, N. P. P.; DIMAURO, C.; RASSU, S. P. G.; STERI, R.; PULINA, G. The mathematical description of lactation curves in dairy cattle. **Italian Journal of Animal Science**, v. 10, n. 51, p. 213-223, 2011.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. (2017). **cluster: Cluster Analysis Basics and Extensions**. R package version 2.0.6, 2017.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome wide dense marker maps. **Genetics**, v. 157, p. 1819-29, 2001.

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

MUIR, B. L.; FATEHI, J.; SCHAEFFER, L. R. Genetic Relationships Between Persistency and Reproductive Performance in First-Lactation Canadian Holsteins. **Journal of Dairy Science**, v. 87, p. 3029-3037, 2004.

NELDER, J. A. Inverse polynomials, a useful group of multi-factor response functions. **Biometrics**, v. 22, p. 128-141, 1966.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, v. 103, n. 482, p. 681-686, 2008.

PEREIRA, R. J.; VERNEQUE, R. S.; LOPES, P. S.; SANTANA JR, M. L.; LAGROTTA, M. R.; TORRENS, R. A.; VERCESI FILHO, A. E.; MACHADO, M. A. Milk yield persistency in Brazilian Gyr cattle based on a rangom regression model. **Genetics and Molecular Research**, v. 11, n. 2, p. 1599-1609, 2012.

PICCARDI, M.; MACCHIAVELLI, E.; FUNES, A. C.; BÓ, G. A.; BALZARINI, M. Fitting milk production curves through nonlinear mixed models. **Journal of Dairy Research**, v. 84, n. 2, p. 146-153, 2017.

PING-WOND, R.; HADJIPAVLOU, G. A. A two-step approach combining the Gompertz growth with genomic selection for longitudinal data. **BMC Proceedings**, v. 4 (S4), 2010.

PINHEIRO, J. C.; BATES, D. M. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. **Journal of Computational and Graphical Statistics**, v. 4, n. 1, p. 12-35, 1995.

PINHEIRO, J.; BATES, D.; DEBROY, S.; SARKAR, D. and R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131, URL: https://CRAN.R-project.org/package=nlme, 2017.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

REKAYA, R.; CARABAÑO, M. J.; TORO, M. A. Bayesian Analysis of Lactation Curves in Holstein-Friesian Cattle. **Journal of Dairy Science**, v. 83, p. 2691-2701, 2000.

SAGHANEZHAD, F.; ATASHI, H.; DADPASAND, M.; ZAMIRI, M. J.; SHOKRI-SANGARI, F. Estimation of Genetic Parameters for Lactation Curve Traits in Holstein Dairy Cows in Iran. **Iranian Journal of Applied Animal Science**, v. 7, n. 4, p. 559-566, 2017.

SHANKS, R. D.; BERGER, P. J.; FREEMAN, A. E.; DICKINSON, F. N. Genetic Aspects of Lactation Curves. **Journal of Dairy Science**, v. 64, n. 9, p. 1852-1860, 1981.

SILVA, F. F.; RESENDE, M. D. V.; ROCHA, G. S.; DUARTE, D. A. S.; LOPES, P. S.; BRUSTOLINI, O. J. B.; THUS, S.; VIANA, J. M. S.; GUIMARÃES, S. E. F. Genomic

growth curves of an outbred pig population. **Genetics and Molecular Biology**, v. 36, n. 4, p. 520-527, 2013.

SILVA, F. F.; ZAMBRANO, M. F. B.; VARONA, L.; GLÓRIA, L. S.; LOPES, P. S.; SILVA, M. V. G. B.; ARBEX, W.; LÁZARO, S. F.; RESENDE, M. D. V.; GUIMARÃES, S. E. F. Genome association study through nonlinear mixed models revealed new candidate for pig growth curves. **Scientia Agricola**, v. 74, n. 1, p. 1-7, 2016.

TEIXEIRA, F. R. F.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; SILVA, F. F.; CRUZ, C. D.; AZEVEDO, C. F.; PAIXÃO, D. M.; BARROSO, L. M.; VERARDO, L. L.; RESENDE, M. D.; GUIMARÃES, S. E. F.; LOPES, P. S. Factor Analysis applied to genome prediction for high-dimensional phenotypes in pigs. **Genetics and Molecular Research**, v. 15, n. 2, p. 1-10, 2016.

TEKERLI, M.; AKINCI, Z.; DOGAN, I.; AKCAN, A. Factor Affecting the Shape of Lactation Curves of Holstein Cows from the Balikesir Province of Turkey. **Journal of Dairy Science**. V. 83, p. 1381-1386, 2000.

VARONA, L.; MORENO, C.; GARCIA-CORTÉS, L. A.; YAGUE, G.; ALTARRIBA, J. Two-step vs. Joint analysis of Von Bertalanffy function. **Journal of Animal Breeding and Genetics**, v. 116, p. 331-338, 1999.

WICKHAM, H. FRANCOIS, R.; HENRY, L.; MULLER, KIRILL. **dplyr: A Grammar of Data Manipulation**. R package version 0.7.3. URL: https://CRAN.R-project.org/package=dplyr, 2016.

WOOD, P. D. P. Algebraic model of the lactation curve in cattle. **Nature**, v. 216, p. 164-165, 1967.

YILMAZ, I.; EYDURAN, E.; KAYGISIZ, A.; JAVED, K. Estimates of Genetic Parameters for Lactation Shape Parameters with Multivariate Statistical Technique in Brown Swiss Cattle. **International Journal of Agriculture and Biology**. v. 13, n. 2, p. 174-178, 2011.

CONSIDERAÇÕES FINAIS

O presente trabalho teve como intuito principal a construção de curvas de lactação de gados da raça Girolando, os quais são de extrema importância no contexto da produção leiteira nacional, com base em parâmetros obtidos por modelos não-lineares mistos. Visou-se também a identificação do valor genético das características de lactação baseadas nessas informações. Dentro desse objetivo geral, várias informações importantes foram obtidas através dessa nova proposta.

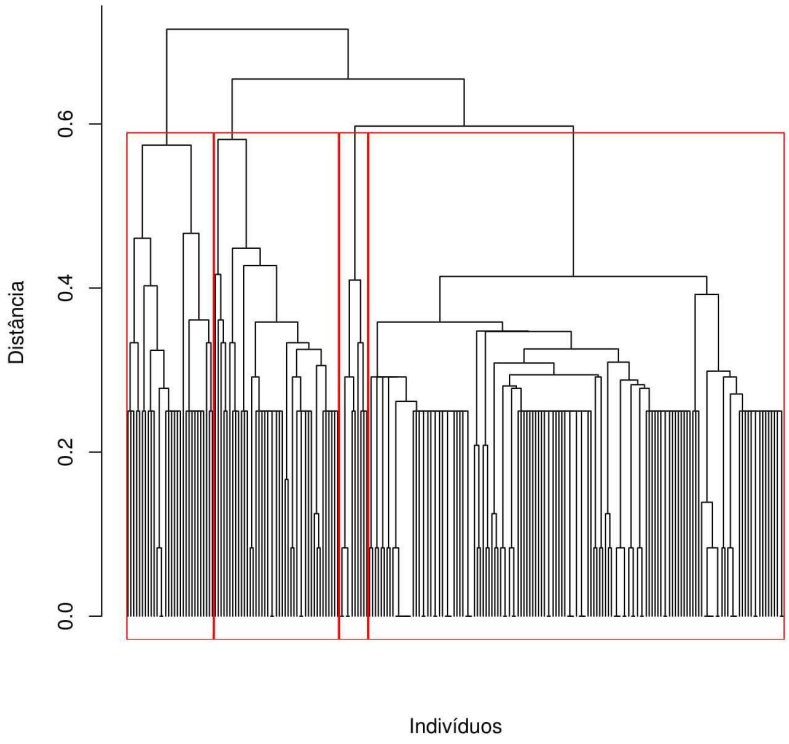
A primeira delas pôde ser observada no capítulo 1, o qual teve como objetivo a identificação da melhor equação para descrever os dados de lactação dos animais em estudo, dentre nove modelos não-lineares comumente utilizados na literatura. A partir da análise das informações, pudemos concluir que o modelo que melhor descreve a lactação de bovinos Girolando é a equação de Wood, de acordo com os critérios de comparação utilizados. Além disso, a partir das informações sobre a lactação obtidas, fomos capazes de identificar um grupo seleto de animais superiores em relação à média do rebanho e até mesmo em relação a outras raças que, em média, produzem mais leite que a raça Girolando.

Já no capítulo 2, a finalidade foi a construção das curvas de lactação genômicas para os mesmos indivíduos na análise baseadas nas informações obtidas no primeiro capítulo. A metodologia utilizada para este propósito foi a Seleção Genômica Ampla, por meio do BLASSO – *Bayesian LASSO*. Após a sua aplicação, tivemos acesso ao comportamento genômico da lactação dos indivíduos do rebanho, que pôde ser evidenciado quando a identificação dos animais superiores e inferiores de acordo com cada característica de lactação, além de permitir o conhecimento da associação entre essas tais características.

Portanto, a abordagem proposta apresentou resultados satisfatórios, permitindo o amplo conhecimento das curvas de lactação fenotípica e genômica do rebanho em estudo e se mostrando como alternativa também para estudos posteriores envolvendo diferentes raças ou espécies.

APÊNDICES

Apêndice I: Dendograma para formação dos grupos de efeitos fixos de acordo com a número de ordenhas (3 grupos), idade (4 grupos), agrupamento genético (3 grupos) e grupos contemporâneos (139 grupos).



Apêndice II: Alguns scripts utilizados na análise.

```
### 1) LEITURA DOS PACOTES ###
library(nlme)

library(dplyr)
library(ggdendro)
library(cluster)
library(BGLR)

### 2) LEITURA DO BANCO DE DADOS ###
dados=read.table("C:\\Users\\filip\\Dropbox\\Doutorado
2018\\Dados\\dados.txt",h=T) #lendo os dados fenotípicos longitudinais
corrigidos
dadosg=read.table("C:\\Users\\filip\\Dropbox\\Doutorado
2018\\Dados\\dados2.txt",h=T) #lendo os dados fenotípicos
longitudinais corrigidos
dados2<-groupedData(prod~dia|id,dados)

### 3) FORMAÇÃO DOS GRUPOS PELO ALGORITMO DE GOWER ###

clust<-cbind(dadosg$ida,dadosg$ord,dadosg$agr,dadosg$cont)
```

```
rownames(clust)=rownames(dadosg)
colnames(clust)=c("ida","ord","agr","cont")

### Análise de agrupamento para formação de efeitos fixos ###
clust2<-
data.frame(cbind(clust[,1],clust[,2],as.factor(clust[,3]),factor(clust
[,4])))
clust2[,3]<-as.factor(clust2[,3])
clust2[,4]<-as.factor(clust2[,4])
colnames(clust2)=c("ida","ord","agr","cont")
distG<-daisy(clust2,metric="gower")
summary(distG)
ac1<-hclust(distG,method="average")
plot(ac1,hang=-1,cex=0.001,xlab="Indivíduos",ylab="Distância de
Gower",main="",sub="")
mojena<-mean(ac1$height)+2.75*sd(ac1$height) # Dendograma
k=length(ac1$height[ac1$height>mojena]) + 1
k # k=5, logo:
rect.hclust(ac1, k = 5)
grupos2<-cutree(ac1, k=5)
gower2<-grupos2
G<-factor(gower2) # Grupos formados

### Estatísticas descritivas por grupo (efeitos fixos) ###

tapply(dados2$prod,dados2$gower2,summary) # summary por grupo
tapply(dados2$prod,dados2$gower2,length) # n por grupo
tapply(dados2$prod,dados2$gower2,sd) # desvio padrão por grupo

##### 4) AJUSTE DOS MODELOS NÃO LINEARES MISTOS (ORDEM) #####

### 4.1) Modelo de Brody (1923) ###

Brody<-function(dia,a,c){
a*exp(-c*dia)}

a1=rep(2.71,5)
a2=rep(0.0125,5)
a3=rep(0.0045,5)
a4<-c(a1,a2)

Brody.nlme.g <- nlme(prod~Brody(dia,a,c),data=dados2,fixed=list(a~G-1,
c~G-1),random=pdDiag(a+c~1),start=list(fixed=a4))

summary(Brody.nlme.g)

### 4.2) Modelo de Sikka (1950) ###

Sikka<-function(dia,a,b,c){
a*exp(b*dia-c*dia^2)}

a1=rep(2.71,5)
a2=rep(0.002,5)
a3=rep(0.00045,5)
a4<-c(a1,a2,a3)

Sikka.nlme.g <- nlme(prod~Sikka(dia,a,b,c),data=dados2,fixed=list(a~G-
1,
b~G-1,c~G-1),random=pdDiag(a+b+c~1),start=list(fixed=a4))

summary(Sikka.nlme.g)

### 4.3) Modelo de Nelder (1966) ###

Nelder<-function(dia,a,b,c){
dia/(a+b*dia+c*dia^2)}

a1=rep(0.1,5)
a2=rep(0.01,5)
a3=rep(0.001,5)
```

```
a4<-c(a1,a2,a3)

Nelder.nlme.g <-
nlme(prod~Nelder(dia,a,b,c),data=dados2,fixed=list(a~G-1,
b~G-1,c~G-1),random=pdDiag(a+b+c~1),start=list(fixed=a4))

summary(Nelder.nlme.g)

### 4.4) Modelo de Wood (1967) ###

Wood<-function(dia,a,b,c){
a*dia^b*exp(-c*dia)} # função de Wood (1967)

a1=rep(2.556667,5)     # 1      # 5.42     # 2.556667
a2=rep(0.04810351,5)   # 0.1    # 0.968    # 0.04810351
a3=rep(0.001334639,5)  # 0.01   # 0.00152  # 0.001334639
a4<-c(a1,a2,a3)

Wood.nlme.g<-nlme(prod~Wood(dia,a,b,c),data=dados2,fixed=list(a+b+c~G-
1),
random=pdDiag(a+b+c~1),start=list(fixed=a4))

summary(Wood.nlme.g)

### 4.5) Modelo de Cobby e Le Du (1978) ###

Cobby<-function(dia,a,b,c){
a-b*dia-a*exp(-c*dia)}

a1=rep(10,5)
a2=rep(0.1,5)
a3=rep(0.01,5)
a4<-c(a1,a2,a3)

Cobby.nlme.g <- nlme(prod~Cobby(dia,a,b,c),data=dados2,fixed=list(a~G-
1,
b~G-1,c~G-1),random=pdDiag(a+b+c~1),start=list(fixed=a4))

summary(Cobby.nlme.g)

### 4.6) Modelo de Dhanoa (1981) - PENDENTE ###

Dhanoa<-function(dia,a,b,c){
a*(dia^(b*c))*exp(-c*dia)}

a1=rep(150,5)          # 50.42
a2=rep(60.379,5)       # 6.79
a3=rep(0.002,5)       # 0.0002
a4<-c(a1,a2,a3)

Dhanoa.nlme.g <-
nlme(prod~Dhanoa(dia,a,b,c),data=dados2,fixed=list(a~G-1,
b~G-1,c~G-1),random=pdDiag(a+b+c~1),start=list(fixed=a4))

summary(Dhanoa.nlme.g)

### 4.7) Papajcsik e Bodero (1988) ###

PB<-function(dia,a,c){
a*dia*exp(-c*dia)}

a1=rep(1,5)
a2=rep(0.002,5)
a3=rep(0.72,5)
a4<-c(a1,a2)

Papajcsik.nlme.g <- nlme(prod~PB(dia,a,c),data=dados2,fixed=list(a~G-
1,
c~G-1),random=pdDiag(a+c~1),start=list(fixed=a4))
```

```
summary(Papajcsik.nlme.g)

### 4.8) Modelo de Rook (1993) - OK ###

Rook<-function(dia,a,b,c,d){
a*(1/(1+b/(c+dia)))*exp(-d*dia)}

a1=rep(10,5)
a2=rep(35,5)
a3=rep(0.1,5)
a4=rep(0.0015,5)
a5<-c(a1,a2,a3,a4)

Rook.nlme.g <- nlme(prod~Rook(dia,a,b,c,d),data=dados2,fixed=list(a~G-
1,
b~G-1,c~G-1,d~G-1),random=pdDiag(a+b+c+d~1),start=list(fixed=a5))

summary(Rook.nlme.g)

### 4.9) Modelo de Cappio-Borlino (1995) ###

Cappio<-function(dia,a,b,c){
a*dia^(b*exp(-c*dia))} # função de Cappio (1995)

a1=rep(7.8,5) #30
a2=rep(0.25,5)  #1
a3=rep(-0.000353,5)  #0.1
a4<-c(a1,a2,a3)

Cappio.nlme.g <-
nlme(prod~Cappio(dia,a,b,c),data=dados2,fixed=list(a~G-1,b~G-1,
c~G-1),random=pdDiag(a+b+c~1),start=list(fixed=a4))

summary(Cappio.nlme.g)

##### 5) Seleção genômica (BLASSO) #####
rm(list = ls())
library(BGLR)

setwd("C:\\Users\\filip\\Desktop\\girolando")

dados<-read.table("dadosdef.txt",h=T)
dadosfen<-dados[,1:8]
### Genótipo + frequências alélicas ###

geno<-as.matrix(dados[,-c(1:8)]) # SNPs
p2=matrix(0,ncol(geno),1)
q2=matrix(0,ncol(geno),1)
for(i in 1:ncol(geno))
{
q2[i,]=(2*length(which(geno[,i]==0))+length(which(geno[,i]==1)))/(2*le
ngth(which(geno[,i]==0))+2*length(which(geno[,i]==1))+2*length(which(g
eno[,i]==2)))
p2[i,]=(length(which(geno[,i]==1))+2*length(which(geno[,i]==2)))/(2*le
ngth(which(geno[,i]==0))+2*length(which(geno[,i]==1))+2*length(which(g
eno[,i]==2)))
}
names(dados[,1:8]) # nomes das variáveis

##### BLASSO - a e b #####

### a ###
feno<-as.matrix(dados$a) # $ variável resposta

BL.a=BGLR(y=feno,ETA=list(list(X=geno,model='BL')),nIter=100000,burnIn
=20000,thin=10)
gebv_BL.a<-geno%*%BL.a$ETA[[1]]$b
cor_BL.a<-cor(gebv_BL.a,feno)
v=BL.a$ETA[[1]]$tau2
Ve=BL.a$varE
```

```
t=matrix(v)*BL.a$varE
Va=sum(2*p2*q2*t)
Vfen=Va+Ve
h2aBL.a=Va/Vfen
h2aBL.a
acurBL.a<-cor_BL.a/sqrt(h2aBL.a)
acurBL.a
va.a<-Va
vfen.a<-Vfen

### b ###
feno<-as.matrix(dados$b) # $ variável resposta

BL.b=BGLR(y=feno,ETA=list(list(X=geno,model='BL')),nIter=100000,burnIn
=20000,thin=10)
gebv_BL.b<-geno%*%BL.b$ETA[[1]]$b # EGBV
cor_BL.b<-cor(gebv_BL.b,feno)
v=BL.b$ETA[[1]]$tau2
Ve=BL.b$varE
t=matrix(v)*BL.b$varE
Va=sum(2*p2*q2*t)
Vfen=Va+Ve
h2aBL.b=Va/Vfen
h2aBL.b # Herdabilidade
acurBL.b<-cor_BL.b/sqrt(h2aBL.b)
acurBL.b
va.b<-Va
vfen.b<-Vfen

### Coeficiente Cohen's Kappa ###
## Identificando os melhores indivíduos segundo cada variável ##
summary(dadosgBL2) # dadosgBL2 é a matriz que contém os EGBVs
head(dadosgBL2)

id.a<-dadosgBL2[order(dadosgBL2$a.gen,decreasing=TRUE),][1:22,]$id
id.b<-dadosgBL2[order(dadosgBL2$b.gen,decreasing=TRUE),][1:22,]$id

## Cohen's Kappa ##
# a e b #
id.a # primeiro grupo (ID)
id.b # segundo grupo (ID)

ab<-cbind(id.a,id.b)

cont<-0
for(i in 1:nrow(ab))
{
  for(j in 1:nrow(ab))
  {
   if(ab[i,1]==ab[j,2])
   cont<-cont+1
   cont<-cont
  }
}
c<-cont/nrow(ab)

kappa.ab<- (c-(22/223))/(1-(22/223))
kappa.ab
```