GERALDO MAGELA DA CRUZ PEREIRA

MEDIDAS ALTERNATIVAS PARA COMPARAÇÃO DE MODELOS E APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA E DE REDUÇÃO DE DIMENSIONALIDADE PARA SELEÇÃO GENÔMICA COM DADOS CENSURADOS

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*.

Orientador: Sebastião Martins Filho

Coorientadora: Renata Veroneze

VIÇOSA - MINAS GERAIS
2020

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

T
P436m
2020

Pereira, Geraldo Magela da Cruz, 1987-
Medidas alternativas para comparação de modelos e aplicação de métodos de aprendizado de máquina e de redução de dimensionalidade para seleção genômica com dados censurados / Geraldo Magela da Cruz Pereira. – Viçosa, MG, 2020.
87 f. : il. (algumas color.) ; 29 cm.

Orientador: Sebastião Martins Filho.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Melhoramento genético - Modelos estatísticos. 2. Aprendizado do computador. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 519.5

GERALDO MAGELA DA CRUZ PEREIRA

MEDIDAS ALTERNATIVAS PARA COMPARAÇÃO DE MODELOS E APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA E DE REDUÇÃO DE DIMENSIONALIDADE PARA SELEÇÃO GENÔMICA COM DADOS CENSURADOS

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: 11 de maio de 2020.

Assentimento:

Geraldo Magela da Cruz Pereira
Autor

Sebastião Martins Filho
Orientador

A Deus, aos meus pais, Geny e José Carlos,

À minha irmã Jisleny e aos meus irmãos José Carlos e Francisco,

À minha irmã de coração, Carolina (in memoriam),

Aos meus sobrinhos, familiares e amigos.

AGRADECIMENTOS

Agradeço a Deus por minha vida, pelas pessoas incríveis que colocou em meu caminho, por me dar força para lutar e conquistar meus objetivos, e por me dar sabedoria para fazer escolhas acertadas nos momentos de incertezas.

Aos meus pais Geny e José Carlos, por todo amor e carinho que recebo, por todas as orações e pensamentos positivos que emanaram, e que me fizeram acreditar que tudo seria possível em minha vida. Agradeço também por todas as lições de vida que me foram dadas, e que sempre carrego comigo para onde quer que eu vá.

À minha irmã Jisleny por nossa amizade, pelas palavras de incentivo e direção, por me apoiar incondicionalmente, e por me ensinar que, agindo com foco e determinação, temos força de vontade suficiente para alcançar qualquer objetivo na vida.

Aos meus irmãos José Carlos e Francisco, por nossa amizade e companheirismo, por nossa convivência e pelos momentos de descontração que compartilhamos, vocês certamente me ajudaram a concluir o Doutorado com mais leveza.

Ao Doutor André Gomes pelo companheirismo, por dividir comigo todas as angustias e felicidades de se cursar um Doutorado, mesmo que em programas de pós-graduação diferentes, por nossas conversas, viagens e idas a barzinhos. Nossa convivência foi fundamental para que este objetivo fosse concluído.

Aos meus amigos Raphael, Leonardo e Patricia, que apesar da distância, trabalho ou pós-graduação, se mostram sempre presentes, dispostos a ouvir, a dar bons conselhos, e a sair para se divertir, pessoas com as quais tenho lembranças e laços que certamente irão durar por toda minha vida.

À amiga Magnória Santos, por nossa amizade de anos, nossos diálogos sobre a vida, de um modo geral, e por me mostrar a importância das coisas simples da vida.

À Universidade Federal de Viçosa (UFV) e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria (PPESTBIO), por me fornecer apoio e a estrutura necessária para realização do Doutorado.

Ao meu orientador, professor Dr. Sebastião Martins Filho, pela confiança, conselhos e paciência durante o desenvolvimento da tese.

À minha coorientadora Renata Veroneze, por sua disponibilidade, conselhos e questionamentos.

Aos membros da banca de qualificação e de defesa do Doutorado, Drs. Leonardo Siqueira Glória, Antônio Policarpo Souza Carneiro, Renata Veroneze, Sebastião Martins Filho, Luiz Fernando Brito, e Vinícius Silva dos Santos, pela disponibilidade, correções, críticas, sugestões e pelo diálogo construtivo.

À ex-secretária do Programa de Pós-Graduação em Estatística Aplicada e Biometria, Carla Zinato Campos, por toda a sua torcida, por nossas conversas sobre temas diversos, por sempre me receber com um sorriso espontâneo e alegre, e por sua competência profissional. Ao atual secretário Júnior por sua competência e suporte durante a realização das atividades do Doutorado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

PEREIRA, Geraldo Magela da Cruz, D.Sc., Universidade Federal de Viçosa, maio de 2020. **Medidas alternativas para comparação de modelos e aplicação de métodos de aprendizado de máquina e de redução de dimensionalidade para seleção genômica com dados censurados.** Orientador: Sebastião Martins Filho. Coorientadora: Renata Veroneze.

Dados censurados são encontrados em diversas características de interesse no melhoramento animal, como por exemplo, tempo ao abate em suínos, idade ao primeiro parto em bovinos, resistência à doença em peixes. A modelagem destas características é comumente realizada via modelos lineares, que podem ou não considerar a natureza censurada dos dados. Os modelos G-BLUP, RR-BLUP e ssGBLUP são exemplos de modelos que não consideram a presença de observações incompletas nos dados. A classe de modelos bayesianos BGLR (*Bayesian Generalized Linear Regression*), possibilita a modelagem de fenótipos censurados. Recentemente tem surgido o interesse na utilização de modelos de sobrevivência para a análise de dados genômicos com observações censuradas. Neste contexto, estudos que avaliem a utilização de medidas mais adequadas para o cálculo da acurácia e do viés, bem como a utilização de métodos de aprendizado de máquina de sobrevivência, não foram encontrados na literatura consultada. O objetivo geral deste estudo foi contribuir para a discussão acerca das metodologias mais indicadas para a comparação de modelos, e para a realização de predições em estudos de seleção genômica com dados censurados simulados e reais de juvenis de dourada (*Sparus aurata*). As metodologias propostas foram comparadas com as metodologias tradicionalmente utilizadas em genômica. Para os dados simulados, foram comparadas as medidas de correlação: de Pearson (CP), maximal (CM) e de Pearson para dados censurados (CPC); e de viés: regressão linear simples e regressão Tobit. A predição de valores genéticos genômicos foi realizada pelos modelos misto de Cox e normal truncado, considerando diferentes cenários. Os resultados mostraram, que principalmente no cenário com herdabilidade de QTL igual à 0,27, as medidas CM e/ou CPC, mostraram-se estatisticamente superiores à CP. O coeficiente de regressão associado aos efeitos marginais para dados censurados e não censurados apresentou valores semelhantes aos obtidos pela regressão linear. Do ponto de vista estatístico, as metodologias propostas são mais adequadas para a análise de dados censurados, visto que em sua formulação, elas consideram a presença de fenótipos não observados. Para os dados reais, foi considerada a utilização dos métodos *Random Survival Forest* (RSF) e *Gradient Boosting Machine* e Análise de Componentes Principais Supervisionados em seleção genômica, sendo estes comparados ao método Regressão Ridge Bayesiana (BRR). Os modelos foram

comparados via validação cruzada *7-fold*, pelas medidas *Area Under the Curve*, *Brier Score*, correlação de Spearman, e pela proporção de indivíduos selecionados, e também pela localização de SNPs ou grupos de ligação relevantes. Os resultados mostraram que, os modelos RSF e BRR, apresentaram valores estatisticamente iguais de habilidade preditiva. O rank dos Top-40 SNPs obtido pela RSF apresentou maior interseção com os ranks obtidos pelos métodos BRR e modelo misto de Cox. A maior correlação de Spearman entre os GEBVs estimados via BRR e as probabilidades de sobrevivência, foi obtida pela RSF. A utilização de subconjuntos de SNPs selecionados pelos métodos propostos, não resultou em diferenças significativas na habilidade preditiva do modelo misto de Cox. Por fim, nota-se que o método RSF, apresenta um desempenho semelhante ao da BRR, sendo possível sua aplicação em estudos genômicos.

Palavras-chave: Seleção genômica ampla. Valores genéticos genômicos. Dados censurados. Modelo misto de Cox. Aprendizado de máquina.

ABSTRACT

PEREIRA, Geraldo Magela da Cruz, D.Sc., Universidade Federal de Viçosa, May, 2020. **Alternative measures for model comparison and application of machine learning and dimensionality reduction methods for genomic selection with censored data.** Adviser: Sebastião Martins Filho. Co-Adviser: Renata Veroneze.

Censored data are found in several characteristics of interest in animal breeding, such as, time to slaughter in pigs, age at first calving in cattle, resistance to disease in fish. The modeling of these characteristics is commonly performed via linear models, which may or may not consider the censored nature of the data. The G-BLUP, RR-BLUP and ssGBLUP models are examples of models that do not consider the presence of incomplete observations in the data. The class of Bayesian models BGLR (Bayesian Generalized Linear Regression), allows the modeling of censored phenotypes. Recently there has been an interest in the use of survival models for the analysis of genomic data with censored observations. In this context, studies evaluating the use of more appropriate measures to calculate accuracy and bias, as well as the use of survival machine learning methods, were not found in the literature consulted. The general objective of this study was to contribute to the discussion about the most suitable methodologies for the comparison of models, and for the realization of predictions in studies of genomic selection with censored data simulated and real of juveniles of Gilthead Sea Bream (*Sparus aurata*). The proposed methodologies were compared with those traditionally used in genomics. For the simulated data, the correlation measures were compared: Pearson (CP), maximal (CM) and Pearson for censored data (CPC); and bias: simple linear regression and Tobit regression. The prediction of genomic breeding values was performed by the mixed Cox and Normal truncated models, considering different scenarios. The results showed that, especially in the scenario with heritability of QTL equal to 0.27, the CM and / or CPC measures were statistically superior to the CP. The regression coefficient associated with the marginal effects for censored and uncensored data showed values similar to those obtained by linear regression. From a statistical point of view, the proposed methodologies are more suitable for the analysis of censored data, since in their formulation, they consider the presence of unobserved phenotypes. For the real data, the use of the Random Survival Forest (RSF) and Gradient Boosting Machine and Supervised Principal Component Analysis methods in genomic selection was considered, these being compared to the Ridge Bayesian Regression (BRR) method. The models were compared via 7-fold cross-validation, by Area Under the Curve, Brier Score, Spearman correlation, and by the proportion of selected individuals, and also by the location of relevant SNPs or link

groups. The results showed that the RSF and BRR models showed statistically equal values of predictive ability. The rank of the Top-40 SNPs obtained by RSF showed a greater intersection with the ranks obtained by the BRR methods and Cox mixed model. The greatest Spearman correlation between the GEBVs estimated via BRR and the survival probabilities, was obtained by RSF. The use of subsets of SNPs selected by the proposed methods did not result in significant differences in the predictive ability of the Cox mixed model. Finally, it is noted that the RSF method has a performance similar to that of the BRR, being possible its application in studies genomics.

Keywords: Genomic wide selection. Genomic breeding values. Censored data. Mixed Cox model. Machine learning.

SUMÁRIO

INTRODUÇÃO GERAL

A herança de características quantitativas pode ser de natureza simples, controlada por poucos genes (QTL - *quantitative trait loci*) com grandes efeitos, ou complexa, quando controlada por muitos genes de efeitos de pequenos à moderados distribuídos pelo genoma (BHAT et al., 2016). A maioria das características fenotípicas de interesse econômico no melhoramento animal são de natureza complexa, como produção de leite e porcentagem de gordura ou proteína no leite de bovinos (HAYES et al., 2010), desempenho, reprodução e parâmetros de longevidade em suínos (SAMORÈ; FONTANESI, 2016), e resistência à doenças em peixes, que é caracterizada pelo tempo de sobrevivência e pela presença de observações censuradas (PALAIOKOSTAS et al., 2016; BARRÍA et al., 2018). A seleção eficiente de indivíduos com desempenho superior para uma dada característica fenotípica, é um dos principais objetivos em programas de melhoramento genético, e pode ser realizada pela predição de valores genéticos dos indivíduos candidatos à seleção.

Uma das primeiras metodologias propostas para a seleção de indivíduos utilizando informações de marcadores moleculares, foi a seleção assistida por marcadores (MAS). A MAS se baseia na utilização de marcadores moleculares associados a QTLs com efeitos relativamente grandes, para a predição de valores genéticos dos indivíduos. Devido ao baixo número de marcadores com associação significativa, esta abordagem geralmente explica uma pequena proporção da variação genética da característica, o que faz com que ela tenha uma capacidade de predição de valores genéticos limitada (GODDARD; HAYES, 2009).

O desenvolvimento das tecnologias de informática, genotipagem e de sequenciamento de DNA, ocorrido nas últimas décadas, possibilitou a produção e a utilização de marcadores moleculares em larga escala, com uma redução considerável no custo de genotipagem. Visando superar as limitações apresentadas pela MAS, Meuwissen, Hayes e Goddard (2001) propuseram uma abordagem chamada seleção genômica ampla, que se baseia na utilização de painéis de marcadores de alta densidade, espalhados pelo genoma. Com tantos marcadores, é provável que a maior parte dos QTLs estejam em desequilíbrio de ligação com pelo menos alguns dos marcadores. Os dados genotípicos são utilizados juntamente com a informação fenotípica, para estimação simultânea dos efeitos de substituição alélica de centenas de milhares de marcadores genéticos, em uma amostra dos animais, denominada população de referência (ou treinamento). Em seguida, baseando-se apenas em informações genotípicas e nos efeitos estimados, são derivados modelos preditivos para a estimação dos valores genéticos genômicos (GEBVs), em

uma segunda amostra, chamada população de validação ou candidatos a seleção (GODDARD; HAYES, 2009; RESENDE et al., 2012).

A alta dimensionalidade dos dados genômicos impõe desafios estatísticos e de bioinformática aos pesquisadores para análise e manipulação dos dados, geralmente, o número de indivíduos é muito menor que o número de variáveis. Outro desafio, é que o desequilíbrio de ligação entre os marcadores gera uma estrutura de dados altamente correlacionada, o que viola o pressuposto da independência dos modelos clássicos de regressão, tornando inviável a sua aplicação para estimação dos efeitos dos marcadores (CHEN; ISHWARAN, 2012; RESENDE et al., 2012).

A literatura apresenta uma ampla variedade de métodos capazes de lidar com estas limitações. De acordo com Resende et al. (2012), o melhor método para predição genômica, é aquele capaz de acomodar uma arquitetura genética complexa, constituída por genes com efeitos variando de pequeno a grande, e deve também realizar uma regularização no processo de estimação dos efeitos, e/ou seleção de marcadores relacionados à característica. Os autores classificam os métodos em três classes: regressão explícita, na qual se encontram os métodos paramétricos de estimação penalizada: RR-BLUP (*Ridge Regression - Best Linear Unbiased Prediction*), LASSO (*Least Absolute Shrinkage and Selection Operator*) e Rede elástica, e de estimação bayesiana: BayesA e BayesB (MEUWISSEN; HAYES; GODDARD, 2001), entre outros; regressão implícita, com destaque para os métodos semiparamétricos de redes neurais artificiais (OKUT et al., 2011; PÉREZ-RODRÍGUEZ et al., 2012) e *Reproducing Kernel Hilbert Space* - RKHS (GIANOLA; FERNANDO; STELLA, 2006; GIANOLA; van KAAM, 2008); e modelos de regressão *Kernel* não paramétrica via modelos aditivos generalizados (GIANOLA; FERNANDO; STELLA, 2006); e por fim, tem-se os métodos de regressão com redução de dimensionalidade, componentes independentes, quadrados mínimos parciais e componentes principais (SOLBERG et al., 2009; RESENDE; SILVA; AZEVEDO, 2014).

Na classe dos métodos não paramétricos, se enquadram os métodos de aprendizado de máquina, *Random Forest* - RF (BREIMAN, 2001), *Gradient Boosting Machine* - GBM (FRIEDMAN, 2001, 2002) e Máquina de Vetores Suporte - SVM (VAPNIK, 2000). Estes métodos constituem uma abordagem alternativa para superar os problemas advindos da utilização de dados genômicos em estudos de predição e de classificação. Objetivando refletir a arquitetura genética de um caractere quantitativo, os modelos paramétricos de regressão bayesiana realizam o processo de regularização através da distribuição *a priori* assumida para efeito dos marcadores (GONZÁLEZ-RECIO; ROSA; GIANOLA, 2014). Já os métodos de

aprendizado de máquina apresentam uma grande flexibilidade, pois não exigem nenhum conhecimento *a priori* a respeito das relações existentes entre as variáveis, e fazem poucas suposições quanto aos dados, o que os tornam atrativos também para respostas que não sejam normalmente distribuídas (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007; GIANOLA et al., 2011, GONZÁLEZ-RECIO; ROSA; GIANOLA, 2014).

Os métodos citados anteriormente têm sido amplamente utilizados em estudos de predição genômica para características fenotípicas de natureza contínua, binária ou ordinal em animais e plantas. No contexto de dados censurados, tem-se o registro da utilização da classe de modelos bayesianos (*Bayesian Generalized Linear Regression* - BGLR), proposto por (PÉREZ; CAMPOS, 2014), que modela fenótipos censurados como sendo amostrados de uma distribuição normal truncada. Esta classe engloba como casos particulares os modelos paramétricos bayesianos citados anteriormente, o método RKHS e as versões bayesianas dos métodos RR-BLUPP e LASSO, dentre outros. Outro modelo citado na literatura para dados censurados é o modelo misto de Cox, proposto e avaliado por Santos et al. (2015), para a predição genômica da característica tempo ao abate de suínos.

O modelo de regressão RKHS é um dos métodos de aprendizado de máquina mais utilizados no melhoramento animal, tal fato se deve a sua abordagem flexível, baseada em uma matriz Kernel geral, criada por meio de uma função de suavização, e que representa a estrutura de covariância entre os indivíduos (GONZÁLEZ-RECIO; ROSA; GIANOLA, 2014). A matriz Kernel (**K**), pode ser substituída por qualquer matriz que seja positiva e definida, como por exemplo, as matrizes de parentesco genômico (**G**) e a matriz de parentesco tradicional baseada em pedigree (**A**). Campos, Gianola e Rosa (2009) mostraram que os modelos genéticos implementados via RKHS, constituem uma abordagem geral para a incorporação da informação de marcadores e de genealogia, e que podem ser aplicados a qualquer modelo genético. Assim como o modelo G-BLUP (VanRaden, 2008), o modelo misto de Cox também utiliza a informação dos marcadores por meio da matriz de parentesco genômico. De acordo com (HABIER; FERNANDO; DEKKERS, 2007) o uso da matriz de parentesco genômico contorna o problema da alta dimensionalidade, e permite explorar o desequilíbrio de ligação entre QTLs e marcadores.

A avaliação da qualidade das predições genômicas de diferentes modelos, geralmente é realizada via validação cruzada. No caso de características fenotípicas contínuas e normalmente distribuídas, geralmente, é utilizada a acurácia, que se baseia na correlação de Pearson entre os valores genéticos genômicos estimados e os valores fenotípicos, e na herdabilidade da

característica. Outra medida utilizada é o viés, estimado como coeficiente angular da regressão linear dos fenótipos nos valores genéticos genômicos estimados (RESENDE et al., 2012). González-Recio e Forni (2011), para predição genômica de características binárias, utilizaram as medidas AUC e a correlação Phi para dados reais, e a correlação de Pearson e a AUC para dados simulados. Na predição genômica de características censuradas, também é utilizada a correlação de Pearson, considerando os tempos até a ocorrência do evento de interesse, como sendo contínuos e normalmente distribuídos.

Como visto, a comparação de modelos em seleção genômica ampla, é fundamentada na correlação de Pearson e na estimação do coeficiente angular da regressão linear simples. Estas medidas não levam em conta a presença de observações censuradas, considerando a resposta como sendo completa para todos os indivíduos, ou seja, supõem que o fenótipo foi registrado para todas as observações. É esperado que metodologias adequadas para avaliação de dados censurados forneçam resultados mais corretos.

Recentemente, Li et al. (2018) propuseram uma medida de correlação própria para a modelagem de dados na presença de censura à esquerda, à direita, intervalar ou com observações perdidas. Esta medida é baseada na verossimilhança perfilada, sendo referida como correlação de Pearson para dados censurados. A correlação maximal (GEBELEIN, 1941; RÉNYI, 1959; BREIMAN; FRIEDMAN, 1985), embora não seja uma medida própria para dados censurados, apresenta uma grande flexibilidade para a modelagem de relações não lineares. O que a torna uma medida interessante em situações em que a correlação entre as duas variáveis é baixa. A regressão Tobit proposta Tobin (1958), é uma alternativa à regressão linear no contexto de dados censurados, e se baseia na utilização de uma variável auxiliar, chamada variável latente, para expressar a variável observada em termos do limiar de censura.

A utilização de métodos de análise de sobrevivência em estudos de seleção genômica ampla foi inicialmente proposta por Santos et al. (2015), ao utilizaram o modelo misto de Cox para a predição de valores genéticos genômicos para a característica tempo ao abate de suínos. Embora a literatura apresente diversas aplicações dos modelos *Random Forest*, *Gradient Boosting Machine* e Análise de Componentes Principais para a seleção genômica de características contínuas, na seleção genômica com dados censurados, estes métodos ainda não foram explorados. Neste contexto, os métodos citados se fundamentam em funções específicas da área de análise de sobrevivência, tais como as funções de sobrevivência, de taxa de falha e de taxa de falha acumulada, o que faz com sejam classificados como métodos de análise de sobrevivência.

Neste sentido, este trabalho tem como objetivos:

i. Avaliar a aplicabilidade das medidas: correlação de Pearson para dados censurados e correlação maximal, como medidas para o cálculo da acurácia de modelos de predição genômica com dados censurados.
ii. Considerar a utilização do coeficiente angular estimado via regressão Tobit como medida para o viés de modelos de predição genômica com dados censurados.
iii. Propor e avaliar a utilização dos métodos *Random Survival Forest* (RSF), *Gradient Boosting Machine* (GBM) e Análise de Componentes Principais Supervisionados (SPCA) em estudos de seleção genômica ampla com dados censurados.
iv. Avaliar a utilização dos subconjuntos de SNPs selecionados pelos métodos RSF, GBM e SPCA no modelo misto de Cox para a predição de valores genéticos genômicos.

Este trabalho está dividido em três capítulos. No capítulo 1 é apresentado o referencial teórico, com os principais conceitos de análise de sobrevivência e com a descrição dos métodos de seleção genômica ampla e de análise de sobrevivência utilizados. No capítulo 2 são propostas e avaliadas medidas alternativas para a avaliação da acurácia e viés dos modelos misto de Cox e normal truncado em estudos de predição genômica com dados censurados simulados. No capítulo 3 é considerada a utilização dos métodos de aprendizado de máquina: *Random Survival Forest* – RSF e *Gradient Boosting Machine* – GBM, e de redução de dimensionalidade: Análise de Componentes Principais Supervisionados - SPCA, em estudos de seleção genômica com dados censurados reais. Por fim, são apresentadas as conclusões gerais do trabalho.

CAPÍTULO I

REFERENCIAL TEÓRICO

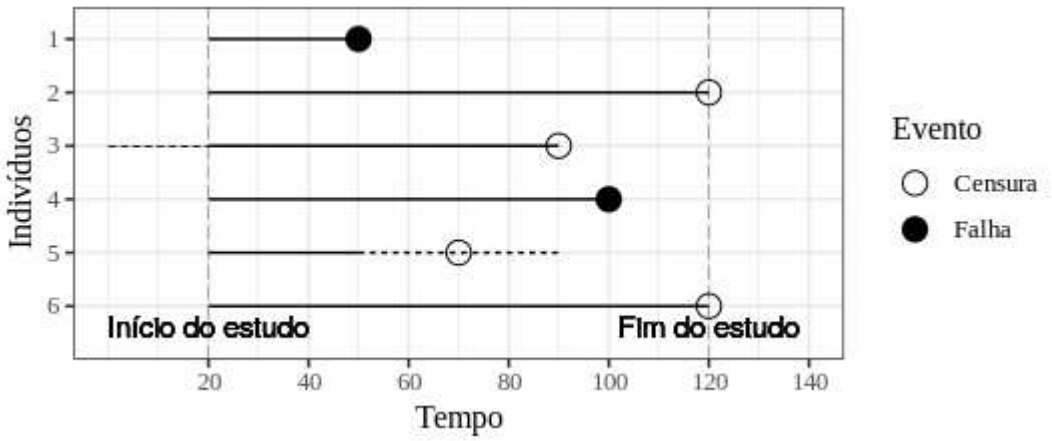
1 ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência é uma área da estatística que possui ferramentas estatísticas e computacionais focadas na modelagem do tempo até a ocorrência de um evento de interesse, denominado tempo de falha ou de sobrevivência. O evento pode ser a morte de um indivíduo, a falha de equipamentos elétricos, divórcio, término ou ingresso em um curso de graduação, registro de ocorrência de uma doença, falência de empresas, entre outros. Esta ampla possibilidade de definição dos eventos, faz com que a análise de sobrevivência seja aplicada em diversas áreas da ciência, como, medicina, confiabilidade, sociologia e economia. Os dados de sobrevivência são caracterizados também pela ocorrência da censura, que se caracteriza quando o evento de interesse não é observado durante o tempo de realização do experimento, ou seja, não se tem o registro exato do tempo de falha, apenas uma informação parcial (COLOSIMO; GIOLO, 2006; EMMERT-STREIB; DEHMER, 2019).

Colosimo e Giolo (2006) definem três tipos de censura: a censura à esquerda, a censura à direita e a censura intervalar. Segundo os autores, a censura à direita é a mais comum em estudos de sobrevivência, e ocorre quando o tempo de sobrevivência é maior que o tempo observado. Já na censura à esquerda, o tempo de sobrevivência é menor do que o tempo observado. E por fim, na censura intervalar, não se sabe ao certo o momento em que a falha ocorreu, só se sabe que o tempo de sobrevivência está contido em um certo intervalo.

A Figura 1 ilustra por meio de um exemplo fictício o significado de cada um dos três tipos de censura. Os sujeitos identificados como 1 e 4 experimentaram o evento de interesse, durante o período de realização do experimento. Por outro lado, os sujeitos 2 e 6, não experimentaram o evento de interesse no tempo de realização do experimento, e podem falhar em algum tempo futuro ou nunca vir a falhar, caracterizando a censura à direita. O sujeito 3 representa o caso da censura à esquerda, e para o indivíduo 5 não se sabe o tempo exato de falha, constituindo um caso de censura intervalar.

Figura 1: Ilustração fictícia dos tipos de censura à esquerda, à direita e intervalar.



Fonte: Autor.

Dados de sobrevivência são comumente representados por $\{(t_i, \delta_i, \mathbf{x}_i);\ i = 1, \ldots, n\}$, em que: t_i é o tempo de falha, δ_i é uma variável indicadora de censura, que assume valor 1, para o i-ésimo tempo de falha, e 0, para o i-ésimo tempo de censura, e $\mathbf{x}_i$ representa um vetor de covariáveis associado a cada um dos n indivíduos. Na censura à direita, caso considerado neste trabalho, tem-se que: $t = \min(T, C)$ e que $\delta = 1$ se $T \leq C$ e $\delta = 0$ se $T > C$, em que T e C são variáveis aleatóricas, representando, respectivamente, os tempos de falha e de censura, sendo o tempo de censura independente do tempo de falha (COLOSIMO; GIOLO, 2006).

1.1 Função de sobrevivência e funções relacionadas

Seja T uma variável aleatória contínua e não negativa que representa o tempo de sobrevivência de um indivíduo em uma dada população, com função densidade de probabilidade $f(t)$ definida no intervalo $[0, \infty)$. A função de distribuição acumulada de T é dada por:

$$F(t) = P(T \leq t) = \int_0^t f(u)du,$$

e descreve a probabilidade de que o evento ocorra até um certo tempo t. A função de sobrevivência, como o próprio nome sugere, descreve a sobrevivência dos indivíduos em função do tempo (t). Ela representa a probabilidade de que um indivíduo sobreviva a um tempo t específico. Esta função é denotada por:

$$S(t) = 1 - F(t) = P(T > t),$$

em que $S(t)$ é uma função monotônica decrescente, contínua à direita, tal que a probabilidade de um indivíduo sobreviver ao tempo zero é 1, ou seja, em $t = 0$, $S(t) = 1$, e a probabilidade de um indivíduo sobreviver em um tempo muito grande é zero, isto é, $\lim_{t\to\infty} S(t) = 0$ (COLOSIMO; GIOLO, 2006).

A função de sobrevivência, pode também ser definida em termos da função taxa de falha ($h(t)$), que representa a taxa de falha instantânea no tempo t condicionada à sobrevivência até o tempo t. $h(t)$ é definida por:

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)},$$

a função $h(t)$ apresenta as seguintes propriedades: é sempre maior ou igual a zero para todo tempo t, não possui limite superior e pode assumir diversas formas (COLOSIMO; GIOLO, 2006).

A função de taxa de falha acumulada $H(t)$ descreve a taxa de falha acumulada para um indivíduo até o tempo t, sendo denotada por (COLOSIMO; GIOLO, 2006):

$$H(t) = \int_0^t h(u)du.$$

1.2 Modelo de riscos proporcionais de Cox

Segundo Lawless (2002), um dos caminhos para representar a heterogeneidade de uma população, é por meio do ajuste de modelos de regressão a um conjunto de variáveis resposta e preditoras. Em dados de sobrevivência, o objetivo é entender e quantificar a relação entre o tempo de falha e as covariáveis, e saber se esta relação é de fato significativa. Neste sentido, o modelo de riscos proporcionais de Cox (ou modelo de Cox), proposto por Cox (1972), é um modelo de regressão semiparamétrico que possibilita modelar a taxa de falha em função de um conjunto de covariáveis. O modelo é expresso pela equação:

$$h(t \mid \mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}),$$

em que: $\mathbf{x}$ e $\boldsymbol{\beta}$ são vetores, respectivamente, de covariáveis e de coeficientes de regressão.

O modelo é composto por um componente não paramétrico, uma função não negativa do tempo $h_0(t)$, não especificada e independente das covariáveis, chamada função base de taxa de falha, obtida quando $\mathbf{x} = \mathbf{0}$, e um componente paramétrico na forma exponencial para os

efeitos dos preditores na taxa de falha, dado por: $r(\mathbf{x};\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}'\mathbf{x})$ (LAWLESS, 2002; COLOSIMO; GIOLO, 2006). Embora o modelo não faça suposições quanto a especificação da função base de taxa de falha, é assumida independência temporal entre as covariáveis, linearidade nas covariáveis, aditividade e proporcionalidade das taxas de falha (EMMERT-STREIB; DEHMER, 2019).

A utilização da função taxa de falha, ao invés da função de sobrevivência ou de densidade de probabilidade no modelo de Cox, se justifica pelo fato da primeira função ser mais sensível a mudanças em um curto período de tempo. Isto faz com que a função taxa de falha seja mais adequada para a definição de um modelo de regressão subjacente (KARIM; ISLAM, 2019). O intercepto β_0 comum em modelos de regressão lineares, não aparece na formulação do modelo de Cox, pois é incluído na função base de taxa de falha. O último pressuposto, taxas de falhas proporcionais é o responsável pelo nome do modelo, deste, segue que a razão das taxas de falha de dois quaisquer indivíduos, digamos m e n, é constante no tempo, e dada por:

$$\frac{h_m(t|\mathbf{x}_m)}{h_n(t|\mathbf{x}_n)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_m)}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_n)} = \exp\big(\boldsymbol{\beta}'(\mathbf{x}_m - \mathbf{x}_n)\big) = k.$$

Observa-se que esta razão é independente do tempo, sendo dependente apenas dos valores das covariáveis. Supondo que $k = 2{,}5$ na equação anterior, segue que o indivíduo m tem duas vezes e meia mais chances de falhar do que o indivíduo n.

Os dados de sobrevivência são utilizados no modelo de Cox para estimação do vetor de parâmetros $\boldsymbol{\beta}$, via método de máxima verossimilhança parcial. Seja $R_i = \{l: T_l \geq T_i\}$ o conjunto de risco que contém os indivíduos em risco no tempo T_i, o estimador de máxima verossimilhança parcial é obtido pela maximização da função de verossimilhança parcial proposta por Cox (1972), dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{l\,\epsilon R_i}\exp(\boldsymbol{\beta}'\mathbf{x}_l)}\right)^{\delta_i}.$$

O estimador de máxima verossimilhança parcial de $\boldsymbol{\beta}$ é consistente, apenas quando a censura é independente, e o modelo é corretamente especificado. No caso de censura dependente, o estimador fornece estimativas viesadas (FLEMING; HARRINGTON, 1991; EMURA; CHEN, 2016). Aplicando-se a função logarítmica em $L(\boldsymbol{\beta})$, tem-se que:

$$l(\boldsymbol{\beta}) = \log\big(L(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{\beta}'\mathbf{x}_i - \log\left(\sum_{l \in R_i} \exp \boldsymbol{\beta}'\mathbf{x}_l \right) \right).$$

Aplicando-se a primeira derivada em $l(\boldsymbol{\beta})$, obtém-se a função escore da forma:

$$U(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i - \frac{\sum_{l \in R_i} \mathbf{x}_l \exp(\boldsymbol{\beta}'\mathbf{x}_l)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right].$$

A matriz Hessiana é obtida pelo cálculo da derivada de segunda ordem de $l(\boldsymbol{\beta})$,

$$H(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \delta_i \left[\frac{\sum_{l \in R_i} \mathbf{x}_l \mathbf{x}'_l \exp(\boldsymbol{\beta}'\mathbf{x}_l)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} - \frac{\sum_{l \in R_i} \mathbf{x}_l \exp(\boldsymbol{\beta}'\mathbf{x}_l)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \left(\frac{\sum_{l \in R_i} \mathbf{x}_l \exp(\boldsymbol{\beta}'\mathbf{x}_l)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right) \right].$$

Para a obtenção de intervalos de confiança e para a realização de testes de hipóteses para os coeficientes de regressão do modelo de Cox, é necessária a utilização das propriedades assintóticas dos estimadores de máxima verossimilhança parcial (FLEMING; HARRINGTON, 1991; COLOSIMO; GIOLO, 2006). Estas propriedades foram demostradas de forma geral por Andersen e Gill (1982), assim, as estatísticas de Wald, da razão de verossimilhança e escore, podem ser utilizadas em inferências via verossimilhança parcial (COLOSIMO; GIOLO, 2006).

2 MODELOS ESTATÍSTICOS PARA PREDIÇÃO GENÔMICA

A seguir são descritos os modelos utilizados neste estudo. Inicialmente são apresentados os modelos lineares G-BLUP e RR-BLUP, sobre as perspectivas frequentista e bayesiana, para respostas contínuas e censuradas. Em seguida é descrito o modelo misto de Cox, que apresenta diversas aplicações no melhoramento animal. E por fim, são discutidos aspectos gerais dos métodos de aprendizado, especificamente, os métodos *Random Survival Forest*, *Gradient Boosting Machine* e Análise de Componentes Principais Supervisionados.

2.1 Modelos G-BLUP e RR-BLUP

O modelo G-BLUP foi introduzido em predições genômicas por Habier, Fernando e Dekkers (2007) e VanRaden (2008), e consiste na substituição da matriz de parentesco tradicional **A** entre os indivíduos, obtida com base na informação de pedigree, pela matriz de parentesco genômica **G** entre os indivíduos, baseada na informação dos marcadores

(RESENDE, 2012). Os valores genéticos genômicos (GBVs) são estimados com base no seguinte modelo misto:

$$\mathbf{y} = \mathbf{1\mu} + \mathbf{Za} + \mathbf{e}, \quad (1)$$

em que: $\mathbf{y}$ é um vetor de fenótipos, $\boldsymbol{\mu}$ é a média geral, $\mathbf{a}$ é o vetor de efeito genético aditivo aleatório, $\mathbf{Z}$ é a matriz de incidência do efeito aleatório, $\mathbf{1}$ é um vetor de uns de incidência para o intercepto $\boldsymbol{\mu}$ e $\mathbf{e}$ é o vetor de resíduos aleatório. Na presença de efeitos sistemáticos, o intercepto é substituído por $\mathbf{X\beta}$, para inclusão de outros efeitos fixos. É assumido para os efeitos genéticos aditivos ($\mathbf{a}$) uma distribuição normal $\mathbf{a} \sim N(0, \mathbf{G}\sigma_a^2)$, sendo $\mathbf{G}$ a matriz de parentesco genômica e σ_a^2 a variância genética aditiva. Para o vetor de resíduos $\mathbf{e}$ é assumida uma distribuição normal $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, sendo $\mathbf{I}$ a matriz identidade e σ_e^2 a variância residual.

A matriz de parentesco genômico ($\mathbf{G}$) segundo a formulação proposta por VanRaden (2008) é dada por:

$$\mathbf{G} = \frac{\mathbf{WW'}}{k},$$

em que: k é um parâmetro de escala dado por $k = 2\sum_{j=1}^{m} p_j q_j$, e $\mathbf{W}$ é a matriz de ordem $n \times m$, cujos elementos são tais que: $w_{ij} = 0 - 2p_j = -2p_j$, $w_{ij} = 1 - 2p_j = q_j - p_j$ e $w_{ij} = 2 - 2p_j = 2q_j$, para os genótipos dos marcadores, mm, Mm e MM, respectivamente, sendo: p_j a frequência alélica de M no locos j, e $q_j = 1 - p_j$ é a frequência alélica de m no locos j (VANRADEN 2008; RESENDE et al., 2012).

Na equação 1, os efeitos fixos são estimados via melhor estimador linear não viesado (BLUE) e os efeitos aleatórios via melhor preditor linear não viesado (BLUP). O BLUP é não viesado ($E(\hat{\mathbf{a}}) = E(\mathbf{a})$), ou seja, o valor esperado das predições é igual ao valor esperado do parâmetro populacional, além disso, dentre as funções lineares do vetor de fenótipos **y,** o BLUP é a que apresenta menores estimativas para o erro quadrático médio (EQM). O estimador BLUE apresenta propriedades similares às apresentadas pelo BLUP (HOWARD; CARRIQUIRY; BEAVIS, 2014). Na prática, são utilizados os BLUE e BLUP empíricos, já que os componentes de variância, associados aos efeitos aleatórios e resíduos são desconhecidos, e precisam ser substituídos por suas estimativas, as quais podem ser obtidas pelo método da máxima verossimilhança restrita (REML) proposto por Patterson e Thompson (1971) (PIEPHO; MÖHRING; MELCHINGER, 2008).

Henderson (1953) mostrou que as estimativas BLUE e BLUP, respectivamente, para os

efeitos fixos e aleatórios, poderiam ser obtidas pela maximização da função de verossimilhança conjunta de $(\mathbf{y}, \mathbf{a})$, dada por:

$$L(\mathbf{y}, \mathbf{a}) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X\beta} - \mathbf{Za})'\mathbf{R^{-1}}(\mathbf{y} - \mathbf{X\beta} - \mathbf{Za})\right]$$
$$\times \frac{1}{(2\pi)^{p/2}|\mathbf{G}|^{1/2}} \exp\left[-\frac{1}{2}\mathbf{a}'\mathbf{G^{-1}}\mathbf{a}\right].$$

Maximizando $L(\mathbf{y}, \mathbf{a})$ através da derivação em relação aos efeitos fixos e aleatórios, e tomando-se as derivadas iguais a zero, obtêm-se o seguinte sistema de equações de modelo misto (EMM):

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}, \quad (2)$$

em que: $\mathbf{R} = \text{Var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$ e $\mathbf{G} = \text{Var}(\mathbf{a})$.

A solução do sistema EMM fornece as estimativas de $\boldsymbol{\beta}$ e $\mathbf{a}$. Segundo Howard, Carriquiry e Beavis (2014), as estruturas de variâncias e covariâncias **R** e **G**, são desconhecidas e precisam ser estimadas juntamente com $\boldsymbol{\beta}$ e $\mathbf{a}$. O processo de estimação dos componentes de variância, pelo chamado método da máxima verossimilhança restrita (REML), envolve a maximização da função de verossimilhança restrita associada a um conjunto específico de combinações lineares dos dados, que depende apenas dos componentes de variâncias **R** e **G**. Assim, o estimador BLUE dos efeitos fixos é $\widehat{\boldsymbol{\beta}} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y}$ e o estimador BLUP dos efeitos aleatórios é dado por: $\hat{\mathbf{a}} = \mathbf{GZ'V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$, com: $\mathbf{V} = \mathbf{ZGZ'} + \mathbf{R}$ (ROBINSON, 1991; RESENDE et al., 2012).

A diferença básica entre os procedimentos de estimação bayesiana e de máxima verossimilhança (ML), está no fato de que na bayesiana, é maximizada a distribuição *a posteriori* do parâmetro, enquanto na ML, maximiza-se a função de verossimilhança. A distribuição *a posteriori* ou distribuição condicional do parâmetro dadas as observações, é proporcional ao produto da função de verossimilhança, que representa toda informação contida nos dados, pela distribuição *a priori* assumida para o parâmetro (RESENDE et al., 2012).

O modelo G-BLUP pode também ser implementado sobre a perspectiva bayesiana, na qual, os parâmetros desconhecidos associados aos efeitos fixos e aleatórios, são tidos como variáveis aleatórias no modelo padrão, sem distinção de natureza. Para o intercepto μ, é assumida uma distribuição *a priori* constante, $p(\mu) \propto \text{constante}$. Já para os componentes de variância σ_a^2 e σ_e^2, é assumida uma distribuição qui-quadrado escalonada invertida com $df > 0$

graus de liberdade e parâmetro de escala $S > 0$. Para o vetor de efeitos genéticos é assumida uma distribuição normal, com média zero e matriz de variâncias e covariâncias $\mathbf{G}\sigma_a^2$. Sejam $H = \{df_a, S_a, df_e, S_e\}$ um conjunto de hiperparâmetros e $\boldsymbol{\theta} = \{\mathbf{a}, \boldsymbol{\mu}, \sigma_a^2, \sigma_e^2\}$ o conjunto de parâmetros a serem estimados. Os parâmetros de interesse podem ser estimados via maximização da distribuição *a posteriori* dos parâmetros desconhecidos do modelo, dado os fenótipos e os hiperparâmetros, pela expressão (PÉREZ; CAMPOS, 2014):

$$p(\boldsymbol{\theta} \mid \mathbf{y}, H) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid H)$$
$$\propto \prod_{i=1}^{n} N(y_i \mid \mu + a_i, \sigma_e^2) \times N(\mathbf{a} \mid 0, \mathbf{G}\sigma_a^2) \times \prod_{j \in \{a,e\}} \chi^{-2}(\sigma_j^2 \mid df_j, S_j).$$

Pérez e Campos (2014) propuseram uma classe de modelos Bayesianos, chamada *Bayesian Generalized Linear Regression* (BGLR), que amplia os modelos de predição genômica, inicialmente utilizados para a predição de fenótipos contínuos e normalmente distribuídos, para fenótipos binários, ordinais e censurados. Variáveis respostas censuradas são tratadas nos modelos bayesianos como parte de um problema de dados perdidos. Nestes modelos, é possível lidar com a censura à esquerda, à direita ou censura intervalar. Segundo os autores, a censura pode ser descrita pela tripla $\{l_i, y_i, s_i\}$, sujeita a condição: $l_i < y_i < s_i$, sendo: y_i o valor fenotípico observado para o i-ésimo indivíduo e, l_i e s_i, respectivamente, os limites inferior e superior, para o valor fenotípico. Em estudos de sobrevivência, y_i representa o tempo até a ocorrência do evento, registrado apenas para indivíduos que não foram censurados, no caso de indivíduos censurados à direita, a tripla é configurada como $\{l_i, NA, Inf\}$, sendo os valores censurados de y_i amostrados de uma distribuição normal truncada. A fórmula da densidade da distribuição condicional dos dados, considerando a presença de observações censuradas é:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) \propto (2\pi\sigma_e^2)^{-n_o/2} \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n_o}(y_{obs_i} - \mathbf{x'}_i\boldsymbol{\beta} - \mathbf{z'}_i\mathbf{a})^2\right\}$$
$$\times \prod_{i=n_o+1}^{n}\left[1 - \Phi\left(\frac{l_i - \mathbf{x'}_i\boldsymbol{\beta} - \mathbf{z'}_i\mathbf{a}}{\sigma_e}\right)\right],$$

em que: y_{obs_i} é o valor fenotípico observado, n é o número total de observações, n_o é o número de observações não censuradas, Φ é a função de distribuição acumulada da normal padrão e l_i é o ponto em que a distribuição é truncada, ou seja, o tempo em que o experimento é encerrado.

Sorensen, Gianola e Korsgaard (1998) sugerem a utilização do algoritmo de aumento de dados (TANNER; WONG, 1987), como uma forma de manipular a densidade de distribuição condicional, visando facilitar a obtenção da densidade conjunta *a posteriori*. Deste modo, é possível obter uma distribuição condicional completa com densidade conhecida e fácil de ser amostrada, facilitando a utilização do amostrador de Gibbs (WANG, 1998; van DYK; MENG, 2001). A distribuição condicional dos dados censurados e não censurados, obtida após o aumento de dados, é dada por:

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\theta}) \propto (2\pi\sigma_e^2)^{-n_o/2} \\ \times \exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{i=1}^{n_o}\left(y_{obs_i} - \mathbf{x'}_i\boldsymbol{\beta} - \mathbf{z'}_i\mathbf{a}\right)^2\right.\right. \\ \left.\left. + \sum_{i=n_o+1}^{n}\left(y_{cen_i} - \mathbf{x'}_i\boldsymbol{\beta} - \mathbf{z'}_i\mathbf{a}\right)^2\right]\right\}, \end{aligned} \tag{3}$$

em que: y_{cen_i} é o tempo de falha não observado para o indivíduo censurado i, com: $y_{cen_i} \geq l_i$, assim, é imposto à variável y_{cen_i}, que seus valores sejam maiores ou iguais do que o ponto de truncamento.

Whittaker, Thompson e Denham (2000) e Meuwissen, Hayes e Goddard (2001) propuseram a utilização do modelo BLUP em estudos genômicos, assumindo que os efeitos dos marcadores são aleatórios. No modelo RR-BLUP, os efeitos dos marcadores podem ser estimados de acordo com a equação mista apresentada para o modelo G-BLUP, com algumas mudanças de notação, a matriz **Z** e o vetor **a,** são substituídos, respectivamente, pela matriz de incidência dos marcadores (**W**) e pelo vetor **m,** que é um vetor aleatório de efeito dos marcadores. É assumido uma distribuição normal para o efeito dos marcadores $\mathbf{m} \sim N(0, \mathbf{I}\sigma_m^2)$ e para o resíduo $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, em que: σ_m^2 é a variância comum para cada efeito de marcador, e a matriz **I** e σ_e^2 como definidos anteriormente. As estimativas dos parâmetros $\boldsymbol{\beta}$ e **m** são obtidas pela expressão:

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{m}}\end{bmatrix} = \begin{bmatrix}\mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{I}\frac{\sigma_e^2}{\sigma_a^2/n_Q}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X'y} \\ \mathbf{W'y}\end{bmatrix},$$

em que: σ_a^2 se refere a variância aditiva total do fenótipo, e $n_Q = \sum_{j=1}^{m} 2p_j(1-p_j)$ com m representando o número de marcadores. Os componentes de variância são desconhecidos e estimados via máxima verossimilhança restrita (REML). O parâmetro de regularização ridge

dos efeitos dos marcadores é representado por $\lambda = \sigma_e^2/(\sigma_a^2/n_Q)$. Este método realiza um encurtamento homogêneo dos efeitos dos marcadores, a qual pode não ser a melhor alternativa para predição genômica, caso alguns marcadores estejam ligados a QTLs, e outros se localizem em regiões que não possuem QTLs (RESENDE et al., 2012; CAMPOS et al., 2013).

Sobre a perspectiva bayesiana, o modelo RR-BLUP pode ser implementado assumindo que os efeitos dos marcadores seguem distribuições normais *a priori* independentes e com variâncias homogêneas. Esta *priori*, induz um encurtamento de mesma extensão para todos os efeitos estimados, semelhante ao que é realizado pela regressão ridge (HOERL; KENNARD, 1970). De modo análogo ao G-BLUP Bayesiano, é assumida uma distribuição qui-quadrado escalonada invertida para a variância dos marcadores e para a variância residual. Substituindo a letra "**a**" por "**m**" nos conjuntos H e $\boldsymbol{\theta}$ apresentados anteriormente, segue que os efeitos dos marcadores, o intercepto e os componentes de variâncias podem ser estimados via maximização da distribuição *a posteriori* dada por (PÉREZ; CAMPOS, 2014):

$$p(\boldsymbol{\theta} \mid \mathbf{y}, H) \propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid H)$$

$$\propto \prod_{i=1}^{n} N\left(y_i \mid \mu + \sum_{j=1}^{m} w_{ij} m_j, \sigma_e^2\right) \times \prod_{j=1}^{m} N\left(m_j \mid 0, \sigma_m^2\right)$$

$$\times \prod_{k \in \{e,m\}} \chi^{-2}\left(\sigma_k^2 \mid df_k, S_k\right).$$

O modelo Regressão Ridge Bayesiana (BRR), quando aplicado a dados censurados, pode ser implementado de modo semelhante ao modelo G-BLUP, considerando a distribuição condicional dos dados censurados e não censurados, obtida após o aumento de dados apresentada na equação (3).

2.2 Modelo misto de Cox

Um dos modelos mais citados e utilizados em análise de sobrevivência, é o modelo de riscos proporcionais de Cox (COX, 1972). Adicionando a este modelo efeitos aleatórios, tem-se o modelo misto de Cox ou modelo de fragilidade. De acordo com Giolo e Demétrio (2011), neste modelo, os termos associados aos efeitos fixos e a fragilidade, atuam multiplicativamente na função base de taxa de falha. Os autores complementam que o termo de fragilidade captura fatores de risco não observados, inerentes à heterogeneidade dos indivíduos, já o termo associado às covariáveis, captura a heterogeneidade observada. O modelo de fragilidade de Cox

pode ser descrito segundo Ripatti e Palmgren (2000) e Therneau, Grambsch e Pankratz (2003) pela equação:

$$\lambda(t) = \lambda_0(t)\exp\{\mathbf{X\beta} + \mathbf{Za}\}, \quad (4)$$

em que: $\lambda_0(t)$ é a função base de taxa de falha; $\mathbf{X}$ e $\mathbf{Z}$ são as matrizes de incidência para os efeitos fixos e aleatórios, respectivamente; $\boldsymbol{\beta}$ é um vetor de efeitos fixos e $\mathbf{a}$ um vetor de efeitos aleatórios tal que: $\mathbf{a} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, sendo $\boldsymbol{\Sigma}$ uma matriz de variâncias e covariâncias.

Therneau (2020) comenta que é possível estender a função de verossimilhança parcial de Cox, para o caso em que efeitos aleatórios são incluídos no modelo. Assumindo que o mecanismo de censura é não informativo e independente, segue que a função de verossimilhança parcial para o modelo misto de Cox, é dada por (GIOLO; DEMÉTRIO, 2011; THERNEAU, 2020):

$$L = \int PL(\boldsymbol{\beta}, \mathbf{a}) \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a}\right\} d\mathbf{a}, \quad (5)$$

sendo: PL a função de verossimilhança parcial do modelo de riscos proporcionais de Cox. Como a integral da equação (5) não apresenta uma forma fechada, Ripatti e Palmgren (2000) sugeriram a utilização da aproximação de Laplace para superar os problemas advindos do cálculo integral multidimensional, e obter as estimativas para os parâmetros de interesse.

2.3 Aprendizado estatístico e de maquina

O grande volume e a complexidade dos dados utilizados em seleção genômica, tornou necessária a integração de várias áreas das ciências, como, Computação, Estatística, Aprendizado de máquina, Matemática, Bioinformática e Genética. Esta integração tem por objetivo uma predição mais acurada de valores não observados dos indivíduos em uma população de teste, fazendo uso de métodos de aprendizado estatístico ou de máquina (GONZÁLEZ-CAMACHO et al., 2018). A teoria de aprendizado de máquina (Machine Learning - ML) surgiu como uma subárea da ciência da computação, que utiliza algoritmos e modelos estatísticos para a realização de tarefas, baseando-se em padrões e em inferências. Os métodos de ML são capazes de aprender e melhorar seu desempenho com a experiência adquirida em uma base de dados de treinamento, utilizada para a construção de modelos matemáticos úteis para a realização de predições ou para a tomada de decisões (BISHOP, 2006; GONZÁLEZ-RECIO; FORNI, 2011; GONZÁLEZ-CAMACHO et al., 2018).

O aprendizado pode ser classificado como supervisionado, não supervisionado ou semi-supervisionado, de acordo com o problema a ser resolvido. No aprendizado supervisionado, o objetivo é utilizar os modelos preditivos obtidos no conjunto de treinamento, composto por variáveis de entrada e de saída, para a realização de predições futuras de uma dada saída, em um novo conjunto de entrada. Os modelos podem ainda ser classificados como de regressão, caso a saída seja contínua, ou de classificação, caso a saída seja categórica. No aprendizado não supervisionado, não há variável de saída, neste caso, o interesse é descrever as associações e padrões entre as variáveis de entrada, objetivando extrair inferências que possibilitem expor uma estrutura escondida dos dados. Por fim, o aprendizado semi-supervisionado, que se encontra entre o aprendizado supervisionado e o não supervisionado, combinando características das duas classes anteriores (YAO; ZHU; WEIGEL, 2006; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAYANTHI; MAHESH, 2018).

Estas técnicas de aprendizado são frequentemente utilizadas em estudos com dados genômicos. O aprendizado supervisionado é aplicado quando se dispõe da informação de marcadores moleculares e das medidas fenotípicas de todos os indivíduos na população de referência. Nesta classe, se enquadram os modelos de árvores de decisão, vizinho mais próximo (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007), *Gradient Boosting Machine*, *Random Forest*, máquina de vetores suporte, redes neurais artificiais (GONZÁLEZ-RECIO; FORNI, 2011; GONZÁLEZ-CAMACHO et al., 2018; PÉREZ-RODRÍGUEZ et al., 2012), dentre outros. Por outro lado, no aprendizado não supervisionado, não há informações fenotípicas para supervisionar o aprendizado dos modelos, um exemplo, é a utilização da análise de componentes principais para o agrupamento de animais com base na similaridade entre seus genótipos (YAO; ZHU; WEIGEL, 2006). Já no aprendizado semi-supervisionado, podemos citar o modelo *self-training* baseado no algoritmo de máquina de vetores suporte, proposto por Yao, Zhu e Weigel (2006) para a predição genômica para a característica ingestão alimentar residual, com animais fenotipados e não fenotipados, em gado leiteiro.

Segundo Pérez-Enciso (2017), o paradigma do modelo linear fundamenta a maioria dos métodos utilizados na teoria do melhoramento animal. Em seleção genômica, os modelos lineares, estão limitados a modelagem de efeitos aditivos, enquanto os modelos de ML são mais flexíveis, capazes de incluir outros tipos de efeitos, como por exemplo, a epistasia. Estes modelos são livres de suposições quanto à distribuição dos dados, não exigindo que os fenótipos sejam contínuos e normalmente distribuídos, o que é comum nos modelos lineares. Os modelos de aprendizado de máquina objetivam encontrar o algoritmo mais eficiente, que faça predições

satisfatórias quando aplicados a grandes bases de dados, a maioria deles não atribui muita importância a noção de modelo. Estes algoritmos têm sido utilizados para obter melhores predições para o mérito genético, para identificar genes relacionados a fenótipos, para estudar as interações existentes entre genes, e entre genes e ambiente (YANG et al., 2010; GONZÁLEZ-CAMACHO et al., 2018).

Nesta seção, são apresentados alguns dos mais conhecidos métodos de aprendizado, a saber, *Random Survival Forest* e *Gradient Boosting Machine*, e o método de redução de dimensionalidade, Análise de Componentes Principais Supervisionados.

2.3.1 *Random Survival Forest*

Introduzida por Breiman (2001), a *Random Forest* (RF) é uma metodologia de aprendizado de máquina supervisionada, não paramétrica e não linear, que não requer nenhuma suposição quanto a relação entre a variável resposta e o conjunto de variáveis preditoras, muito utilizada para a resolução de problemas de classificação e de regressão. É uma metodologia robusta, que se fundamenta na utilização de um conjunto de árvores visando a otimização da acurácia preditiva, pela estabilização das estimativas do modelo (ISHWARAN et al., 2008; EHRLINGER, 2016). Utilizando-se dos mesmos princípios descritos por Breiman (2001), Ishwaran et al. (2008) estenderam a RF para a análise de dados censurados à direita, o que originou o método *Random Survival Forest* (RSF). Esta base comum, faz com que a RSF e RF compartilhem várias propriedades que são fundamentais para análise de dados complexos, tais como, facilidade em lidar com dados que possuem muito mais variáveis do que indivíduos e a capacidade de incorporar relações não lineares entre as variáveis resposta e preditoras e suas interações.

A *Random Survival Forest* pode ser utilizada como modelo preditivo para a taxa de falha em dados de sobrevivência. Cada uma de suas árvores realiza predições do risco individualmente, a predição final é dada pela combinação dos resultados de todas as árvores de decisão (MOGENSEN; ISHWARAN; GERDS, 2012). A RSF é baseada no método *bagging* (*bootstrap aggregated*), que se fundamenta na retirada de b amostras *bootstrap* ($b = 1, \dots, B$) do conjunto de treinamento. E, em seguida, para cada uma destas amostras, é construída uma árvore de decisão (BOU-HAMAD; LAROCQUE; BEN-AMEUR, 2011). Cada árvore de decisão é obtida pela partição recursiva dos indivíduos em cada um de seus nós, de modo a maximizar as diferenças de sobrevivência entre dois nós filhos. A divisão dos indivíduos em

grupos é feita com base nas variáveis preditoras. Dentre as m variáveis, são selecionadas aleatoriamente p variáveis, em cada nó, a variável que implicar na melhor divisão de acordo com algum critério, será utilizada para a divisão binária dos indivíduos. As regras de decisão mais comuns são: teste log-rank e escore log-rank (ISHWARAN et al., 2008; BOU-HAMAD; LAROCQUE; BEN-AMEUR, 2011).

Segundo Mogensen, Ishwaran e Gerds (2012) em cada nó terminal da árvore, a função taxa de falha acumulada condicional é calculada via estimador de Nelson-Aalen, pela expressão:

$$\widehat{H}_b(t|\mathbf{x}) = \int_0^t \frac{\widetilde{N}_b(du, \mathbf{x})}{\widetilde{Y}_b(u, \mathbf{x})},$$

em que: $\widetilde{N}_b(s, \mathbf{x})$ é o número de eventos até o tempo s e $\widetilde{Y}_b(s, \mathbf{x})$ é o número de indivíduos em risco no tempo s. A função de sobrevivência conjunta, que agrega todas as árvores de decisão, é dada por:

$$\hat{S}(t|\mathbf{x}) = \exp\left\{-\frac{1}{B}\sum_{b=1}^{B} \widehat{H}_b(t|\mathbf{x})\right\}.$$

Predições em um novo conjunto de entrada são realizadas pelas combinações das saídas das árvores de sobrevivência com base nas amostras *bootstrap* do conjunto de treinamento. O pacote *randomForestSRC* (ISHWARAN; KOGALUR, 2020) possibilita a implementação da RF de Breiman para a solução de tarefas de sobrevivência, regressão e de classificação.

2.3.2 *Gradient Boosting Machine*

Proposto por Friedman (2001), *Gradient Boosting Machine* (GBM) é um poderoso algoritmo de aprendizado supervisionado, com sucesso comprovado em diversas aplicações, fato que o tornou popular e largamente utilizado na prática (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; LU et al., 2019). O GBM é um método *ensemble*, não paramétrico, em que o modelo preditivo é obtido por expansões aditivas, com ajuste sequencial de classificadores de base (*weak learners*) (FRIEDMAN, 2001). O algoritmo *boosting* foi originalmente proposto por Breiman (1996a, 1996b), e utilizado para solução de problemas de classificação e de regressão. Ridgeway (1999) propôs uma adaptação do *boosting* para a família exponencial e para modelos de regressão de riscos proporcionais, possibilitando sua utilização em dados de

sobrevivência.

O *bagging* e *boosting* se baseiam na ideia de criar um conjunto de classificadores por meio da reamostragem dos dados. A diferença básica entre estes procedimentos, é que no *boosting* a reamostragem é planejada de modo a obter dados de treinamento mais informativos a cada iteração. Em problemas de regressão, o ajuste de cada árvore de decisão é adequado de acordo com os resíduos estimados para a árvore anterior, um parâmetro de encurtamento é utilizado para limitar a contribuição de uma árvore recém adicionada para predição. A combinação dos classificadores base gera um classificador final (ou comitê) com performance preditiva superior a de qualquer um dos classificadores bases que constituem o modelo (FRIEDMAN, 2001; POLIKAR, 2006). O modelo GBM é apresentado a seguir.

Seja $(x_i, y_i)_{i=1}^{N}$ o conjunto de dados de treinamento, com: $\mathbf{x} = (x_1, \dots, x_d)$, em que $\mathbf{x}$ representa as variáveis preditoras e $\mathbf{y}$ a variável resposta, e seja F uma função que relaciona $\mathbf{x}$ e $\mathbf{y}$. O objetivo do GBM é reconstruir a relação de dependência funcional desconhecida entre $\mathbf{x}$ e $\mathbf{y}$, pela estimação da função $F(x, \theta)$, que minimiza a função de perda média dos dados de treinamento dada por $\sum_{i=1}^{N} \Phi(y_i, F(x_i; \theta)$. Assumindo para F a forma de uma expansão aditiva, segue que:

$$F(\mathbf{x}) = \sum_{m=0}^{M} \boldsymbol{\beta}_m f(\mathbf{x}; \theta_m),$$

em que: $\boldsymbol{\beta}_m$ são os coeficientes de expansão e $f(\mathbf{x}; \theta_m)$ é uma função real multivariada de $\mathbf{x}$, caracterizada pelo conjunto de parâmetros θ_m, também conhecida como classificador base ou *weak learner*, com: $m = 1, \dots, M$ representando o número de iterações (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; NATEKIN; KNOLL, 2013). Os parâmetros β e θ são estimados pela expressão:

$$(\beta, \theta) = \underset{\beta,\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \Phi\big(y_i, F_{m-1}(x_i) + \beta f(x_i; \theta)\big).$$

De acordo com Friedman (2002) e Chen et al. (2013), a estimação destes parâmetros pode ser feita por meio de uma aproximação da equação anterior em dois passos. Primeiro é ajustado o classificador base $f(x_i; \theta_m)$ pela estimação do parâmetro θ, em seguida, são estimados os coeficientes de expansão β. A minimização da função de perda é realizada via algorítmo *steepest descent*, atribuindo-se a função nula para f na iteração zero ($f_0 \equiv 0$), em cada uma das iterações seguintes, é computado o pseudo-resíduo (g^m), ou gradiente negativo da

função de perda em relação a F_m, dados por:

$$g_i^m = -\left[\frac{\partial \Phi\left(y_i; F_m(x_i)\right)}{\partial F_m(x_i)}\right].$$

O classificador base com melhor ajuste aos pseudos resíduos é obtido pela minimização dos erros quadráticos médios: $\theta_m = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} (g_i^m - f(x_i; \theta_m))^2$. Por fim, os coeficientes de expansão são estimados pela expressão: $\beta_m = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Phi(y_i, F_{m-1}(x_i) + \beta f(x_i; \theta_m))$. Em cada iteração a função F é atualizada por $F_m(x) = F_{m-1}(x) + v\beta_m f(x_i; \theta_m)$, em que: v representa um parâmetro de encurtamento $0 < v \leq 1$, que tem por função evitar o super ajustamento (FRIEDMAN, 2002; CHEN et al., 2013).

A função de perda deve ser uma função derivável. São exemplos as funções de: perda quadrática, perda absoluta, perda Huber, perda exponencial, dentre outras (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Para aplicação do GBM em análise de sobrevivência, Ridgeway (1999) propôs a utilização do negativo do logaritmo da verossimilhança parcial como função de perda. Assim, tem-se:

$$\Phi(y, F) = -\sum_{i=1}^{N} \delta_i \left\{ F(x_i) - \log\left(\sum_{j:t_j \geq t_i} e^{F(x_j)} \right) \right\}.$$

2.3.3 Análise de Componentes Principais Supervisionados

Bair e Tibshirani (2004) propuseram a análise de regressão via componentes principais supervisionados (SPCA), para o estudo de dados em que o número de variáveis preditoras excede muito o número de observações. Esta metodologia pode ser aplicada para regressão e análise de sobrevivência, caso em que a variável resposta apresenta observações incompletas devido a presença de censura. Neste cenário, modelos de regressão tradicionais podem fornecer resultados incorretos, enquanto a SPCA soluciona estas dificuldades pela redução de dimensionalidade, e com a utilização do modelo de regressão apropriado, de acordo com a tarefa a ser realizada. Esta técnica é semelhante à análise de componentes principais (PCA), a diferença principal entre as duas metodologias, é que na SPCA, os componentes principais são aplicados ao um subconjunto das variáveis, enquanto no PCA todas as variáveis são utilizadas para derivação dos componentes principais (BAIR et al., 2006).

A ideia básica do SPCA é realizar a análise de componentes principais utilizando-se

apenas das variáveis que estão fortemente correlacionadas com a variável resposta. Seja $\mathbf{X}_{n\times p}$ a matriz de variáveis, com n e p, representando, respectivamente, o número de indivíduos e o número de variáveis. Segundo Bair et al. (2006), a SPCA se inicia com o cálculo da estatística escore de Cox, para cada uma das variáveis preditoras $(s_1, s_2, \dots, s_p)$. Este escore mede o efeito univariado de cada variável sobre o tempo de sobrevivência. Seja $l_j(\beta)$ o logaritmo da verossimilhança parcial que relaciona uma variável $\mathbf{X}_j$ à resposta **y**, com: β indicando o coeficiente de regressão univariado do modelo de Cox, e sejam $U_j(\beta_0)$ e $I_j(\beta_0)$, tais que (BAIR et al., 2006):

$$U_j(\beta_0) = \left.\frac{dl}{d\beta}\right|_{\beta\equiv\beta_0} \text{ e } I_j(\beta_0) = -\left.\frac{d^2 l_j}{d\beta^2}\right|_{\beta\equiv\beta_0},$$

segue que a estatística escore de Cox para uma certa variável preditora j é dada por:

$$s_j = \frac{U_j(0)^2}{I_j(0)}.$$

Apenas as variáveis cujo valor absoluto dos escores superem o limiar θ, estimado via validação cruzada, são mantidas em análise. Estas variáveis são utilizadas para compor a matriz reduzida dos dados $\mathbf{X}^{\theta}_{n\times p_1}$, em que: p_1 é o número de variáveis filtradas. A decomposição em valores singulares de $\mathbf{X}^{\theta}_{n\times p_1}$ é: $\mathbf{X}^{\theta}_{n\times p_1} = \mathbf{U}_\theta \mathbf{D}_\theta \mathbf{V}_\theta^T$. As dimensões de **U**, **D** e **V**, são, respectivamente, $n \times m$, $m \times m$, $p_1 \times m$, e $m = \min(n, p_1)$. Em seguida, são obtidos os componentes principais supervisionados da matriz reduzida, dados por: $\mathbf{U}_\theta = (u^1_\theta, u^2_\theta, \dots, u^m_\theta)$. Por fim, o primeiro componente principal, ou alguns dos primeiros componentes principais, são utilizados no modelo de riscos proporcionais de Cox para a predição da resposta (BAIR et al., 2006).

REFERÊNCIAS

ANDERSEN, P. K.; GILL, R. Cox's Regression Model for Counting Processes: A large sample study. **Annals of Statistics,** v. 10, p. 1100-1200, 1982.

BAIR, E.; TIBSHIRANI, R. Semi-supervised methods to predict patient survival from gene expression data. **PLoS Biology,** v. 2, n. 4, p. 511-522, 2004.

BAIR, E.; HASTIE, T.; PAUL, D.; TIBSHIRANI, R. Prediction by Supervised Principal Components. **Journal of the American Statistical Association,** v. 101, n. 473, p. 119-137,

2006.

BARRÍA, A.; CHRISTENSEN, K. A.; YOSHIDA, G. M.; CORREA, K.; JEDLICKI, A.; LHORENTE, J. P.; DAVIDSON, W. S.; YÁÑEZ, J. M. Genomic Predictions and Genome Wide Association Study of Resistance Against Piscirickettsia salmonis in Coho Salmon (Oncorhynchus kisutch) Using ddRAD Sequencing. **G3,** v. 8, p. 1183-1194, 2018.

BHAT, J. A.; ALI, S.; SALGOTRA, R. K.; MIR, Z. A.; DUTTA, S.; JADON, V.; TYAGI, A.; MUSHTAQ, M.; JAIN, N.; SINGH, P. K.; SINGH, G. P.; PRABHU, K. V. Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. **Frontiers in Genetics,** v. 7, p. 1-11, 2016.

BISHOP, C. M. **Pattern Recognition and Machine Learning.** 1st ed. Singapore: Springer, 2006. 738 p.

BREIMAN, L. Bagging predictors. **Machine Learning,** v. 24, n. 2, p. 123-140, 1996a.

BREIMAN, L. Heuristics of instability and stabilization in model selection. **The annals of statistics,** v. 24, n. 2, p. 2350-2383, 1996b.

BREIMAN, L. Random forest. **Machine Learning,** v. 45, n. 1, p. 5-32, 2001.

BOU-HAMAD, I.; LAROCQUE, D.; BEN-AMEUR, H. A review of survival trees. **Statistics Surveys,** v. 5, p. 44-71, 2011.

CAMPOS, G.; GIANOLA, D.; ROSA, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science,** v. 87, n. 6, p. 1883-1887, 2009.

CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, v. 193, p. 327-345, 2013.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics,** v. 99, p. 323-329, 2012.

CHEN, Y.; JIA, Z.; MERCOLA, D.; XIE, X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. **Computation and Mathematical Methods in Medicine,** p. 1-8, 2013.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência.** 1. ed. São Paulo: Edgard Blücher, 2006. 369 p.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society. Series B (Methodological),** v. 34, n. 2, p. 187-202, 1972.

EHRLINGER, J. ggRandomForests Exploring random forest survival. R Vignette, 2016.

EMMERT-STREIB, F.; DEHMER, M. Introduction to Survival Analysis in Practice. **Machine Learning and Knowledge Extraction,** v. 1, n. 1, p. 1013-1038, 2019.

EMURA, T.; CHEN, Y. H. Gene selection for survival data under dependent censoring, a copula-based approach. **Statistical Methods in Medical Research,** v. 25, n. 6, p. 2840-2857, 2016.

EMURA, T.; MATSUI, S.; RONDEAU, V. **Survival Analysis with Correlated Endpoints:** Joint Frailty-Copula Models. Singapore: Springer, 2019. 118 p.

FREUND, Y.; SCHAPIRE, R. Experiments with a new boosting algorithm. **In:** Machine Learning: Proceedings of the Thirteenth International Conference, 1996. p. 148-156.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics,** v. 29, n. 5, p. 1189-1232, 2001.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational Statistics & Data Analysis,** v. 38, n. 4, p. 367-378, 2002.

FLEMING, T. R.; HARRINGTON, D. P. **Counting processes and survival analysis.** New York: Wiley, 1991. 429 p.

GEBELEIN, H. Das statistische problem der correlation als variation und eigenwertproblem und sein zusammenhang mit der ausgleichrechnung. **Zeitschrift für angewandte Mathematik und Mechanik,** v. 21, n. 6, p. 364-379, 1941.

GIANOLA, D.; FERNANDO, R. L.; STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics,** v. 173, p. 1761-1776, 2006.

GIANOLA, D., van KAAM, J. B. C. H. M., Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics,** v. 178, p. 2289-2303, 2008.

GIANOLA, D.; OKUT, H.; WEIGEL, K. A.; ROSA, G. J. M. Predicting complex quantitative traits with Bayesian neural networks: A case study with Jersey cows and wheat. **BMC Genetics,** v. 12, n. 87, p. 1-14, 2011.

GIOLO, S. R.; DEMÉTRIO, C. G. B. A frailty modeling approach for parental effects in animal breeding. **Journal of Applied Statistics,** v. 38, n. 3, p. 619-629, 2011.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programs. **Nature Reviews Genetics,** v. 10, p. 381-391, 2009.

GONZÁLEZ-CAMACHO, J. M.; ORNELLA, L.; PÉREZ-RODRÍGUEZ, P.; GIANOLA, D.; DREISIGACKER, S. CROSSA, J. Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance. **The Plant Genome,** v. 11, n. 2, p. 1-15, 2018.

GONZÁLEZ-RECIO, O.; FORNI, S. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. **Genetics Selection Evolution,** v. 43, n. 7, p. 1-12, 2011.

GONZÁLEZ-RECIO, O.; ROSA, G. J. M.; GIANOLA, D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science,** v.

166, p. 217-231, 2014.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. The impact of genetic relationship information on genome-assisted breeding values. **Genetics,** v. 177, p. 2389-2397, 2007.

HASTIE, T. J.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning:** Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. 745 p.

HAYES, B. J.; PRYCE, J.; CHAMBERLAIN, A. J.; BOWMAN, P. J.; GODDARD, M. E. Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Color, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. **PLoS Genetics,** v. 6, n. 9, p. 1-11, 2010.

HENDERSON, C. R. 1953 Estimation of variance and covariance components. **Biometrics,** v. 9, p. 226-252, 1953.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics,** v. 42, n. 1, p. 80-86, 1970.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. **G3**, v. 4, p. 1027-1046, 2014.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. Random survival forests. **Annals of Applied Statistics,** v. 2, n. 3, p. 841-860, 2008.

ISHWARAN, H.; KOGALUR, U. B. Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 2.9.3., 2020. Disponível em: https://cran.r-project.org/web/packages/randomForestSRC/index.html. Acesso em: 3 abr. 2020.

JAYANTHI, K.; MAHESH, C. A Study on machine learning methods and applications in genetics and genomics. **International Journal of Engineering &Technology,** v. 7, n. 1.7, p. 201-204, 2018.

KARIM, M. R.; ISLAM, M. A. **Reliability and Survival Analysis.** Singapore: Springer, 2019. 252 p.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. **Informatica,** v. 31, n. 3, p. 249-268, 2007.

LAWLESS, J. F. **Statistical Models and Methods for Lifetime Data.** 2nd ed. New Jersey: Wiley-Interscience, 2002. 663 p.

LI, Y.; GILLESPIE, B. W.; SHEDDEN, K.; GILLESPIE, J. A. Profile Likelihood Estimation of the Correlation Coefficient in the Presence of Left, Right or Interval Censoring and Missing Data. **The R Journal,** v. 10/2, p. 159-179, 2018.

LU, H.; KARIMIREDDY, S. P.; PONOMAREVA, N.; MIRROKNI, V. Accelerating Gradient Boosting Machine. **ArXiv,** 2019.

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. 1819-1829, 2001.

MOGENSEN, U. B.; ISHWARAN, H.; GERDS, T. A. Evaluating random forests for survival analysis using prediction error curves. **Journal of Statistical Software,** v. 50, n. 11, p. 1-23, 2012.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics,** v. 7, p. 1-21, 2013.

OKUT, H.; GIANOLA, D.; ROSA, G. J. M.; WEIGEL, K. A. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. **Genetics research,** v. 93, n. 3, p. 189-201, 2011.

PALAIOKOSTAS, C.; FERRARESSO, S.; FRANCH, R.; HOUSTON, R. D.; BARGELLONI, L. Genomic Prediction of Resistance to Pasteurellosis in Gilthead Sea Bream (Sparus aurata) Using (Sparus aurata) 2b-RAD Sequencing. **G3,** v. 6, p. 3693-3700, 2016.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika,** v. 58, p. 545-554, 1971.

PÉREZ, P.; CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics,** v. 198, n. 2, p. 483-495, 2014.

PÉREZ-ENCISO, M. Animal breeding learning from machine learning. **Journal of Animal Breeding and Genetics,** v. 134, p. 85-86, 2017.

PÉREZ-RODRÍGUEZ, P.; GIANOLA, D.; GONZÁLEZ-CAMACHO, J. M.; CROSSA, J.; MANÈS, Y.; DREISIGACKER, S. Comparison Between Linear and Non-parametric Regression Models for Genome Enabled Prediction in Wheat. **G3,** v. 2, p. 1595-1605, 2012.

PIEPHO, H. P.; MÖHRING, J.; MELCHINGER, A. B. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica,** v. 161, p. 209-228, 2008.

POLIKAR, R. Ensemble based systems in decision making. **IEEE Circuits and Systems Magazine,** v. 6, p. 3, p. 21-45, 2006.

RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial.** Viçosa, MG: Editora UFV, 2012, 291 p.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência.** Viçosa: Editora Suprema, 2014. 881 p.

RIDGEWAY, G. The state of boosting. **Computing Science and Statistics,** v. 31, p. 172-181, 1999.

RIPATTI, S.; PALMGREN, J. Estimation of multivariate frailty models using penalized partial likelihood. **Biometrics,** v. 56, p. 1016-1022, 2000.

ROBINSON, G. K. That BLUP is a good thing: The estimation of random effects. **Statistical Science,** v. 6, n. 1, p. 15-32, 1991.

SAMORÈ, A. B.; FONTANESI, L. Genomic selection in pigs: state of the art and perspectives. **Italian Journal of Animal Science,** v. 15, n. 2, p. 211-232, 2016.

SANTOS, V. S.; MARTINS, F. S.; RESENDE, M. D.; AZEVEDO, C. F.; LOPES, P. S.; GUIMARÃES, S. E.; GLÓRIA, L.; SILVA, F. F. Genomic selection for slaughter age in pigs using the Cox frailty model. **Genetics and Molecular Research,** v. 14, n. 4, p. 12616-12627, 2015.

SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selections Evolution,** v. 41, n. 29, p. 1-8, 2009.

SORENSEN, D. A.; GIANOLA, D.; KORSGAARD, I. R. Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. **Acta Agriculturae Scandinavica A-Animal Sciences,** v. 48, n. 4, p. 222-229, 1998.

TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation. **Journal of the American statistical Association,** v. 82, n. 398, p. 528-540, 1987.

THERNEAU, T. M.; GRAMBSCH, P. M.; PANKRATZ, V. S. Penalized survival models and frailty. **Journal of Computational and Graphical Statistics,** v. 12, p. 156-175, 2003.

THERNEAU, T. M. coxme: Mixed Effects Cox Models. R-package description., p. 1–14, 2020. Disponível em: https://cran.r-project.org/web/packages/coxme/vignettes/coxme.pdf. Acesso em: 24 mar. 2020.

TOBIN, J. Estimation of relationships for limited dependent variables. **Econometrica**, v. 26, n. 1, p. 24-36, 1958.

van DYK, D. A.; MENG, X. L. The Art of Data Augmentation. **Journal of Computational and Graphical Statistics,** v. 10, n. 1, p. 1-50, 2001.

VanRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Sciences,** v. 91, p. 4414-4423, 2008.

VAPNIK, V. N. **The Nature of Statistical Learning Theory.** 2. ed. [S.l.]: Springer, 2000. 314 p.

YANG, P.; YANG, Y. H.; ZHOU, B. B.; ZOMAYA, A. Y. A review of ensemble methods in bioinformatics. **Current Bioinformatics,** v. 5, n. 4, 296-308, 2010.

YAO, C.; ZHU, X.; WEIGEL, K. A. Semi-supervised learning for genomic prediction of novel traits with small reference populations: an application to residual feed intake in dairy cattle.

Genetics Selection Evolution, v. 48, n. 84, p. 1-9, 2006.

WANG; C. S. Implementation issues in Bayesian analysis in animal breeding, **In:** Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, 11-16 January 1998, University of New England, Armidale, Australia, 1998. p. 481−488.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetics Research,** v. 75, p. 249-252, 2000.

CAPÍTULO II

MEDIDAS ALTERNATIVAS PARA AVALIAÇÃO DA ACURÁCIA E DO VIÉS DA PREDIÇÃO GENÔMICA COM OBSERVAÇÕES CENSURADAS

RESUMO

Este trabalho teve como objetivo propor e comparar medidas mais adequadas para estimação da acurácia e viés para a predição genômica com dados censurados, considerando os modelos misto de Cox e normal truncado. Foram simuladas informações fenotípicas censuradas (10%, 40% e 70%) para quatro características: C1: $h^2 = 0,07$ e h^2 de QTL = 0,07, C2: $h^2 = 0,07$ e h^2 de QTL = 0, C3: $h^2=0,27$ e h^2 de QTL = 0,27 e C4: $h^2=0,27$ e h^2 de QTL = 0. O genoma foi constituído por 52.885 marcadores e 88 QTLs, distribuídos aleatoriamente por 29 cromossomos. A estimação dos efeitos dos marcadores foi realizada para uma população de treinamento de 6.000 animais. Outros 3.000 animais foram utilizados como população de validação. Os valores genéticos genômicos (GBVs) foram estimados via modelo misto de Cox e modelo normal truncado. A acurácia dos modelos foi calculada com base na correlação (de Pearson - CP, maximal - CM ou de Pearson para dados censurados - CPC) entre os GEBVs e o fenótipo incompleto (censurado), e verdadeiros valores genéticos genômicos (TBV). O viés foi estimado via regressão linear simples e regressão Tobit. As correlações maximal e de Pearson para dados censurados mostraram-se estatisticamente superiores à correlação de Pearson para a característica C3 com 10 e 40% de censura, para 70% de censura, a CPC superou as outras duas medidas. Para as demais características, as medidas propostas de acurácia foram superiores ou estatisticamente iguais à CP. A correlação de Pearson, na maioria das vezes, apresentou estimativas que se distanciaram consideravelmente dos valores estimados com base nos fenótipos completos. Este fato ficou mais evidente para as porcentagens de censura iguais a 40 e 70%. Os coeficientes associados aos efeitos marginais, estimados para a regressão Tobit apresentaram estimativas próximas das obtidas para a regressão linear simples, enquanto o coeficiente relativo à variável latente apresentou comportamento quase inalterado com o aumento da censura na maioria das vezes. Do ponto de vista estatístico, o uso de metodologias próprias para dados censurados deve ser priorizado, até mesmo para baixas porcentagens de censura.

Palavras-chave: seleção genômica, simulação, modelo misto de Cox, modelo normal truncado.

ABSTRACT

This study aimed to propose and compare the most appropriate measures for estimating accuracy and bias for genomic prediction with censored data, considering the mixed Cox and truncated normal models. Censored phenotypic information (10%, 40% and 70%) were simulated for four characteristics: C1: $h^2 = 0.07$ and h^2 of QTL = 0.07, C2: $h^2 = 0.07$ and h^2 of QTL = 0, C3: $h^2 = 0.27$ and h^2 of QTL = 0.27 and C4: $h^2 = 0.27$ and h^2 of QTL = 0. The genome consisted of 52,885 markers and 88 QTLs, randomly distributed across 29 chromosomes. The estimation of the effects of the markers was performed for a training population of 6,000 animals. Another 3,000 animals were used as a validation population. Genomic breeding values (GBVs) were estimated using the Cox mixed model and truncated normal model. The accuracy of the models was calculated based on the correlation (Pearson - CP, maximal - CM or Pearson for censored data - CPC) between the GEBVs and the incomplete (censored) phenotype, and true genomic breeding values (TBV). The bias was estimated via simple linear regression and Tobit regression. Maximal and Pearson correlations for censored data were statistically superior to Pearson's correlation for characteristic C3 with 10 and 40% censorship, for 70% censorship, CPC surpassed the other two measures. For the other characteristics, the proposed measures of accuracy were superior or statistically equal to the CP. Pearson's correlation, in most cases, presented estimates that differed considerably from the estimated values based on the complete phenotypes. This fact was more evident for the censorship percentages equal to 40 and 70%. The coefficients associated with the marginal effects, estimated for the Tobit regression, presented estimates close to those obtained for the simple linear regression, while the coefficient related to the latent variable showed almost unchanged behavior with the increase in censorship in most cases. From a statistical point of view, the use of proprietary methodologies for censored data should be prioritized, even for low censorship percentages.

Key words: genomic selection, simulation, mixed Cox model, truncated normal model.

1 INTRODUÇÃO

Com a disponibilidade de painéis de marcadores de alta densidade, a seleção genômica ampla proposta por Meuwissen et al. (2001), tornou-se uma poderosa ferramenta no melhoramento genético, devido a sua alta performance na predição de valores genéticos para características complexas. Muitos métodos foram propostos para implementação da seleção genômica, em grande parte desses métodos, a informação dos marcadores é incorporada a modelos de predição, nos quais os fenótipos são regressados considerando um conjunto de marcadores como variáveis explicativas em um modelo de regressão (CAMPOS et al., 2013).

Características que possuem observações censuradas, como idade ao primeiro parto e ao abate, são frequentemente avaliadas em programas de melhoramento genético animal, os quais, de modo geral, tratam essas características como não censuradas. Na presença de censura, o que se sabe é que o valor para uma dada característica é maior ou menor do que um certo limiar, ou pertencente a um intervalo, não sendo possível observar um valor exato para resposta. Santos et al. (2015) avaliaram a característica tempo ao abate de suínos, neste caso, a censura se caracterizou pela presença de animais que não obtiveram o peso mínimo necessário para serem abatidos, sendo assim considerados como observações censuradas. Outro exemplo de censura na produção animal é a seleção de indivíduos geneticamente superiores para resistência a doenças, o que resultaria em maior tempo de sobrevivência dos animais (ALEMU et al., 2016; PALAIOKOSTAS et al., 2016; VALLEJO et al., 2016). Nestes estudos, uma metodologia comum é desafiar os indivíduos durante o período de teste, sendo comum que alguns indivíduos não morram durante o período de realização do experimento, o que caracteriza a presença de dados censurados.

Métodos para predição genômica de características censuradas vem sendo propostos ao longo dos últimos anos. Kärkkäinen e Sillanpää (2013) propuseram um modelo bayesiano hierárquico, chamado modelo de limiar bayesiano, para análise de dados em escala binária, ordinal ou que apresentem observações censuradas. Pérez e Campos (2014) propuseram uma classe de modelos, denominada *Bayesian Generalized Linear Regression* (BGLR), que permitem modelar respostas normais, binárias, ordinais ou censuradas. Ademais, Santos et al. (2015) foram os pioneiros ao utilizarem o modelo misto de Cox para a predição genômica para fenótipos censurados.

As comparações entre diversos modelos em seleção genômica ampla é comumente realizada utilizando a acurácia e o viés de predição. A acurácia pode ser definida como a

correlação de Pearson entre valor genético verdadeiro e os valores genéticos genômicos preditos. O viés, por sua vez, é estimado por meio do ajuste de um modelo de regressão linear simples. Contudo, essas medidas podem não ser adequadas quando o fenótipo é parcialmente observado, ou seja, há presença de observações censuradas. De acordo com Li et al. (2018), quando estimadores tradicionais para dados completos, são aplicados a dados censurados, eles reduzem a precisão das estimativas e podem introduzir viés. Segundo os autores, abordagens que incorporam observações incompletas, são geralmente mais poderosas. Assim, o uso da correlação de Pearson para avaliar a habilidade preditiva no quadro de dados censurados, pode implicar em conclusões equivocadas.

Estudos que avaliem medidas alternativas à correlação de Pearson e à regressão linear, respectivamente, para a avaliação da acurácia e viés no contexto de dados censurados, não foram encontrados na literatura consultada. Considerando o exposto, objetivou-se propor e comparar medidas mais adequadas para estimação da acurácia e do viés na predição genômica com dados censurados, considerando os modelos misto de Cox e normal truncado sob diferentes cenários de censura e arquitetura genética da característica.

2 MATERIAL E MÉTODOS

2.1 Dados simulados

As informações fenotípicas e genotípicas usadas neste estudo foram simuladas por meio do software QMSim (SARGOLZAEI; SCHENKEL, 2009). Os parâmetros utilizados no processo de simulação foram semelhantes aos utilizados por Brito et al. (2011). A população histórica foi constituída por 1.000 gerações com tamanho inicial de 2.000 indivíduos cada. Em seguida foi provocado um gargalo genético, de forma que nas 1.020 gerações seguintes, o tamanho da população decresceu gradualmente de 2.000 para 1.500, com o intuito de criar um desequilíbrio de ligação (LD) inicial.

Posteriormente, foi criada uma população expandida pelo acasalamento aleatório de indivíduos da última geração da população histórica, gerando mais oito gerações, com cinco crias por fêmea. Por fim, foi criada uma população recente, obtida pelo acasalamento de 120 machos com 6.000 fêmeas, selecionados aleatoriamente da última geração da população expandida, gerando outras dez gerações, com uma cria por fêmea. A proporção de indivíduos machos foi mantida igual a 0,5 nas populações expandida e recente.

O genoma foi simulado com 52.885 marcadores moleculares SNPs e 88 QTLs, distribuídos por 29 cromossomos, compondo um tamanho total de 2.740cM. Os SNPs foram igualmente espaçados e distribuídos aleatoriamente nos cromossomos. O efeito dos QTLs foi amostrado de uma distribuição gama com parâmetro de forma igual a 0,4 e parâmetro de escala determinado internamente pelo software QMSim de acordo com a variância genética simulada. Para marcadores e QTLs foi assumida uma menor frequência alélica (MAF) de 0,10, sendo considerada uma taxa de mutação de 1×10^{-5}, para marcadores e QTLs na população histórica.

O genoma foi escolhido de modo a mimetizar o genoma bovino, e no processo de simulação buscou-se representar a característica fenotípica idade ao primeiro parto. Para caracterização da população base, foram simuladas quatro características fenotípicas: C1 e C2, com herdabilidade de QTL, respectivamente, iguais a 0,07 e 0, e a herdabilidade e a variância fenotípica, de acordo com os valores estimados por Costa (2017) via modelo linear - limiar, respectivamente, iguais a 0,07 e 17,64. As demais características, tratam-se de extrapolações para o caso de características fictícias com herdabilidade igual a 0,27, e herdabilidade de QTL de 0,27 (C3) e 0 (C4).

As observações censuradas foram obtidas pela escolha de um valor fenotípico (C), como limiar para a censura à direita, assim, valores fenotípicos maiores que C, foram considerados como observações censuradas. Sendo $\mathbf{Y_c}$ o vetor de fenótipos simulados pelo QMSim, a partir de $\mathbf{Y_c}$ e C foram criadas outras duas variáveis fenotípicas, dadas por: $\mathbf{y_c} = \min(\mathbf{Y_c}, C)$, sendo C, calculado por: $C = Q(\mathbf{Y_c}, \text{probabilidade} = 1 - PC)$, em que: PC é a proporção de observações censuradas e Q é o quantil da distribuição normal; e δ uma variável indicadora de censura, que assume o valor 1, se $Y_c \leq C$ ou 0, se $Y_c > C$.

A variável censurada foi construída de modo a se obter fenótipos com 10, 40 e 70% de observações censuradas. Assim, foi avaliado neste estudo um total de 12 cenários, obtidos por dois valores de herdabilidade, duas herdabilidades de QTL e três proporções de dados censurados. Todo o processo de simulação foi repetido 10 vezes.

2.2 Métodos estatísticos

Apenas os indivíduos das gerações oito, nove e dez da população recente foram utilizados para fins de predição genômica, sendo cada geração composta por 6.000 indivíduos, totalizando 18.000 animais. As informações genotípicas desses animais foram utilizadas para a obtenção da matriz de parentesco genômico (**G**). Como a característica simulada é restrita às

fêmeas, apenas a informação fenotípica delas, foi utilizada para ajuste dos modelos. Assim, as fêmeas das gerações oito e nove foram consideradas como população de treinamento e as fêmeas da geração dez como população de validação.

O modelo misto de Cox e o modelo normal truncado foram utilizados para predizer os valores genéticos aditivos (**a**) de cada indivíduo. O modelo misto de Cox é descrito, segundo Ripatti e Palmgren (2000) e Therneau, Grambsch e Pankratz (2003), pela equação:

$$\lambda(t) = \lambda_0(t)\exp\{\mathbf{X\beta} + \mathbf{Za}\},$$

em que: $\lambda_0(t)$ é a função base de taxa de falha; **X** e **Z** são as matrizes de incidência para os efeitos fixos e aleatórios, respectivamente; **β** é um vetor de efeitos fixos e **a** um vetor de efeitos genéticos aditivos, para o qual é assumida uma distribuição normal com média zero e matriz de covariância $\mathbf{G}\sigma_a^2$, sendo σ_a^2 a variância genética aditiva.

A função de verossimilhança parcial para o modelo misto de Cox, é dada por (GIOLO; DEMÉTRIO, 2011):

$$L = \int PL(\boldsymbol{\beta}, \mathbf{a}) \frac{1}{\sqrt{2\pi|\mathbf{G}|}} \exp\left[-\frac{1}{2}\mathbf{a}'^{\mathbf{G}^{-1}}\mathbf{a}\right] d\mathbf{a}, \qquad (1)$$

sendo: PL a função de verossimilhança parcial do modelo de riscos proporcionais de Cox e os demais termos como definidos anteriormente. Como a integral da equação (1) não apresenta uma forma fechada, Ripatti e Palmgren (2000) sugeriram a utilização da aproximação de Laplace para obter o logaritmo da função de verossimilhança, e superar os problemas advindos do cálculo integral multidimensional. O modelo misto de Cox foi ajustado via pacote *coxme* (THERNEAU, 2020) do software R (R Development Core Team, 2019).

O modelo Normal truncado foi ajustado via modelo de regressão *Reproducing Kernel Hilbert Spaces* bayesiano (RKHS) do pacote BGLR do software R (R Development Core Team, 2019), considerando a matriz Kernel (**K**), como sendo igual a matriz de parentesco genômico (**G**). Segundo Pérez e Campos (2014), o ajuste de modelos lineares, na presença de características fenotípicas censuradas, é feito considerando que os fenótipos são amostrados de uma distribuição normal truncada. O seguinte modelo linear misto foi considerado: $\mathbf{y_c} = 1\boldsymbol{\mu} + \mathbf{Za} + \mathbf{e}$, em que: $\mathbf{y_c}$ é o vetor de fenótipos observados; $\boldsymbol{\mu}$ é o vetor de intercepto; **Z** é a matriz de incidência que relaciona fenótipos com o efeito aleatório animal; **a** é o vetor de valores genéticos aditivos, com distribuição $N(0, \mathbf{G}\sigma_a^2)$, sendo σ_a^2 a variância genética aditiva e **e** o vetor de resíduos com distribuição $N(0, \mathbf{I}\sigma_e^2)$. Para a variância genética aditiva e para variância

residual foram assumidas distribuições qui-quadrado escalonada invertida, com parâmetro de escala e grau de liberdade, escolhidos conforme indicado por Pérez e Campos (2014).

A matriz de parentesco genômico **G** foi utilizada nos modelos misto de Cox e normal truncando, com a formulação proposta por VanRaden (2008), dada por: $\mathbf{G} = \mathbf{WW}'/2\sum_{j=1}^{m} p_j q_j$, em que: **W** é uma matriz de ordem $n \times m$, composta pelos elementos: $w_{ij} = 0 - 2p_j = -2p_j$, $w_{ij} = 1 - 2p_j = q_j - p_j$ e $w_{ij} = 2 - 2p_j = 2q_j$, associados, respectivamente, aos genótipos, mm, Mm e MM dos marcadores, com: p_j sendo a frequência alélica de M no locos j, e $q_j = 1 - p_j$ a frequência alélica de m no locos j.

O algoritmo *Gibbs sampler* foi utilizado para a estimação, considerando 160.000 iterações, sendo as 20.000 primeiras descartadas, com espaçamento de 10 entre as amostras. A análise de convergência foi realizada por meio do critério de Geweke do pacote BOA (SMITH, 2007) do software R (R Development Core Team, 2019).

2.3 Medidas para avaliação dos modelos

O modelo misto de Cox e o modelo normal truncado foram ajustados considerando-se as informações fenotípicas, pela dupla $(\mathbf{y_c}, \delta)$ e genotípicas pela matriz de parentesco genômico (**G**). Os modelos também foram ajustados aos dados completos $\mathbf{Y_c}$, considerando um cenário onde a porcentagem de censura é igual a zero.

A acurácia (Ac) e o viés ($b_{(\widehat{GEBV},y)}$), são duas medidas tradicionalmente utilizadas para a comparação de modelos em seleção genômica ampla. Para estimar Ac e viés considerando os GEBVs obtidos para a variável fenotípica com observações completas ($\mathbf{Y_c}$), podemos utilizar a correlação de Pearson, entre os valores genéticos genômicos estimados ($\widehat{GEBVs}$) e os verdadeiros valores genéticos genômicos (TBVs), ou seja, $Ac = cor(\widehat{GEBV}, TBV)$. O viés pode ser obtido por meio da estimação do coeficiente angular da regressão linear dos TBVs em função dos $\widehat{GEBVs}$, ou pela expressão: $b_{(\widehat{GEBV},TBV)} = cov(\widehat{GEBV}, TBV)/\sigma^2_{\widehat{GEBV}}$. Como na prática estas medidas também são utilizadas para fenótipos incompletos, em um cenário onde não se conhece os TBVs, o viés e a acurácia também foram estimados considerando $\mathbf{y_c}$, ou seja, $Ac = cor(\widehat{GEBV}, \mathbf{y_c})/\sqrt{h^2}$ e $b_{(\widehat{GEBV},\mathbf{y_c})}$. Os valores $r_{(\widehat{GEBV},TBV)}$ e $b_{(\widehat{GEBV},TBV)}$ foram utilizados como referência para comparar as medidas alternativas de acurácia e viés, na presença de observações incompletas.

Neste trabalho, propomos duas medidas alternativas para a estimação da acurácia: a

correlação maximal (GEBELEIN, 1941; RÉNYI, 1959; BREIMAN; FRIEDMAN, 1985) e a correlação de Pearson para dados censurados (LI et al., 2018), a serem comparadas com a correlação de Pearson para variáveis contínuas.

A correlação maximal se baseia na determinação de transformações, possivelmente não lineares para duas variáveis, sujeitas à média zero e variância um, com o intuito de maximizar a correlação de Pearson entre estas duas transformações (FEIZI et al., 2017). Assim, sendo X e Y duas variáveis aleatórias reais, a correlação maximal entre elas é definida por: $\rho^*(X,Y) = \max_{f,g}\rho(f(X),g(Y))$, em que: ρ é o coeficiente de correlação de Pearson, e o máximo é tomado no conjunto de funções mensuráveis $f, g: \mathbb{R} \to \mathbb{R}$, com $0 < \mathrm{Var}\, f(X) < \infty$ e $0 < \mathrm{Var}\, g(Y) < \infty$ (BREIMAN; FRIEDMAN, 1985; BLÁZQUEZ; MIÑO, 2014). Para estimar a correlação maximal, foi utilizado o algoritmo proposto por Breiman e Friedman (1985), implementado no pacote *acepack* (SPECTOR et al., 2016) do software R.

Neste trabalho utilizou-se também uma medida própria para a correlação de dados censurados. Li et al. (2018) descrevem a implementação do método de máxima verossimilhança perfilada, para estimação da correlação de Pearson para dados bivariados com censura ou valores perdidos. Os autores propõem um modelo geral, para o caso da censura intervalar, que inclui como casos particulares, a censura à direita, à esquerda, valores observados exatos e dados com observações perdidas. A função de verossimilhança univariada é definida por:

$$L(\theta) = \prod_{i=1}^{k_1} F_\theta\left(x_i^{superior}\right) \cdot \prod_{i=k_1+1}^{k_2} \left[F_\theta\left(x_i^{superior}\right) - F_\theta\left(x_i^{inferior}\right)\right] \cdot \prod_{i=k_2+1}^{k_3} \left[1 - F_\theta\left(x_i^{inferior}\right)\right] \cdot \prod_{i=k_3+1}^{n} f_\theta(x_i),$$

os autores consideram que cada um dos casos pertence ao intervalo $(x_i^{inferior}, x_i^{superior}]$. Para a censura à direita, à esquerda, intervalar, para valores exatos e observações perdidas, são considerados, respectivamente, as desigualdades: $-\infty < x_i^{inferior} < x_i^{superior} = \infty$, $-\infty = x_i^{inferior} < x_i^{superior} < \infty$, $-\infty < x_i^{inferior} < x_i^{superior} < \infty$, $-\infty < x_i^{inferior} = x_i = x_i^{superior} < \infty$ e $-\infty = x_i^{inferior} < x_i^{superior} = \infty$ em que: k_1 representa o número de observações censuradas à esquerda com limite de detecção $x_i^{superior}, i = 1, \ldots, k_1$; seguidos por $k_2 - k_1$ censuras intervalares com limites $(x_i^{inferior}, x_i^{superior}]$, $i = k_1 + 1, \ldots, k_2$; $k_3 - k_2$ valores censurados à direita em $x_i^{inferior}$, $i = k_2 + 1, \ldots, k_3$ e $n - k_3$ valores exatos x_i, $i =$

$k_3 + 1, \ldots, n$ (LI et al., 2018).

Para dados censurados à direita, os autores comentam que a função de verossimilhança univariada pode ser escrita em termos da função de densidade, para valores não censurados, e da função de sobrevivência ($S_\theta(x) = P(X > x)$) para valores censurados à direita. Neste caso, θ representa um vetor de parâmetros, que inclui o coeficiente de correlação. A correlação de Pearson para dados censurados (CPC) foi obtida por meio do pacote *clikcorr* (LI et al., 2018) do software R.

A acurácia esperada (Ace) foi estimada com base nas matrizes de parentesco genômica (**G**) e na matriz aditiva associada ao pedigree (**A**), segundo equação proposta por Wientjes, Veerkamp e Calus (2013):

$$\text{Ace} = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}},$$

em que: N_p é o número de indivíduos na população de treinamento, genotipados e fenotipados, h^2 é a herdabilidade da característica e M_e é o número efetivo de locos, obtido por: $M_e = 1/\text{Var}(\mathbf{D})$, com: $\mathbf{D} = \mathbf{G} - \mathbf{A}$.

O viés foi estimado pelo coeficiente angular da regressão linear simples e da regressão Tobit considerando a variável latente e os efeitos marginais associados as observações censuradas e não censuradas, pela regressão dos fenótipos nos valores genéticos genômicos. O modelo de regressão Tobit ou regressão censurada é definido por:

$$y_i = \begin{cases} y_i^* = \mathbf{x_i}\boldsymbol{\beta} + \varepsilon_i, \text{se } y_i^* < t_c \\ t_c, \text{se } y_i^* \geq t_c \end{cases},$$

com: $i = 1,2,\ldots,n$, n o número de observações e $\varepsilon_i \sim N(0, \sigma^2)$, sendo: y_i^* uma variável latente não observada, y_i a variável dependente observada, $\mathbf{x_i}$ um vetor de variáveis regressoras, $\boldsymbol{\beta}$ um vetor de parâmetros desconhecidos e t_c o tempo de censura. Para obter o efeito marginal no valor esperado para **y**, associado as observações censuradas e não censuradas, o coeficiente angular $\boldsymbol{\beta}$ foi multiplicado pela probabilidade de que a variável resposta seja maior ou igual do que o tempo de censura, ou seja, $\partial E(y|\mathbf{x})/\partial x_j = \beta_j \times \Phi(t_c - \bar{\mathbf{x}}'\boldsymbol{\beta}/\sigma)$, em que: Φ é a função de distribuição acumulada da normal padrão (TOBIN, 1958; LONG, 1997).

Como medida para avaliar o desvio entre os valores estimados de acurácia, com base nos fenótipos e nos GEBVs e com base nos TBVs e GEBVs, foi utilizada a medida diferença

relativa, dada por: $\Delta = 100 \times \left|\mathrm{Ac}_{(\widehat{\mathrm{GEBV}},y_c)} - \mathrm{Ac}_{(\widehat{\mathrm{GEBV}},\mathrm{TBV})}\right| / \left|\mathrm{Ac}_{(\widehat{\mathrm{GEBV}},\mathrm{TBV})}\right|$, adaptada de Baccan et al. (2001).

3 RESULTADOS E DISCUSSÃO

Os valores estimados para a acurácia dos modelos, com base na correlação de Pearson (CP), correlação maximal (CM) e correlação de Pearson para dados censurados (CPC) em diferentes cenários são apresentados na Tabela 1. Observaram-se valores positivos e negativos, respectivamente, para a acurácia dos modelos normal truncado e misto de Cox, quando estimadas via correlação de Pearson. Segundo Hou et al. (2009), o sinal destes valores foi devido ao fato de que, no modelo misto de Cox é modelado o risco de ocorrência do evento, já no modelo normal truncado, modelou-se diretamente o tempo até a ocorrência do evento. Assim, os valores genéticos genômicos preditos pelo modelo misto de Cox e pelo modelo normal truncado, foram inversamente proporcionais, e apresentaram escalas diferentes, já que os valores genéticos genômicos estimados pelo modelo misto de Cox foram relacionados ao risco de ocorrência do evento de interesse (HOU et al., 2009; SANTOS et al., 2015). Como a correlação maximal só pode assumir valores no intervalo fechado de zero a um, foram percebidos valores positivos e de magnitude semelhante para os dois modelos na maioria dos cenários avaliados.

A confiabilidade da predição genômica esperada para as herdabilidades de 0,07 e 0,27, de acordo com a formula teórica de acurácia, foram de, respectivamente, 0,41±0,05 e 0,65±0,06. Assim, as medidas de correlação que apresentaram valores mais próximos ao esperado, foram consideradas as melhores medidas. Nos cenários em que a herdabilidade de QTL foi igual a herdabilidade da característica (C1 e C3), observou-se que na maioria das vezes, as medidas CM e CPC apresentaram estimativas estatisticamente iguais, sendo estas diferentes das obtidas para CP, ou apenas a medida CPC diferiu significativamente de CP. As únicas exceções ocorreram para C3 com 70% de censura, em que a medida CPC foi estatisticamente superior as medidas CP e CM. A CM, estimada com base nos valores preditos pelo modelo normal truncado, para C1 com 10% de censura mostrou-se estatisticamente superior as outras medidas. Com a herdabilidade de QTL igual a zero, foram observadas poucas diferenças significativas entre as medidas de acurácia. Para a característica C2, a única diferença significativa encontrada foi entre as medidas CM e CP, para o modelo normal truncado com 70% de censura. Para a característica C4 com 70% de censura, considerando-se os GEBVs

obtidos pelos modelos normal truncado e de Cox, notou-se que a medida CPC diferiu estatisticamente da medida CP, sendo as estimativas obtidas pelas medidas CM e CP estatisticamente iguais. Com 10% de censura, a CM mostrou-se estatisticamente superior às demais medidas, apenas para o modelo normal truncado. De modo geral, as estimativas obtidas pela CP foram menores que as estimadas para a CM e CPC, e estas por sua vez, apresentaram valores similares na maioria dos cenários.

Tabela 1 - Valores médios estimados para a acurácia (Ac) dos modelos misto de Cox e normal truncado, considerando os fenótipos incompletos (y_c), para 10, 40 e 70% de censura, utilizando-se das correlações de Pearson (CP), maximal (CM) e de Pearson para dados censurados (CPC), como medida de correlação entre os fenótipos e os valores genéticos genômicos estimados (GEBVs) e para a correlação de Pearson e maximal entre verdadeiros valores genéticos genômicos (TBVs) e GEBVs, com 0% de censura.

$Ac = cor(\widehat{GEBV}, y_c)/\sqrt{h^2}$												
PC	C1						C2					
	Modelo misto de Cox			Modelo normal truncado			Modelo misto de Cox			Modelo normal truncado		
	CP	CM	CPC	CP	CM	CPC	CP	CM	CPC	CP	CM	CPC
10%	-0,30 Aa	0,32 Aa	-0,30 Aa	0,30 Aa	0,34 Ba	0,31 Aa	-0,13 Aa	0,19 Aa	-0,13 Aa	0,12 Aa	0,16 Aa	0,12 Aa
40%	-0,27 Aa	0,29 Bba	-0,29 Ba	0,27 Aa	0,31 Aba	0,30 Ba	-0,10 Aa	0,16 Aa	-0,11 Aa	0,10 Aa	0,16 Aa	0,11 Aa
70%	-0,20 Ab	0,26 Ba	-0,27 Ba	0,21 Ab	0,25 Bb	0,27 Ba	-0,09 Aa	0,16 Aa	-0,12 Aa	0,07 Aa	0,15 Ba	0,10 ABa
	C3						C4					
10%	-0,44[a] Aa	0,44[a] ABa	-0,45[a] Ba	0,46[b] Aa	0,46[b] Aba	0,46[b] Ba	-0,19 Aa	0,20 Aa	-0,19 Aa	0,20 Aa	0,21 Ba	0,20 Aa
40%	-0,40[a] Ab	0,41[a] ABb	-0,43[a] Bb	0,41[b] Ab	0,43[b] Bb	0,44[b] Bb	-0,18 Aa	0,20 Aa	-0,20 Aa	0,19 Aa	0,20 Aa	0,21 Aa
70%	-0,31[a] Ac	0,33[a] Bc	-0,39[a] Cb	0,32[b] Ac	0,34[b] Bc	0,41[a] Cb	-0,16 Aa	0,17 ABa	-0,20 Ba	0,16 Ab	0,18 Ab	0,20 Ba
$Ac = cor(\widehat{GEBV}, TBV)$												
	C1						C2					
0%	-0,38	0,38	-0,38	0,41	0,41	0,41	-0,14	0,15	-0,14	0,13	0,14	0,13
	C3						C4					
0%	-0,54	0,54	-0,54	0,56	0,57	0,56	-0,19	0,20	-0,19	0,21	0,21	0,21

Fonte: Autores. PC: Porcentagem de censura; C1 – h^2 = 0,07 e h^2 de QTL = 0,07; C2 – h^2 = 0,07 e h^2 de QTL = 0; C3 – h^2=0,27 e h^2 de QTL = 0,27 e C4 – h^2=0,27 e h^2 de QTL = 0. Médias seguidas por uma mesma letra, maiúscula nas linhas e minúscula nas colunas, não diferem estatisticamente pelo teste t pareado a 5% de probabilidade com a correção de Bonferroni. As diferenças significativas entre os modelos de Cox e normal truncado são indicadas pela presença das letras a e b sobrescritas.

Várias medidas de correlação foram propostas ao longo dos anos com o objetivo de quantificar a dependência entre duas variáveis, em um contexto mais geral, em que a dependência pode ser linear ou não linear, considerando diferentes distribuições e tipos de relação de dependência. Reshef et al. (2011) apresentaram uma medida de dependência entre duas variáveis contínuas, chamada coeficiente de informação maximal (MIC) que, com tamanho amostral suficiente, é capaz de capturar associações funcionais e não funcionais. Ao comparar o MIC com as medidas correlação maximal (RÉNYI, 1959; BREIMAN, FRIEDMAN, 1985), de Pearson (GALTON, 1888; PEARSON, 1920) e de Spearman (SPEARMAN, 1904), os autores notaram que para a relação de dependência com ruído linear, estas medidas apresentaram desempenho semelhante, com coeficientes de determinação iguais a 1. Para ruídos não lineares, tais como, cúbico, exponencial, categórico e parabólico, o MIC e a correlação maximal apresentaram resultados similares. Quando ruídos senoidais foram considerados, o MIC mostrou-se mais adequado do que todas as medidas comparadas, dentre elas, a informação mútua via KDE - *kernel density estimators* (MOON; RAJAGOPALAN; LALL, 1995) e via estimador de Kraskov (KRASKOV; STÖGBAUER; GRASSBERGER, 2004).

Santos et al. (2014) realizaram um estudo para comparar métodos estatísticos utilizados para identificar o tipo de relação de dependência entre variáveis aleatórias. Neste estudo foram avaliadas as medidas: correlação de Pearson, de Spearman, de Kendall (KENDALL, 1938), da distância (SZEKELY; RIZZO; BAKIROV, 2007), informação mútua via KDE, MIC (RESHEF et al., 2011), entre outras. Foram utilizados dados simulados e reais, sendo as comparações entre os métodos realizadas por meio da curva ROC (*Receiver Operating Characteristic*). Os autores mostraram que, as correlações de Pearson, Spearman e Kendall podem detectar apenas relações lineares ou monotônica não lineares, estritamente crescente ou estritamente decrescente, e que estas três medidas apresentam uma performance similar. Já os demais métodos, foram capazes de identificar relações lineares, e também relações não monotônicas e não funcionais. Segundo os autores, caso as hipóteses de linearidade ou de monotonicidade sejam satisfeitas, a aplicação das correlações de Spearman e Kendall devem ser preferidas, já que elas são capazes de identificar relações lineares e monotônica não lineares com alto poder.

Deebani e Kachouie (2018) avaliaram o desempenho das medidas de correlação de Pearson, de Spearman, da distância, correlação maximal e o MIC, considerando dados simulados semelhantes aos utilizados por Reshef et al. (2011), sendo estes, estendidos com a

adição de diferentes níveis de ruídos: nenhum, baixo (5%), moderado (20%) e alto (40%). Para a relação linear entre as variáveis, com nenhum ou baixo ruído, os cinco métodos apresentaram scores próximos ou iguais a um. Com o aumento do ruído, a correlação maximal mostrou-se superior ao MIC, e levemente superior às medidas correlação de Pearson, de Spearman e da distância. Para a relação exponencial, monotônica estritamente crescente, scores iguais a um foram obtidos pela correlação maximal e pelo MIC, sendo o score da correlação de Spearman igual a 0,89, nos cenários com adição de ruído. Nestes cenários, o score das correlações de Pearson e da distância foram iguais a 0,20.

A correlação de Pearson é uma medida utilizada para quantificar a dependência linear entre duas variáveis. Esta medida é frequentemente utilizada em seleção genômica para a estimação da acurácia de modelos preditivos, sendo sua aplicação adequada quando as variáveis apresentam uma relação linear e são normalmente distribuídas. As correlações de Spearman e de Kendall, são medidas que vão além de quantificar a dependência linear entre duas variáveis, elas medem a associação monotônica entre as variáveis. Estas medidas não possuem o pressuposto da linearidade, podendo ser utilizadas também para variáveis em escala ordinal, o que as tornam mais adequadas, quando o fenótipo é o tempo de sobrevivência. A correlação maximal não apresenta preferência por relações lineares ou monotônicas, e também não requer nenhum pressuposto quanto a distribuição dos dados. Ela é capaz de caracterizar completamente a independência entre duas variáveis, já que uma correlação igual a zero implica na independência das variáveis. Por fim, para a identificação de relações de dependência não lineares ou não funcionais, os métodos baseados no ranqueamento ou na teoria da informação mostraram-se mais adequados (SANTOS et al., 2014).

Os maiores valores médios em valor absoluto para acurácia foram observados para a proporção de censura de 10%. Com o aumento da porcentagem de censura, observou-se perda da capacidade preditiva, o que pode ser explicado pela redução da informação fenotípica utilizada para a predição dos GEBVs. Apenas para a característica C3, esta redução na acurácia foi significativa, nos outros cenários, não foram observadas diferenças significativas, ou a diferença foi significativa apenas entre as classes com 10 e 40% e a classe de 70%. Este fato foi percebido para as três medidas, sendo a redução na acurácia mais evidente para a correlação de Pearson. Estes resultados estão de acordo com os obtidos por Kärkkäinen e Sillanpää (2013), que ao considerar as porcentagens de censura de 20, 50 e 80%, observaram que a acurácia do modelo de limiar bayesiano, calculada pela correlação de Pearson, diminui gradualmente com o aumento da porcentagem de censura.

A correlação de Pearson foi mais afetada pela diminuição da porcentagem de censura, pelo fato de não utilizar a informação presente nas observações censuradas para estimação. Ao utilizar a correlação de Pearson e Maximal para resposta contínua, considerou-se os fenótipos das observações censuradas como sendo exatos, o que implicou na redução da informação utilizada. Embora a correlação maximal não faça uso das observações censuradas, ela é mais flexível que a CP, o que possibilita a modelagem de relações mais complexas. De acordo com Li et al. (2018), dados bivariados parcialmente observados aparecem em diferentes formas. Caso uma das variáveis apresente observações totalmente perdidas, embora estas observações não forneçam uma informação direta a respeito do coeficiente de correlação, elas são informativas quanto as distribuições marginais. Na presença de observações censuradas, e considerando que nenhuma das variáveis é completamente perdida, as observações contribuem com informação a respeito da correlação, e das distribuições marginais.

Diversas medidas de correlação foram propostas para a estimação da correlação na presença de variáveis com observações censuradas à esquerda, à direita ou intervalar. Newton e Rudel (2007) propuseram uma medida de correlação baseada em estimadores de máxima verossimilhança (ML) para análise de dados com diferentes pontos de censura à esquerda. Esta medida foi comparada com o coeficiente de correlação tau-b de Kendall (OAKES, 1982) para dados censurados, e com as correlações de Pearson, Spearman e tau de Kendall, considerando as observações censuradas iguais à metade do limite de detecção ou como observações perdidas. Os autores mostraram via estudo de simulação, que para tamanhos amostrais maiores ou iguais a 100, a medida de correlação baseada em ML cujo os desvios-padrão foram estimados separadamente, apresenta os melhores resultados, podendo ser utilizada para proporções de censura variando de 60 a 90%, sendo este desempenho dependente do valor paramétrico da correlação. A maioria das medidas avaliadas foram altamente viesadas para a proporção de censura variando de 0 a 90%.

A correlação de Pearson para dados censurados, estimada via verossimilhança perfilada, proposta por Li et al. (2018) é uma medida geral de correlação que pode ser aplicada a dados com censura à esquerda, à direita, intervalar e com dados perdidos, sendo assumida para os dados uma distribuição normal ou t de Student. Os autores avaliaram o desempenho desta medida de correlação em um estudo de simulação considerando diferentes tamanhos amostrais, porcentagens de censura e distribuições subjacentes. Para dados gerados a partir da distribuição normal bivariada, foi mostrado que a probabilidade de cobertura do intervalo de confiança fornece valores satisfatórios para a censura à direita com pontos fixos ou aleatórios, e que o

estimador proposto é não viesado para censura aleatória à direita. Caso as distribuições marginais subjacentes tenham caudas mais pesadas que a da distribuição normal, e as correlações sejam altas, a utilização da distribuição t de Student bivariada fornece probabilidades de cobertura melhores do que as obtidas com a distribuição normal bivariada.

Nas características C1 e C3, toda variância genética foi explicada pelos QTLs, já as características C2 e C4 foram de natureza poligênica. As três medidas mostraram-se igualmente úteis quanto ao discernimento entre os modelos, ou seja, a eficiência média do modelo normal truncado relativa ao modelo misto de Cox, apresentou valores próximos para as três medidas de acurácia. Para as características C1 e C3, esses valores foram em média de 3 e 4%, respectivamente. Apenas para C2, o modelo misto de Cox mostrou-se levemente superior ao modelo normal truncado. Neste caso, a eficiência relativa do método normal truncado foi em média 6% inferior à do modelo de Cox. Para a característica C4, também poligênica, o resultado anterior não se repetiu, o modelo normal truncado mostrou-se 6% mais eficiente que o modelo de Cox. De modo geral, os modelos misto de Cox e normal truncado apresentaram um desempenho muito semelhante.

As acurácias calculadas com base nos GEBVs e TBVs (dados completos) foram menores que as acurácias esperadas, e, na maioria das vezes, maiores que as acurácias estimadas com base nos fenótipos incompletos. Em estudos envolvendo dados reais censurados, as únicas informações disponíveis a respeito da característica são o fenótipo e o status de censura. Nestes casos, para avaliar a acurácia dos modelos, geralmente, é estimada a correlação de Pearson entre os valores fenotípicos e os valores genéticos genômicos preditos. Ao se analisar dados simulados dispomos dos verdadeiros valores genéticos (TBVs), e ao correlacionarmos os TBVs com os GEBVs, podemos comparar os valores estimados, com os valores simulados. Neste estudo, ao se correlacionar os valores genéticos genômicos preditos com os TBVs e com os fenótipos incompletos, pôde-se medir o quanto as correlações estimadas com os fenótipos incompletos se distanciaram da estimada com os TBVs, por meio do cálculo da medida de diferença relativa (Δ).

Na Tabela 2 são apresentados os valores obtidos para a diferença relativa em cada um dos cenários. Na maioria das vezes, para os modelos misto de Cox e normal truncado, os resultados mostraram que a correlação de Pearson para dados censurados apresentou uma menor diferença relativa, quando comparada com as correlações de Pearson e maximal. Em todos os cenários, a diferença relativa das medidas CPC e CM foram menores do que as obtidas para CP. As diferenças relativas médias calculadas foram de 12,45, 13,95% e 12,68%,

respectivamente, para CPC, CM e CP, com 10% de censura. Com o aumento da censura, a diferença média associada à CP aumentou para 22,78% e 37,88%, respectivamente, para 40% e 70% de censura, enquanto para CPC a diferença relativa média aumentou para 16,80% e 20,63%. A CM apresentou valores de distância relativa média semelhantes aos obtidos para a medida CPC, sendo uma diferença maior encontrada para 10% de censura. Os modelos misto de Cox e normal truncado são modelos que conhecidamente apresentam um desempenho melhor para características poligênicas, do que para características que são governadas por genes com efeitos de moderado a grande. Este fato, provavelmente justifica uma menor diferença relativa para as características C2 e C4. De modo geral, os valores de acurácia estimados com base na medida CPC, foram mais próximos dos valores estimados com base nos TBVs, do que os estimados com as medidas CP e CM.

Tabela 2 - Valores médios estimados para a diferença relativa (Δ) em porcentagem, considerando as acurácias estimadas com base nos fenótipos incompletos (y_c) e nos verdadeiros valores genéticos genômicos (TBV), fundamentando-se nas correlações de Pearson (CP), maximal (CM) e de Pearson para dados censurados (CPC), considerando os valores genéticos genômicos estimados pelos modelos misto de Cox e normal truncado, para as características fenotípicas com 10, 40 e 70% de censura.

	Δ											
PC	C1						C2					
	Modelo misto de Cox			Modelo normal truncado			Modelo misto de Cox			Modelo normal truncado		
	CP	CM	CPC	CP	CM	CPC	CP	CM	CPC	CP	CM	CPC
10%	21,05	15,79	21,05	24,39	17,07	24,39	7,14	26,67	7,14	7,69	14,29	7,69
40%	28,95	23,68	23,68	34,15	24,39	26,83	28,57	6,67	21,43	23,08	14,29	15,38
70%	47,37	31,58	28,95	48,78	39,02	34,15	35,71	6,67	14,29	46,15	7,14	23,08
	C3						C4					
10%	18,52	18,52	16,67	17,86	19,30	17,86	0	0	0	4,76	0	4,76
40%	25,93	24,07	20,37	26,79	24,56	21,43	5,26	0	5,26	9,52	4,76	0
70%	42,59	38,89	27,78	42,86	40,35	26,79	15,79	15	5,26	23,81	14,29	4,76

Fonte: Autores. PC: Porcentagem de censura; C1 – h^2 = 0,07 e h^2 de QTL = 0,07; C2 – h^2 = 0,07 e h^2 de QTL = 0; C3 – h^2=0,27 e h^2 de QTL = 0,27 e C4 – h^2=0,27 e h^2 de QTL = 0.

A Tabela 3 apresenta os valores médios estimados para o coeficiente angular do modelo de regressão linear simples, o efeito marginal relativo às observações censuradas e não censuradas e o coeficiente angular relativo a variável latente para a regressão Tobit, considerando todos os cenários avaliados, para os GBVs estimados pelos modelos de Cox e normal truncado. Os coeficientes estimados para os modelos de Cox e normal truncado diferiram nitidamente quanto a magnitude e o sinal das estimativas. Este fato foi devido a

relação inversa, e a diferença de escala existente entre os GBVs estimados pelos dois modelos. Observou-se, também, que os coeficientes angulares estimados pela regressão dos TBVs nos GEBVs com base nos dados completos foram maiores ou iguais aos valores estimados na presença de observações censuradas em todos os cenários.

Tabela 3 - Valores médios estimados para os coeficientes angulares obtidos pela regressão dos fenótipos incompletos (y_c), nos valores genéticos genômicos estimados (GEBV) via modelos misto de Cox e Normal truncado, pelas regressões linear simples e Tobit, considerando diferentes valores de herdabilidade e herdabilidade de QTL. Para a regressão Tobit, o primeiro valor representa o efeito marginal associado as observações censuradas e não censuradas, e o segundo valor o efeito associado a variável latente.

$b_{(y_c,\widehat{GEBV})} = reg(y_c \sim \widehat{GEBV})$								
PC	C1				C2			
	Modelo misto de Cox		Modelo normal truncado		Modelo misto de Cox		Modelo normal truncado	
	LR	TB	LR	TB	LR	TB	LR	TB
10%	-4,08	-4,09; -4,48	0,63	0,63; 0,69	-2,26	-2,25; -2,46	0,26	0,26; 0,29
40%	-1,96	-1,95; -3,09	0,44	0,45; 0,71	-1,20	-1,23; -1,94	0,18	0,19; 0,30
70%	-0,81	-0,83; -2,51	0,23	0,23; 0,69	-0,95	-0,82; -2,50	0,09	0,09; 0,29
	C3				C4			
10%	-2,99	-3,03; -3,22	0,87	0,88; 0,94	-2,41	-2,41; -2,56	0,60	0,60; 0,64
40%	-1,95	-1,94; -2,85	0,64	0,64; 0,94	-1,53	-1,51; -2,20	0,46	0,45; 0,66
70%	-0,87	-0,87; -2,35	0,35	0,34; 0,93	-0,73	-0,71; -1,88	0,27	0,27; 0,70
$b_{(TBV,\widehat{GBV})} = reg(TBV \sim \widehat{GEBV})$								
	C1				C2			
0%	-5,23	-5,23	0,71	0,71	-2,80	-2,80	0,30	0,30
	C3				C4			
0%	-3,27	-3,27	0,92	0,92	-2,61	-2,61	0,61	0,61

Fonte: Autores. PC: Porcentagem de censura; C1: h² = 0,07 e h² de QTL = 0,07; C2: h² = 0,07 e h² de QTL = 0; C3: h²=0,27 e h² de QTL = 0,27 e C4: h²=0,27 e h² de QTL = 0.

O coeficiente angular e o coeficiente associado ao efeito marginal estimados, respectivamente, pela regressão linear simples e pela regressão Tobit, apresentaram valores próximos em todos os cenários avaliados. Estes coeficientes mostraram-se inversamente relacionados a proporção de censura para o modelo normal truncado, e diretamente relacionado a proporção de censura para o modelo misto de Cox, ou seja, no modelo normal truncado, à medida que a censura aumenta, o coeficiente diminui, já no modelo misto de Cox, notou-se um aumento no valor do coeficiente, com o aumento da censura. O coeficiente angular estimado com base na variável latente, apresentou um comportamento nitidamente crescente, com o

aumento da censura, apenas para o modelo misto de Cox nos cenários C1, C3 e C4. De modo geral, os coeficientes estimados para o modelo normal truncado com base na variável latente sofreram menos variação com o aumento da censura, enquanto os estimados pela regressão linear e com base nos efeitos marginais foram drasticamente influenciados pela censura, principalmente com a mudança de 40 para 70% de censura. Estes dois últimos coeficientes, mostraram-se ser mais adequados para a avaliação do viés na presença de observações censuradas, pois com a redução de informação fenotípica utilizada para a estimação dos GBVs, espera-se que o viés da predição varie consideravelmente, já que a censura pode atingir até 70% das observações. Segundo Long (1997), o modelo de regressão Tobit faz uso de toda informação disponível, o que inclui as observações censuradas, assim, o modelo é capaz de fornecer estimativas consistentes, enquanto o modelo de regressão linear, ao tratar todos os valores fenotípicos como observados, fornece estimativas inconsistentes para os parâmetros.

As estimativas dos parâmetros no modelo de regressão linear simples são obtidas via mínimos quadrados ordinais, com a suposição de linearidade nos dados, normalidade e homogeneidade dos resíduos. Já o modelo de regressão Tobit é baseado em estimadores de máxima verossimilhança, e possui os mesmos pressupostos que a regressão linear simples. Supondo que o pressuposto da normalidade seja atendido, Amemiya (1973) apresentou resultados a respeito da consistência e da normalidade assintótica dos estimadores de ML. Lumley et al. (2002) mostrou por meio de um estudo de simulação, que para uma amostra suficientemente grande, a regressão linear simples pode apresentar um desempenho razoável mesmo para dados não normais. De acordo com Amore e Murtinu (2020), o modelo de regressão Tobit é mais sensível à quebra dos pressupostos de normalidade e de homogeneidade dos resíduos do que o modelo de regressão linear simples.

Por meio de uma inspeção gráfica, foi constatado que no presente estudo, os resíduos não foram normalmente distribuídos, e que o pressuposto da homogeneidade dos resíduos não foi aceito para todas as repetições e porcentagens de censura. Lewis e McDonald (2014) verificaram que para dados simulados com resíduos normais, com mil observações, e com 25% de censura, o modelo Tobit mostrou-se o modelo mais eficiente dentre todos os modelos considerados, sendo este, menos viesado que a regressão linear simples, e com raiz do erro quadrático médio (RMSE) aproximadamente quatro vezes menor do que o obtido para a regressão linear simples. Para resíduos com distribuições normal mista e lognormal, os dois modelos mostraram-se altamente viesados, sendo que, para a normal mista os modelos

apresentaram valores semelhantes de viés e de RMSE e para a lognormal, a regressão linear simples apresentou menores valores de viés e de RMSE.

Devido a diferença de escala existente entre os GEBVs obtidos pelos modelos de Cox e normal truncado, o viés não é uma boa medida para comparar diretamente estes modelos, o que não prejudica o objetivo do estudo de avaliar medidas alternativas para estimação do viés, em predição genômica, no contexto de dados censurados. Santos et al. (2015), utilizaram a correlação de Spearman como medida de acurácia para dados com observações censuradas, e o coeficiente Kappa como uma medida para a avaliação da concordância do ranqueamento dos indivíduos de acordo com os GBVs estimados pelo modelo misto de Cox e pelo modelo linear misto com dados completos ou imputados.

4 CONCLUSÃO

Neste estudo, consideramos dois novos métodos para estimação da acurácia, a correlação maximal e de Pearson para dados censurados, e uma nova medida para estimação do viés, a regressão Tobit, no contexto de dados censurados. Do ponto de vista estatístico, estas metodologias devem ser preferidas por lidarem de modo adequado com a presença de observações incompletas nos dados. Os resultados mostraram que as medidas propostas de correlação apresentaram estimativas significativamente superiores às apresentadas pela correlação de Pearson, principalmente para a característica com herdabilidade de QTL igual a 0,27. O coeficiente associado ao efeito marginal do modelo Tobit apresenta valores próximos aos obtidos pela regressão linear simples. O coeficiente associado à variável latente na regressão Tobit, na maioria das vezes, sofre poucas variações com as mudanças de proporções de censura, enquanto a regressão linear simples é extremamente afetada pela presença de observações censuradas.

REFERÊNCIAS

ALEMU, S. W.; CALUS, M. P. L.; MUIR, W.M.; PEETERS, K.; VEREIJKEN, A.; BIJAMA, P. Genomic prediction of survival time in a population of brown laying hens showing cannibalistic behavior. **Genetics Selection Evolution,** v. 48, n. 68, p. 1-10, 2016.

AMEMIYA, T. Regression analysis when the dependent variable is truncated normal. **Econometrica**, v. 41, n. 6, p. 997-1016, 1973.

AMORE, M. D.; MURTINU, S. Tobit models in strategy research: Critical issues and

applications. **Global Strategy Journal,** p. 1-25, 2020.

BACCAN, N.; ANDRADE, J. C.; GODINHO, O. E. S.; BARONE, J. S. **Química Analítica Quantitativa Elementar.** 3. ed. São Paulo: Editora Edgard Blucher, 2001.

BLÁZQUEZ, F. L.; MIÑO, B. S. Maximal correlation in a non-diagonal case. **Journal of Multivariate Analysis,** v. 131, p. 265-278, 2014.

BREIMAN, L.; FRIEDMAN, J. H. Estimating optimal transformations for multiple regression and correlation. **Journal of the American statistical Association,** v. 80, n. 391, p. 580-598, 1985.

BRITO, F. V.; NETO, J. B.; SARGOLZAEI, M.; COBUCI, J. A.; SCHENKEL, F. S. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics**, v. 12, n. 80, p. 1-10, 2011.

CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole genome regression and prediction methods applied to plant and animal breeding. **Genetics,** v. 193, n. 2, p. 327-345, 2013.

COSTA, E. V. **Modelos bayesianos multicaracterísticos para dados censurados na avaliação genética de características reprodutivas em bovinos nelore.** 2017. 42f. Tese (Doutorado em Zootecnia) - Curso de Pós-graduação em Zootecnia, Universidade Federal de Viçosa.

DEEBANI, W.; KACHOUIE, N. N. Ensemble Correlation Coefficient. **In:** International Symposium on Artificial Intelligence and Mathematics - ISAIM, 2018, Fort Lauderdale, FL.

FEIZI, S.; MAKHDOUMI, A.; DUFFY, K.; KELLIS, M.; MEDARD, M. Network Maximal Correlation. **arXiv.org,** 2017.

GALTON, F. Co-relations and their measurement, chiefly from anthropometric data. **Proceedings of the Royal Society of London**, v. 45, p. 135-45, 1888.

GEBELEIN, H. Das statistische problem der correlation als variation und eigenwertproblem und sein zusammenhang mit der ausgleichrechnung. **Zeitschrift für angewandte Mathematik und Mechanik,** v. 21, n. 6, p. 364-379, 1941.

GIOLO, S. R.; DEMÉTRIO, C. G. B. A frailty modeling approach for parental effects in animal breeding. **Journal of Applied Statistics**, v. 38, n. 3, p. 619-629, 2011.

HOU, Y.; MADSEN, P.; LABOURIAU, R.; ZHANG, Y.; LUND, M. S.; SU, G. Genetic analysis of days from calving to first insemination and days open in Danish Holsteins using different models and censoring scenarios. **Journal Dairy Science,** v. 92, n. 3, p. 1229-1239, 2009.

KÄRKKÄINEN, H. P.; SILLANPÄÄ, M. J. Fast Genomic Predictions via Bayesian G-BLUP and Multilocus Models of Threshold Traits Including Censored Gaussian Data. **G3,** v. 3, p. 1511-1523, 2013.

KENDALL, M. A new measure of rank correlation. **Biometrika,** v. 30, n. 1/2, p. 81-93, 1938.

KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. **Physical Review E,** v. 69, n. 6, p. 1-16, 2004.

LEWIS, R. A.; McDONALD, J. B. Partially Adaptive Estimation of the Censored Regression Model. **Econometric Reviews**, v. 33, n. 7, p. 732-750, 2014.

LI, Y.; GILLESPIE, B. W.; SHEDDEN, K.; GILLESPIE, J. A. Profile Likelihood Estimation of the Correlation Coefficient in the Presence of Left, Right or Interval Censoring and Missing Data. **The R Journal,** v. 10/2, p. 159-179, 2018.

LONG, J. S. **Regression Models for Categorical and Limited Dependent Variables.** Thousand Oaks, CA: Sage Publications, 1997. 297p.

LUMLEY, T.; DIEHR, P.; EMERSON, S.; CHEN, L. The importance of the normality assumption in large public health data sets. **Annual Review of Public Health,** v. 23, n. 1, p. 151-169, 2002.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics,** v. 157, n. 4, p. 1819-1829, 2001.

MOON, Y.; RAJAGOPALAN, B.; LALL, U. Estimation of mutual information using kernel density estimators. **Physical Review E,** v. 52, n. 3, p. 2318-2321, 1995.

NEWTON, E.; RUDEL, R. Estimating correlation with multiply censored data arising from the adjustment of singly censored data. **Environmental science & technology,** v. 41, n. 1, p. 221-228, 2007.

OAKES, D. A concordance test for independence in the presence of censoring. **Biometrics**, v. 38, p. 451-455, 1982.

PALAIOKOSTAS, C.; FERRARESSO, S.; FRANCH, R.; HOUSTON, R. D.; BARGELLONI, L. Genomic Prediction of Resistance to Pasteurellosis in Gilthead Sea Bream (Sparus aurata) Using 2b-RAD Sequencing. **G3,** v. 6, p. 3693-3700, 2016.

PEARSON, K. Notes on the history of correlation. **Biometrika,** v. 13, n. 1, p. 25-45, 1920.

PÉREZ, P.; CAMPOS, G. de los Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483-495, 2014.

RÉNYI, A. On measures of dependence. **Acta Mathematica Hungarica,** v. 10, n. 34, p. 441-451, 1959.

RESHEF, D. N.; RESHEF, Y. A.; FINUCANE, H. K.; GROSSMAN, S. R.; McVEAN, G.; TURNBAUGH, P. J.; LANDER, E. S.; MITZENMACHER, M.; PARDIS C. SABETI, P. C. Detecting Novel Associations in Large Datasets. **Science,** v. 334, n. 6062, p. 1518-1524, 2011.

RIPATTI, S.; PALMGREN, J. Estimation of multivariate frailty models using penalized partial likelihood. **Biometrics**, v. 56, p. 1016-1022, 2000.

SANTOS, S. S.; TAKAHASHI, D. Y.; NAKATA, A.; FUJITA, A. A comparative study of statistical methods used to identify dependencies between gene expression signals. **Briefings in Bioinformatics,** v. 15, n. 6, p. 906-918, 2014.

SANTOS, V. S.; MARTINS, F. S.; RESENDE, M. D.; AZEVEDO, C. F.; LOPES, P. S.; GUIMARÃES, S. E.; GLÓRIA, L.; SILVA, F. F. Genomic selection for slaughter age in pigs using the Cox frailty model. **Genetics and Molecular Research,** v. 14, n. 4, p. 12616-12627, 2015.

SARGOLZAEI, M.; SCHENKEL, F. S. QMSim: A large-scale genome simulator for livestock. **Bioinformatics**, v. 25, n. 5, p. 680-681, 2009.

SMITH, B. J. boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. **Journal of Statistical Software,** v. 21, p. 1-37, 2007.

SPEARMAN, C. "General intelligence", objectively determined and measured. **The American Journal of Psychology,** v. 15, n. 2, p. 201-92, 1904.

SPECTOR, P.; FRIEDMAN, J.; TIBSHIRANI, R.; LUMLEY, T. acepack: ACE and AVAS methods for choosing regression transformations. **R package version 1.4.1,** 2016.

SZEKELY, G.; RIZZO, M.; BAKIROV, N. Measuring and testing dependence by correlation of distances. **The Annals of Statistics,** v. 35, n. 6, p. 2769-2794, 2007.

THERNEAU, T. M.; GRAMBSCH, P. M.; PANKRATZ, V. S. Penalized survival models and frailty. **Journal of Computational and Graphical Statistics,** v. 12, p. 156-175, 2003.

THERNEAU, T. M. coxme: Mixed Effects Cox Models. R package version 2.2-16, 2020. Disponível em: https://cran.r-project.org/web/packages/coxme/index.html. Acesso em: 22 jul. 2020.

TOBIN, J., Estimation of Relationships for Limited Dependent Variables. **Econometrica,** v. 26, n. 1, p. 24-36, 1958.

VALLEJO, R. L.; LEEDS, T. D.; FRAGOMENI, B. O.; GAO, G.; HERNANDEZ, A. G.; MISZTAL, I.; WELCH, T. J.; WIENS, G. D.; PALTI, Y. Evaluation of Genome-Enabled Selection for Bacterial Cold-Water Disease Resistance Using Progeny Performance Data in Rainbow Trout: Insights on Genotyping Methods and Genomic Prediction Models. **Frontiers in Genetics**, v. 7, p. 1-13, 2016.

VanRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Sciences,** v. 91, p. 4414-4423, 2008.

WIENTJES, Y. C. J.; VEERKAMP, R. F.; CALUS, M. P. L. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. **Genetics**, v. 193, n. 2, p. 621-631, 2013.

CAPÍTULO III

SELEÇÃO GENÔMICA AMPLA VIA MÉTODOS DE APRENDIZADO DE MÁQUINA E DE REDUÇÃO DE DIMENSIONALIDADE COM DADOS CENSURADOS

RESUMO

O objetivo deste estudo foi avaliar a aplicabilidade das metodologias Random Survival Forest (RSF), Gradient Boosted Machine (GBM) e Análise de Componentes Principais Supervisionados (SPCA) em seleção genômica ampla, comparando-as com o método Regressão Ridge Bayesiana (BRR). Foram utilizados dados de uma população de 777 juvenis de douradas (*Sparus aurata*), genotipados por 12085 marcadores SNPs. A variável resposta foi o tempo em dias decorridos do início do experimento até a morte pela doença *Pasteurellosis*. Ao término do experimento, animais que não falharam foram considerados como observações censuradas. Os modelos foram comparados pelas medidas AUC (*Area under the Receiver Operating Characteristic Curve*), BS (*Brier-Score*), $\mathrm{DAUC} = |\mathrm{AUC} - 0{,}50|$, correlação de Spearman e pela proporção de indivíduos selecionados, estimados via validação cruzada 7 - fold, e também pela localização de SNPs ou grupos de ligação relevantes. Os resultados mostraram que, em termos de habilidade preditiva, os modelos RSF (AUC: 0,5777±0,0576, BS: 0,0442±0,0034, DAUC: 0,0806±0,0528) e BRR (AUC: 0,4141±0,0425, BS: 0,0419±0,0037, DAUC: 0,0859±0,0426) foram semelhantes. O modelo GBM apresentou o maior valor para a medida BS (0,26±0,1231), sendo este estatisticamente igual ao valor estimado para o SPCA, e diferente dos estimados para a RSF e BRR. O modelo SPCA mostrou o menor valor para DAUC (0,0347±0,0253), no entanto, este valor só diferiu estatisticamente do valor estimado para a BRR. Quanto a localização dos top-40 SNPs, observou-se que na maioria das vezes eles se encontram em um mesmo grupo de ligação. O modelo RSF foi o que apresentou maior interseção com o rank dos top-40 SNPs obtido pelos métodos BRR (5 SNPs) e modelo misto de Cox (17 SNPs). Foi estimada uma correlação de Spearman alta (-0,7538±0,0456) entre os GEBVs estimados via BRR e as probabilidades de ocorrência do evento preditas pelo modelo RSF. Além do mais, as maiores porcentagens de concordância no ranqueamento dos 20% e 40% indivíduos selecionados foram obtidas entre os métodos GBM e RSF (20% - 0,7180±0,1039), RSF e BRR (20% - 0,6290±0,0794), RSF e BRR (40% - 0,7732±0,0319) e,

GBM e RSF (40% - 0,6937±0,0833). A utilização de diferentes densidades de SNPs selecionados pelos métodos RSF, GBM e SPCA, não implicou em uma diferença significativa na habilidade preditiva do modelo misto de Cox. Os resultados sugerem que a RSF, constitui a alternativa com maior potencial para aplicação em estudos de seleção genômica com observações censuradas, dentre os métodos considerados.

Palavras-chave: seleção genômica ampla, métodos de aprendizado de máquina, observações censuradas, valores genéticos genômicos, modelo misto de Cox.

ABSTRACT

The aim of this study was to evaluate the applicability of the Random Survival Forest (RSF), Gradient Boosted Machine (GBM) and Supervised Principal Component Analysis (SPCA) methodologies in genomic selection, comparing them with the Bayesian Ridge Regression (BRR) method. Data from a population of 777 sea bream juveniles were used, genotyped by 12085 SNPs markers. The response variable was the time in days from the beginning of the experiment until death due to *Pasteurellosis*. At the end of the experiment, animals that did not die were considered censored observations. The models were compared using the AUC (*Area under the Receiver Operating Characteristic Curve*), BS (*Brier-Score*), $\mathrm{DAUC} = |\mathrm{AUC} - 0{,}50|$, Spearman correlation and the proportion of selected individuals, estimated via 7 - fold cross validation, and also by the location of relevant SNPs or link groups. The results showed that, in terms of predictive ability, the RSF (AUC: 0.5777 ± 0.0576, BS: 0.0442 ± 0.0034, DAUC: 0.0806 ± 0.0528) and BRR (AUC: 0.4141 ± 0.0425, BS: 0.0419 ± 0.0037, DAUC: 0.0859 ± 0.0426) were similar. The GBM model presented the highest value for the BS measure (0.26 ± 0.1231), which is statistically equal to the estimated value for the SPCA, and different from those estimated for the RSF and BRR. The SPCA model showed the lowest value for DAUC (0.0347 ± 0.0253), however this value only differed statistically from the estimated value for BRR. As for the location of the top-40 SNPs, it was observed that in most cases they are in the same link group. The RSF model was the one that presented the greatest intersection with the rank of the top-40 SNPs obtained by the BRR methods (5 SNPs) and Cox's mixed model (17 SNPs). A high Spearman correlation (-0.7538 ± 0.0456) was estimated between the GEBVs estimated via BRR and the event probabilities predicted by the RSF model. Furthermore, the highest percentages of agreement in the ranking of the 20% and

40% selected individuals were obtained between GBM and RSF (20% - 0.7180 ± 0.1039), RSF and BRR (20% - 0.6290) ± 0.0794), RSF and BRR (40% - 0.7732 ± 0.0319) and, GBM and RSF (40% - 0.6937 ± 0.0833). The use of different SNP densities selected by the RSF, GBM and SPCA methods, did not imply in significant difference in the predictive ability of Cox's mixed model. The results suggest that RSF is the alternative with the greatest potential for application in studies of genomic selection with censored observations, among the methods considered.

Key words: genomic wide selection, machine learning methods, censored records, genomic breeding values, mixed Cox model.

1 INTRODUÇÃO

As doenças infecciosas representam um grande obstáculo à produção para muitas das espécies utilizadas na aquicultura, pois causam altas taxas de mortalidade, reduzindo o crescimento das populações de peixes e moluscos (HOUSTON, 2017; LAFFERTY et al., 2015). Para medir como uma determinada espécie responde a um patógeno específico, pode-se realizar o teste de desafio à doença. Este teste se baseia em expor os animais, em ambientes controlados, à presença do patógeno, sendo registradas diariamente ou com uma frequência maior, as taxas de mortalidade. A característica fenotipada por meio destes testes apresenta em geral, uma herdabilidade de moderada a alta, o que a torna um alvo interessante para programas de melhoramento animal, devido ao seu grande potencial de ganho genético para a resistência a doenças (ODEGARD et al., 2011).

Os fenótipos assim obtidos, caracterizam-se como dados censurados, pois ao término do experimento, uma parte dos animais pode não experimentar o evento de interesse, sendo caracterizados como observações incompletas. Estes dados são definidos por um par de variáveis, a primeira indicando o tempo até a morte do indivíduo (tempo de sobrevivência), e a segunda, indicando se o indivíduo veio a óbito ou não (variável indicadora de censura). Estudos de seleção e de associação genômica ampla vêm sendo conduzidos em aquicultura, visando, respectivamente, comparar a acurácia de modelos de avaliação genômica e/ou identificar variantes genéticas associadas ao tempo de sobrevivência e/ou a variável indicadora de censura (PALAIOKOSTAS et al., 2016; BARRÍA et al., 2018; PALAIOKOSTAS et al., 2018; ROBLEDO et al., 2018).

Diversos modelos têm sido utilizados para a predição de valores genéticos genômicos (GEBVs) e estimação de componentes de variância, para a característica tempo de sobrevivência em aquicultura. Palaiokostas et al. (2016) fizeram uso dos métodos bayesianos, Regressão Ridge Bayesiana (BRR), BayesA, BayesB e BayesC, para a modelagem da resistência à doença Pasteurellosis em douradas - gilthead sea bream (*Sparus aurata*). Estes métodos encontram-se implementados no pacote BGLR (PÉREZ; CAMPOS, 2014) do software R, e modelam os dados censurados como amostrados de uma distribuição normal truncada. Barría et al. (2018) avaliaram separadamente o tempo até a morte e a variável indicadora de falha (morte), devido à doença *Piscirickettsia salmonis* em uma população de salmão coho, por meio dos métodos: G-BLUP, ssGBLUP e WssGBLUP, implementados pela família de programas BLUPf90 (MISZTAL et al., 2016). Saura et al. (2019), utilizaram análise de sobrevivência e modelos lineares, para elucidar as bases genéticas da resistência e tolerância à escuticociliatose em pregados (turbot), baseando-se em informações genômicas e genealógicas. Para estimar os componentes de variância das características, Saura et al. (2019) utilizaram o modelo misto de riscos proporcionais, considerando as abordagens semi-paramétrica (Cox) e paramétrica (Weibull) pelo software Survival Kit V.6.1 (MÉSZÁROS et al., 2013).

Métodos de aprendizado de máquina e de redução de dimensionalidade, tais como, *Random Forest* (RF), *Gradient Boosting Machine* (GBM) e Análise de Componentes Principais (PCA), vem sendo aplicados para a predição de GEBVs e identificação de SNPs importantes, para características de natureza binária e contínua em programas de melhoramento animal (NADERI et al., 2016; DU et al., 2018; LI et al., 2018). Li et al. (2018) avaliaram a utilização de subconjuntos de SNPs, selecionados pelos métodos de aprendizado de máquina RF, GBM e XgBoost, para a predição genômica via modelo G-BLUP, avaliando impacto de diferentes densidades de marcadores moleculares na acurácia e na estimação de componentes de variância. Em aquicultura, estes métodos têm sido pouco explorados para características contínuas, e ainda é um campo inexplorado para características com observações incompletas, tais como dados censurados. Métodos como Random Survival Forest (RSF), GBM e Análise de Componentes Principais Supervisionados (SPCA) podem ser utilizados para a predição em dados onde a resposta é o tempo de sobrevivência, contudo estes métodos não têm sido utilizados para a predição em estudos de seleção genômica ampla.

Considerando o exposto, este trabalho tem por objetivo principal, comparar as metodologias RSF, GBM, SPCA e BRR na avaliação genômica do tempo de sobrevivência de

douradas acometidos pela doença Pasteurellosis. Adicionalmente, pretende-se: identificar os SNPs de maior relevância na predição de sobrevivência dos peixes via métodos de aprendizado de máquina, redução de dimensionalidade e GWAS para dados censurados via modelo misto de Cox, e avaliar a utilização de subconjuntos de SNPs, selecionados pelos métodos RSF, GBM e SPCA no modelo misto de Cox para a predição de valores genéticos genômicos.

2. MATERIAL E MÉTODOS

2.1 Descrição dos dados

Neste estudo foram utilizadas informações fenotípicas e genotípicas de uma população de dourada (*Sparus aurata*), disponibilizadas online por Palaiokostas et al. (2016). Os dados podem ser obtidos como tabelas S1 (fenótipos), S5 (localização dos SNPs no mapa de ligação) e S6 (genótipos) em https://www.g3journal.org/content/6/11/3693. Estes dados constituem um subconjunto do grupo de juvenis de douradas, com os quais Antonello et al. (2009) realizaram um experimento para estimar a herdabilidade da característica fenotípica resistência à doença Pasteurellosis. Os peixes foram provenientes do cruzamento fatorial entre 32 machos e 35 fêmeas, o que originou 825 juvenis. Deste total, 48 animais foram excluídos por apresentarem uma taxa de dados perdidos maior que 30%, restando assim, 777 juvenis.

O fenótipo consiste nos registros diários de mortalidade dos juvenis devido à doença Pasteurellosis, sendo registrado o tempo até a morte dos juvenis. Animais encontrados vivos após 19 dias foram considerados como observações censuradas. Deste modo, a informação fenotípica foi representada pelo par, tempo de sobrevivência e variável indicadora de censura. Os animais foram genotipados por 12085 marcadores SNPs, distribuídos por 24 grupos de ligação. O controle de qualidade dos marcadores SNPs foi realizado considerando uma call rate de 90% e uma menor frequência alélica (MAF) de 10%. Após o controle de qualidade, 7833 marcadores SNPs foram mantidos para as análises posteriores.

2.2 Métodos estatísticos

2.2.1 *Random Survival Forest*

Proposta por Ishwaran et al. (2008), a Random Survival Forest (RSF) é uma metodologia de aprendizado de máquina não-paramétrica, obtida pela generalização do método *Random Forest* (BREIMAN, 2001), para a modelagem de dados de sobrevivência. Inicialmente, o conjunto total de SNPs é subdividido de forma a se obter a melhor divisão do conjunto, com base em algum critério de divisão (log-rank, log-rank score, entre outros). Em seguida, cada árvore é construída usando uma amostra bootstrap dos indivíduos em estudo.

O algoritmo utilizado para implementação da RSF pode ser sumarizado da seguinte forma: (1) primeiro são retiradas aleatoriamente B (número de árvores - ntree) amostras bootstrap do conjunto de dados original, de modo que cada amostra contenha aproximadamente dois terços do total de indivíduos. O terço restante é denominado dado *out-of-bag* (OOB), e é utilizado para a predição e validação; (2) a árvore de decisão binária é construída para cada amostra bootstrap, pela seleção aleatória de um subconjunto de covariáveis (mtry), em cada um dos nós da árvore. A divisão do conjunto de variáveis é realizada por meio de um critério de divisão para dados censurados, de modo a maximizar as diferenças de sobrevivência entre dois nós vizinhos (filhos); (3) cada árvore cresce até o tamanho máximo, de modo que o nó terminal atinja o número mínimo de eventos com tempos de sobrevivência exclusivos; (4) calcula-se uma estimativa para o risco acumulado, combinando as informações de todas as árvores. Uma estimativa para o risco é gerada para cada indivíduo presente nos nós terminais; (5) estima-se a taxa de erro, computada a partir dos dados OOB (ISHWARAN et al., 2008).

Como critério para a divisão do conjunto de observações em cada nó da árvore de sobrevivência binária foi utilizado a estatística de log-rank. Seja i denotando um indivíduo, com $i = \{1, \dots, n\}$, em que n denota o número total de observações em um dado nó h. Considerando o SNP1, denotado por x, a regra de divisão se baseia em encontrar o ponto de corte c, tal que: $x \leq c$ (indivíduo é direcionado ao nó 1) ou $x > c$ (indivíduo é direcionado ao nó 2), que maximiza a diferença de sobrevivência entre os indivíduos dos dois nós. Este critério se fundamenta na estatística log-rank, dada por:

$$L(x, c) = \frac{\sum_{i=1}^{N} \left(d_{i,1} - Y_{i,1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{Y_{i,1}}{Y_i} \left(1 - \frac{Y_{i,1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}},$$

em que: $d_{i,j}$ e $Y_{i,j}$ são respectivamente, o número de mortes e o número de indivíduos sob risco no tempo t_i nos nós filhos $j = 1, 2$ e N o número de tempos de falhas distintos $t_1 < t_2 < \cdots < t_N$. Neste critério, o valor $|L(x, c)|$ é utilizado como medida para a divisão em cada um dos nós,

quanto maior o valor, maior será a diferença entre os grupos (ISHWARAN et al., 2008). O processo foi repetido para todas as variáveis, até que fosse encontrada uma variável x* e o respectivo ponto de corte c*, tais que: $|L(x^*,c^*)| \geq |L(x,c)|$, para qualquer variável x e ponto de corte c. Este procedimento foi repetido em cada nó, até que os nós terminais fossem alcançados.

Considerando, sem perda de generalidade o nó h. A taxa de falha acumulada neste nó é calculada pelo estimador de Nelson-Aalen, dado por:

$$\widehat{\Lambda}_h(t) = \sum_{t_{i,h} \leq t} \frac{d_{i,h}}{Y_{i,h}},$$

em que: $t_{i,h}$ são tempos distintos de sobrevivência no nó h; $d_{i,h}$ número de eventos e $Y_{i,h}$ o número de indivíduos sob risco no tempo $t_{i,h}$. Deste modo, a cada nó terminal se associa uma estimativa para a taxa de falha acumulada. Assim, é estimada uma sequência de taxas de falhas para cada árvore binária, sendo a taxa de falha acumulada de cada árvore obtida pelo agrupamento das sequências de taxa de falha.

A estimativa da taxa de falha de um indivíduo específico i (na amostra b), considerando o conjunto de variáveis explicativas (x_i), é denotada por $\widehat{\Lambda}_b(t|x_i)$, e pode ser obtida percorrendo a árvore, na tentativa de identificar em qual dos nós terminais o indivíduo está contido. Caso o indivíduo seja encontrado no nó h, a estimativa da taxa de falha acumulada será igual a $\widehat{\Lambda}_b(t|x_i) = \widehat{\Lambda}_h(t)$. A taxa de falha acumulada conjunta é obtida pela soma das taxas de falha acumuladas estimadas para cada amostra bootstrap, dividida pelo número de árvores, e pode ser obtida por:

$$\widehat{\Lambda}(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\Lambda}_b(t|x_i).$$

Os dados OOB foram utilizados para ajuste dos parâmetros da RSF, pelo cálculo da taxa de erro OOB, e para obtenção da medida de importância das variáveis (VIMP), associado a cada SNP. De acordo com Breiman (2001) e Ishwaran (2007), o cálculo do VIMP para um determinado SNP é realizado distribuindo os casos OOB dentro da bolsa (in bag) da árvore de sobrevivência. Caso seja encontrada uma divisão para o SNP, a escolha entre um dos nós filhos é feita aleatoriamente. Em seguida, é calculada a taxa de falha acumulada média das árvores em questão. Segundo os autores, o VIMP é dado pela diferença entre o erro de predição no conjunto original e o erro de predição do novo conjunto, construído via randomização. Valores

altos e positivos de VIMP estão relacionados à SNPs com maior poder preditivo, já valores negativos ou nulos estão associados à SNPs com baixo ou nenhum poder preditivo.

2.2.2 *Gradient Boosting Machine*

Gradient Boosting Machine (GBM) é uma poderosa ferramenta estatística para a construção de modelos de predição, que combina sequencialmente vários classificadores base (weak learners), com o intuito de melhorar o desempenho preditivo do modelo final (FRIEDMAN; TIBSHIRANI, 2000). Seu funcionamento se baseia em três elementos principais: (1) uma função de perda a ser minimizada; (2) um modelo de predição ou classificador base (weak learner) e (3) um modelo aditivo que combina os modelos de predição de modo a minimizar a função de perda.

O objetivo é obter uma função $\mathbf{y} = F(\mathbf{x}; \boldsymbol{\beta})$, capaz de classificar (ou mapear) as observações pertencentes ao conjunto de treinamento $T = \{x_i, y_i\}_{i=1}^{n}$, em que: $\boldsymbol{\beta}$ é o vetor de parâmetros associados a F, estimados pela minimização da função de perda $\sum_{i=1}^{n} \Phi(y_i, F(x_i; \boldsymbol{\beta}))$. É assumido para F uma expansão aditiva da forma $F(\mathbf{x}) = \sum_{m=0}^{M} \rho_m f(\mathbf{x}, \boldsymbol{\tau_m})$, sendo f um classificador base com peso $\boldsymbol{\rho}$, e vetor de parâmetros $\boldsymbol{\tau}$, de modo que $\{\rho_m, \tau_m\}_{m=1}^{M}$ componha o conjunto de parâmetros $\boldsymbol{\beta}$. Os classificadores são aprendidos segundo um processo stagewise aditivo: (1) seja $f_0(x)$ um classificador base (estimador inicial); (2) considere m o número de iterações com $m = 1, \ldots, M$, os parâmetros $\boldsymbol{\rho}$ e $\boldsymbol{\tau}$ são estimados pela expressão $(\boldsymbol{\rho}, \boldsymbol{\tau}) = \underset{\rho,\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \Phi(y_i, F_{m-1}(x_i) + \boldsymbol{\rho} f(x_i; \boldsymbol{\tau}))$. A aproximação de (2) é realizada em dois passos. Primeiro passo: é ajustada $f(x_i; \tau_m)$ via mínimos quadrados por $\tau_m = \underset{\tau}{\operatorname{argmin}} \sum_{i=1}^{n} (g_{im} - f(x_i; \tau_m))^2$, com: $g_{im} = -\left[\frac{\partial \Phi(y_i; F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$. No segundo passo, são estimados os pesos ($\boldsymbol{\rho}$) pela expressão: $\rho_m = \underset{\boldsymbol{\rho}}{\operatorname{argmin}} \sum_{i=1}^{n} \Phi(y_i, F_{m-1}(x_i) + \boldsymbol{\rho} f(x_i; \tau_m))$. Por fim, a F é atualizada pela fórmula $F_m(x) = F_{m-1}(x) + v\rho_m f(x_i; \tau_m)$, em que: v representa um parâmetro de encurtamento $0 < v \leq 1$, que tem por função evitar o super ajustamento (FRIEDMAN, 2002; CHEN et al., 2013).

Ridgeway (1999) propôs uma adaptação do GBM para a análise de dados censurados. O autor assume como função de perda o negativo do logaritmo da verossimilhança parcial, dada por:

$$\Phi(y, F) = -\sum_{i=1}^{n} \delta_i \left\{ F(x_i) - \log\left(\sum_{j:t_j \geq t_i} e^{F(x_j)} \right) \right\}.$$

A influência relativa mede a contribuição de cada SNP na minimização da função de perda. Proposta por Friedman (2001), para métodos baseados em árvore, a influência relativa aproximada de uma variável x_j, em uma dada árvore, é obtida por: $\widehat{J}_j^2 = \sum_{\text{Divisão em } x_j} \widehat{I}_t^2$, em que: $\widehat{I}_t^2$ é a melhoria empírica ao se utilizar x_j como ponto de divisão. Em seguida, calcula-se o valor médio para a influência relativa da variável x_j, considerando todas as árvores geradas pelo algoritmo boosting (RIDGEWAY, 2019).

2.2.3 Análise de Componentes Principais Supervisionados

Introduzida por Bair e Tibshirani (2004), a análise de regressão via componentes principais supervisionados ou Análise de Componentes Principais Supervisionados (SPCA), se baseia na seleção de variáveis relacionadas ao tempo de sobrevivência. Esta abordagem utiliza as variáveis significativas em uma análise de componentes principais, e a partir destas, realiza a previsão de sobrevivência dos indivíduos.

Bair et al. (2006) descrevem os passos para implementação do método SPCA. Seja $X_{n\times p}$ uma matriz composta por p variáveis, medidas em n indivíduos. Assumindo que a resposta são os tempos de sobrevivência e o status de censura, o método se fundamenta nas etapas: (1) Computar os coeficientes (β_j) do modelo de regressão univariado de Cox, $h(t) = h_0 \exp(x_j \beta_j)$, com: $j = 1, \dots, p$; (2) Obter a matriz reduzida dos dados, composta pelas variáveis cujo os coeficientes univariados ultrapassaram em valor absoluto um certo limiar θ, estimado via validação cruzada; (3) Computar os primeiros componentes principais da matriz reduzida ($c_1$variáveis mais significativas); (4) Usar os primeiros c_2 componentes $u_{\theta,l}$, no modelo de regressão de Cox, $h(t) = h_0 \exp(u_{\theta,1}\alpha_1 + \cdots + u_{\theta,c_2}\alpha_{c_2})$, em que: $u_{\theta,l} = f(x_k)$, com: $k = 1, \dots, c_1$.

Bair et al. (2006) utilizam o preditor $u_{\theta,1}$, primeiro componente principal supervisionado, para acessar a contribuição individual de cada variável. Os autores definem o escore de importância como sendo o produto interno entre cada variável e $u_{\theta,1}$, dado por: $imp_j = \langle x_j, u_{\theta,1} \rangle$, ou $imp_j = cor(x_j, u_{\theta,1})$ no caso das variáveis x_j estarem padronizadas.

Assim, é possível ranquear as variáveis em ordem decrescente de contribuição para a predição da resposta, características com maior valor de $|\mathrm{imp_j}|$ são as mais importantes nos modelos preditivos.

2.2.4 Regressão Ridge Bayesiana

A regressão ridge bayesiana (BRR) assume que todos os coeficientes de regressão possuem uma variância comum. De acordo com Pérez e Campos (2014), para realização do ajuste de modelos lineares, quando se dispõe de tempos de sobrevivência e de uma variável indicadora de censura, considera-se que os fenótipos foram amostrados de uma distribuição normal truncada. O seguinte modelo linear foi considerado: $y_j = \mu + \sum_{i=1}^{k} x_{ij} m_i + e_j$, em que: y_j é o fenótipo do indivíduo j; μ é o intercepto; m_i é o efeito aleatório do marcador i; e_j é o erro da observação j, e x_{ij} é o genótipo do indivíduo j, para o marcador i, codificado como 0, 1 ou 2. As distribuições *a priori* assumidas para o efeito aleatório dos marcadores e para a variância genética associada aos marcadores, foram, respectivamente, $m_i \mid \sigma^2_m \sim N(0, \sigma^2_m)$ e $\sigma^2_m \mid df_m, S_m \sim \chi^{-2}(df, S)$, com: com $df_m > 0$ graus de liberdade e parâmetro de escala $S_m > 0$ (PÉREZ; CAMPOS, 2014). O ajuste do modelo BRR foi realizado via MCMC, com: 160000 iterações, burn-in de 20000 e thin igual a 10. O critério de Geweke do pacote BOA (SMITH, 2007) foi utilizado para a realização da análise de convergência no software R (R Development Core Team, 2019).

2.2.5 Modelo misto de Cox

O modelo misto de Cox, em estudos de seleção genômica ampla e de associação genômica, se fundamenta na equação:

$$\lambda(t) = \lambda_0(t)\exp\{\mathbf{X\beta} + \mathbf{Za}\},$$

em que: $\lambda_0(t)$ é a função base de taxa de falha; $\mathbf{X}$ e $\mathbf{Z}$ são as matrizes de incidência, relacionadas, respectivamente, aos efeitos fixos ($\boldsymbol{\beta}$) e genéticos aditivos ($\mathbf{a}$), sendo $\mathbf{a}$ normalmente distribuído com média zero e matriz de covariância $\mathbf{G}\sigma^2_a$. A estimação dos parâmetros é realizada com base na verossimilhança parcial do modelo de riscos proporcionais de Cox, com uso da aproximação de Laplace (RIPATTI; PALMGREN, 2000; THERNEAU; GRAMBSCH; PANKRATZ, 2003).

O modelo misto de Cox foi inicialmente ajustado considerando a matriz de parentesco genômica obtida com a utilização de todos os SNPs que passaram pelo controle de qualidade. Em seguida, os SNPs foram ranqueados de acordo com os coeficientes de importância obtidos pelos métodos RSF, GBM e SPCA, com base nestes ranks, foram utilizadas proporções dos SNPs mais relevantes, 1%, 2,5%, 5%, 10% e SNPs com coeficiente de importância positivo, para a obtenção da matriz de parentesco genômica em diferentes densidades de SNPs. Para cada uma destas densidades de SNPs, o modelo misto de Cox foi utilizado para a predição de valores genéticos genômicos, e para estimação da variância genética e da herdabilidade.

A herdabilidade (h^2) do tempo de sobrevivência foi calculada com base na variância aditiva (σ_a^2) estimada via modelo misto de Cox e na proporção de observações censuras (c), pela equação proposta por Yazdi et al. (2002), dada por:

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + 1/(1-c)},$$

com c igual a 4,7% para os dados utilizados neste estudo.

2.3 Critérios para a comparação dos modelos

O desempenho preditivo dos modelos, ao longo do tempo, foi avaliado pelas medidas *Brier-Score* (BS) e a AUC (*Area under the Receiver Operating Characteristic Curve*). De acordo com (GRAF et al., 1999), BS no tempo t, pode ser estimado pela expressão:

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N} w_i(t)[\hat{S}(t|x_i) - y_i(t)]^2,$$

em que: $y_i(t) = I(t_i \geq t)$ é o verdadeiro status do indivíduo i no tempo t, $\hat{S}(t|x_i)$ representa a probabilidade de sobrevivência do indivíduo i no tempo t, dado o conjunto de variáveis preditoras, N o número de indivíduos e $w_i(t)$ denota o peso associado ao i-ésimo indivíduo. Os pesos são estimados com base no estimador Kaplan-Meier para a função de sobrevivência (S), de modo que:

$$w_i(t) = \begin{cases} \delta_i / S(y_i) \text{ se } y_i \leq t \\ 1/S(y_i) \text{ se } y_i > t \end{cases}.$$

AUC foi estimada pela utilização do c-index, que segundo Harrell, Lee e Mark (1996) mede a probabilidade de concordância entre os casos preditos e os observados, entre todos os pares de respostas. A AUC em função do tempo é dada por:

$$\mathrm{AUC}(t) = \mathrm{P}\left(\hat{y}_i < \hat{y}_j \middle| y_i < t, y_j > t\right) = \frac{1}{\mathrm{num}(t)} \sum_{i:y_i<t} \sum_{j:y_j>t} \mathrm{I}\left(\hat{y}_i < \hat{y}_j\right),$$

em que: $\mathrm{num}(t)$ representa o número de pares comparáveis no tempo t.

A AUC é uma medida que apresenta valores no intervalo [0,1], sendo o valor um indicando um modelo com predições perfeitas e 0,5 um modelo com predições aleatórias. Valores entre zero e 0,5 indicam que o modelo prediz o evento de forma contrária a esperada. Já o Brier Score apresenta valor 0 para predições perfeitas, 0,25 para predições equivalentes a um palpite aleatório, e 1 para um baixo poder de predição.

A validação cruzada (7 – fold) foi utilizada para a avaliação dos modelos. Os conjuntos de teste e de validação foram definidos de modo que as proporções de observações censuradas nos dois conjuntos fossem aproximadamente iguais. Os 777 animais foram divididos aleatoriamente em sete conjuntos, compostos por 111 animais cada (14% da população). Os conjuntos de teste foram utilizados para o ajuste dos modelos. Em seguida, estes modelos foram utilizados para a realização de predições nos conjuntos de validação. No processo de validação cruzada, foram geradas sete estimativas para as medidas, AUC, BS, correlação de Spearman e porcentagens de indivíduos selecionados em comum pelos métodos RSF, GBM, SPCA e BRR dois a dois, sendo uma estimativa para cada conjunto de validação. Para comparação dos modelos foram considerados os valores médios destas medidas.

Todas as análises foram realizadas no software R (R Core Tem, 2019). Os pacotes randomsurvivalSRC (ISHWARAN; KOGALUR, 2018), GBM (RIDGEWAY, 2017) e superpc (BAIR; TIBSHIRANI, 2015) foram utilizados, respectivamente, para o ajuste dos modelos RSF, GBM e SPCA, e o pacote BGLR (PÉREZ; DE LOS CAMPOS, 2014) para ajuste do modelo Regressão Ridge Bayesiana (BRR). O modelo misto de Cox foi ajustado para predição genômica via pacote coxme (THERNEAU, 2020), e para associação genômica via pacote coxmeg (HE, 2019).

3 RESULTADOS E DISCUSSÃO

3.1 Predição genômica

Os métodos RSF, GBM e SPCA foram ajustados aos dados genômicos, sendo os parâmetros dos modelos RSF e GBM escolhidos de forma a estabilizar o erro out-of-bag (OOB) e os parâmetros do SPCA via inspeção gráfica pela plotagem da estatística do teste da razão de verossimilhanças em função dos thresholds e do número de componentes principais. A Tabela 1 mostra os parâmetros selecionados para o ajuste dos modelos.

Tabela 1: Parâmetros utilizados para o ajuste dos modelos RSF, GBM e SPCA.

Método	RSF			GBM		SPCA	
Parâmetro	ntree	Mtry	Nodesize	ntree	shrinkage	n.comp	th
Valor	1500	88,50	3	1500	0,001	1	0,4822

Fonte: Autor. ntree: número total de árvores; mtry: número de SNPs selecionados aleatoriamente como candidatos para a divisão em um dado nó; nodesize: número médio de casos únicos em um nó terminal; shrinkage: taxa de aprendizado; n.comp: número de componentes principais; th: threshold.

A Tabela 2 apresenta os valores médios estimados para AUC e BS (desvios-padrão), e os valores absolutos médios dos desvios da AUC em relação a 0,5, estimados via validação cruzada (7 – fold), para os métodos RSF, GBM, SPCA e BRR. Foram estimados valores de AUC maiores que 0,5 para os métodos RSF e GBM, e menores que 0,5 para os métodos SPC e BRR. Os valores estimados para os métodos RSF e GBM não diferiram estatisticamente, já os estimados para SPCA e BRR apresentaram uma diferença estatística. As medidas DAUC e BS variaram, respectivamente, nos intervalos de 0,0347 a 0,0859, e de 0,0419 a 0,26.

Tabela 2: Valores médios estimados para AUC (area under the operating characteristic curve), DAUC (valor absoluto dos desvios da AUC em relação a 0,5) e BS (Brier Score), e seus respectivos desvios-padrão, para os métodos RSF, GBM, SPCA e BRR, considerando o processo de validação cruzada 7-fold.

	AUC	DAUC	BS
RSF	$0,5777^{a}$ (0,0576)	$0,0806^{ab}$ (0,0528)	$0,0442^{a}$ (0,0034)
GBM	$0,5577^{a}$ (0,0568)	$0,0662^{b}$ (0,0447)	$0,2600^{b}$ (0,1231)
SPCA	$0,4664^{b}$ (0,0269)	$0,0347^{b}$ (0,0253)	$0,0906^{b}$ (0,0098)
BRR	$0,4141^{c}$ (0,0425)	$0,0859^{a}$ (0,0426)	$0,0419^{a}$ (0,0037)

Fonte: Autor. RSF: Random Survival Forest; GBM: Gradient Boosting Machine; SPCA: Análise de Componentes Principais Supervisionados; BRR: Regressão Ridge Bayesiana. Valores estimados para as medidas seguidos por uma mesma letra nas colunas, indicam não

haver uma diferença significativa entre as médias pelo teste t pareado a 5% de probabilidade.

A divisão dos modelos segundo a AUC é devida ao tipo de valores preditos pelos modelos na população de teste. Para os modelos RSF e GBM, foram preditos valores para a função de distribuição ($F(t) = 1 - S(t)$). No caso do GBM, os valores foram preditos em escala do logaritmo da função taxa de falha. No modelo SPCA, foram preditos scores, de modo que, altos valores indicam uma pior sobrevivência, ou seja, alto risco de ocorrência do evento de interesse. Para estimação da AUC no modelo SPCA, foi utilizada uma medida complementar ao score obtido via SPCA. No modelo BRR, foram modelados os valores genéticos genômicos preditos (GEBVs), que apresentam uma relação direta com a função de sobrevivência e com os tempos de falha, e uma relação inversa com a função taxa de falha.

As medidas AUC e BS foram escolhidas para avaliação dos modelos, dada a impossibilidade de se estimar diretamente os efeitos dos SNPs pelos métodos RSF, GBM e SPCA, o que inviabiliza a predição dos valores genéticos genômicos, e a posterior utilização da correlação de Pearson para o cálculo da acurácia dos modelos. A AUC foi utilizada como medida para a comparação do desempenho de modelos preditivos em diversos estudos genômicos, dentre os quais podemos citar (GANZÁLEZ-RECIO; FORNI, 2011; NADERI; YIN; KÖNIG, 2016).

O modelo BRR apresentou o maior valor de DAUC, seguido pelos métodos RSF, GBM e SPCA. Os métodos RSF e BRR mostraram um desempenho semelhante em relação à medida DAUC, não sendo encontradas diferenças significativas entre os valores estimados. Os valores estimados de DAUC para os métodos RSF, GBM e SPCA não diferiram estatisticamente, sendo o valor estimado para o método BRR maior do que os estimados para os métodos GBM e SPCA. Quanto ao BS, não foram observadas diferenças significativas entre os métodos RSF e BRR, sendo os maiores valores médios estimados em ordem crescente para os métodos SPCA e GBM, iguais estatisticamente, e diferentes dos estimados para os métodos RSF e GBM. Considerando que o método BRR é um modelo tradicionalmente utilizado em estudos de predição genômica, ele foi tomado como referência para a comparação com os demais modelos.

Não foram encontrados na literatura estudos que avaliassem os métodos RSF, GBM e SPCA, pelas medidas AUC e BS, comparando-os ao modelo BRR para características censuradas. Entretanto, alguns estudos foram realizados para características fenotípicas contínuas ou binárias, considerando dados reais e simulados. Ogutu, Piepho e Schulz-Streeck (2011) avaliaram a acurácia (correlação de Pearson) preditiva dos modelos *Random Forest* (RF), *Gradient Boosting* (Boosting) e Máquina de Vetores Suporte (SVM), para a predição de

valores genéticos genômicos utilizando dados simulados. Os resultados obtidos mostraram que os métodos *Boosting* e SVM, apresentaram valores de acurácia (0,503) próximos dos obtidos pela regressão ridge BLUP (0,53), e uma acurácia ainda menor para a RF (0,466). Grinberg, Orhobor e King (2019) analisaram o uso de métodos de aprendizado de máquina (GBM, SVM, Ridge Regression, RF, LASSO, Rede Elástica) e G-BLUP para a predição de fenótipos em estudos genômicos com leveduras, arroz e trigo, investigando o desempenho dos modelos pela proporção da variância explicada por eles (R^2). De modo geral, para os fenótipos relacionados aos dados de levedura, o método GBM apresentou um melhor desempenho, seguido pelos métodos LASSO e G-BLUP. Já para os fenótipos de arroz e de trigo, os métodos SVM e G-BLUP foram os melhores. Os estudos apresentados mostram não haver um consenso quanto à superioridade dos métodos de aprendizado de máquina em relação aos métodos usualmente utilizados na predição genômica.

Na Tabela 3 são apresentados os valores médios estimados para o coeficiente de correlação de Spearman, e as porcentagens de indivíduos selecionados em comum pelos métodos RSF, GBM, SPCA e BRR dois a dois. Foram observados valores positivos e negativos para a correlação de Spearman, tais valores se devem ao fato de haver uma relação direta entre os valores preditos pelos métodos RSF e GBM; e SPCA e BRR, e uma relação inversa entre os valores preditos pelos modelos RSF e SPCA; RSF e BRR; GBM e SPCA; GBM e BRR.

O maior valor estimado para a correlação de Spearman foi obtido entre os métodos RSF e GBM (0,5866), seguido pelas correlações entre GBM e BRR (-0,4091), e RSF e BRR (-0,3965), já o menor valor (-0,0662), foi obtido pela correlação entre GBM e SPCA, considerando 40% dos animais da população de teste. Considerando todos os animais da população de teste para o cálculo da correlação de Spearman, notou-se um aumento da correlação, sendo obtidos valores fortes de correlação (0,60 a 0,79) para os pares RSF x GBM, RSF x BRR e GBM x BRR. Para os demais pares foram constatados valores moderados (0,40 à 0,59) ou fracos (0,20 à 0,39) de correlação.

A frequência relativa de concordância entre os top 20% e 40% dos animais variou, respectivamente, entre 0,3456 (SPCA e BRR) e 0,7180 (RSF e GBM); e entre 0,5468 (SPCA e BRR) e 0,7732 (RSF e BRR). Com a mudança de top 20% para 40%, houve um aumento na concordância entre os métodos, exceto para o par GBM e RSF, que sofreu uma pequena redução. Os maiores valores de concordância, nos dois ranks, foram observados entre os métodos, RSF e GBM, RSF e BRR, GBM e BRR. Dado o tamanho da população em estudo, não se pode considerar proporções menores de indivíduos selecionados, o que poderia conduzir

a estimativas incorretas para a correlação de Spearman. Os resultados da correlação de Spearman e concordância mostram que o método utilizado na avaliação possui grande influência na seleção dos indivíduos, podendo, portanto, resultar em diferentes ganhos genéticos para a característica estudada.

Tabela 3: Correlação de Spearman entre 40% dos animais selecionados (acima da diagonal) e entre todos os animais (abaixo da diagonal), segundo a menor probabilidade de ocorrência do evento (RSF e GBM), maiores valores para complementar do Cox escore (SPCA) e maiores GEBVs (BRR); e a porcentagem de animais em comum entre os métodos considerando 20% (acima da diagonal) e 40% (abaixo da diagonal) dos animais selecionados.

Modelo	RSF	GBM	SPCA	BRR
		Correlação de Spearman		
RSF	-	0,5866 (0,1268)	-0,2519 (0,1678)	-0,3965 (0,1549)
GBM	0,6701 (0,0939)	-	-0,0662 (0,1610)	-0,4091 (0,2133)
SPCA	-0,5388 (0,1527)	-0,4335 (0,1213)	-	0,1215 (0,2389)
BRR	-0,7538 (0,0456)	-0,6576 (0,0512)	0,3788 (0,1499)	-
		Frequência relativa de indivíduos em comum		
RSF	-	0,7180 (0,1039)	0,4729 (0,0911)	0,6290 (0,0794)
GBM	0,6937 (0,0833)	-	0,4726 (0,1086)	0,6160 (0,0713)
SPC	0,6362 (0,0985)	0,5913 (0,0767)	-	0,3456 (0,0755)
BRR	0,7732 (0,0319)	0,7031 (0,0494)	0,5468 (0,0969)	-

Fonte: Autor. RSF: *Random Survival Forest*; GBM: *Gradient Boosting Machine*; SPCA: Análise de Componentes Principais Supervisionados; BRR: Regressão Ridge Bayesiana.

Lázaro et al. (2019) compararam modelos bayesianos, aplicados a dados censurados para estimação de parâmetros genéticos para a característica idade ao primeiro parto de fêmeas da raça Brahman. Os valores estimados para a correlação de Spearman entre os GEBVs preditos pelos métodos variaram entre 0,82 e 0,97, o que indica uma correlação muito forte. Já as proporções de indivíduos em comum nos ranques dos top1% e top10%, variaram respectivamente, nos intervalos de 32,79 e 82,96%, e de 59,48 e 89,12%.

3.2 Identificação de SNPs importantes e GWAS

Na Figura 1 são plotados os coeficientes de importância dos SNPs para os métodos RSF,

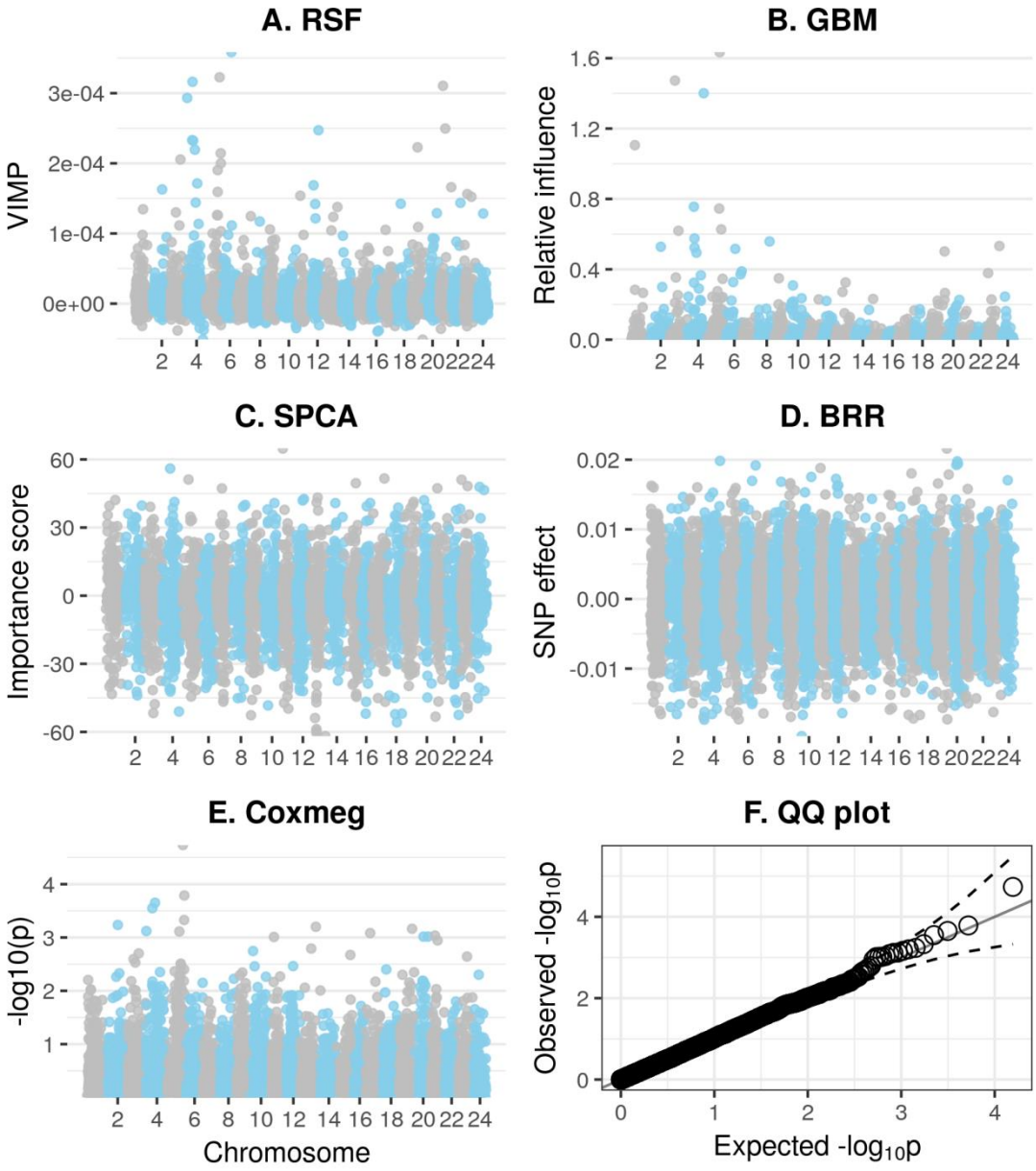
GBM e SPCA e os efeitos de substituição alélica dos SNPs estimados via BRR e os gráficos Manhattan Plot e QQ plot obtidos via modelo misto de Cox. No modelo de Cox, observou-se que nenhum dos SNPs superou o limiar de significância de 5,195, considerando a correção de Bonferroni ao nível de significância de 5% (Figuras 1E e 1F). Palaiokostas et al. (2016) também não identificaram SNPs significativamente associados a resistência à doença Pasteurellosis. Os autores fizeram uso do pacote rrBLUP (ENDELMAN, 2011), que considera apenas o tempo de sobrevivência, ignorando a variável indicadora de censura.

Analisando-se os 40 SNPs mais relevantes, ranqueados de acordo com os maiores valores de importância (RSF, GBM e SPCA), efeito (BRR) e p-valor (Cox), percebeu-se que na maioria das vezes, eles se localizaram em um mesmo grupo de ligação (GL). Contrastando os métodos RSF, GBM e SPCA, com os métodos BRR e coxmeg, foram evidenciados, respectivamente, 14 (GL1-6, GL8, GL11-13, GL19-21 e GL24), 12 (GL1-6, GL8-10, GL13, GL19 e GL24) e 16 (GL1-5, GL8, GL10-13, GL15, GL17, GL19-21 e GL24); e 13 (GL2-5, GL11-13, GL18-21, GL23 e GL24), 10 (GL2-5, GL9, GL10, GL13, GL19, GL23 e GL24) e 16 (GL2-5, GL10-13, GL15, GL17-21, GL23 e GL24) regiões simultaneamente.

Quando confrontadas as regiões evidenciadas pelos métodos de aprendizado de máquina com as evidenciadas pelo modelo misto de Cox, os grupos de ligação: GL2-5, GL13, GL19, GL23 e GL24, foram destacados pelos três métodos (Figuras 1A, 1B e 1E). Ao se confrontar os métodos SPCA e BRR, com o modelo misto de Cox, foram realçados concomitantemente os grupos, GL2-5, GL10-13, GL15, GL17, GL19-21 e GL24 (Figuras 1C, 1D e 1E). Por fim, os cinco métodos evidenciaram simultaneamente os: GL2-5, GL13, GL19 e GL24.

Embora os métodos evidenciem regiões semelhantes, eles apresentaram um perfil dos top 40 SNPs bem diferente. Comparando o rank obtido pelos métodos RSF, GBM e SPC, respectivamente, com os obtidos pelos modelos BRR e misto de Cox, notou-se, 5, 3 e 0; e 17, 8 e 1 SNPs na interseção dos ranks. O SNP mais próximo de ser significativo pelo modelo misto de Cox, foi o M2789 localizado no grupo de ligação 5. Este SNP foi o segundo no rank da RSF, e não apareceu no rank dos demais métodos. Os métodos de aprendizado de máquina, destacaram com clareza grupos de ligação localizados no início, meio e fim do genoma. Estas mesmas regiões foram evidenciadas pelo modelo misto de Cox.

Figura 1: A. Manhattan plot dos valores absolutos de importância dos SNPs (VIMP) estimados via *Random Survival Forest* (RSF). **B.** Manhattan plot dos valores absolutos de influência relativa dos SNPs estimados via *Gradient Boosting Machine* (GBM). **C.** Manhattan plot dos valores absolutos de escore de importância dos SNPs estimados via Análise de Componentes Principais Supervisionados (SPCA). **D.** Manhattan plot dos valores absolutos dos efeitos dos SNPs estimados via Regressão Ridge Bayesiana (BRR). **E.** Manhattan plot mostrando os valores de -log 10(p) para cada SNP, obtidos via coxmeg. **F.** QQ plot mostrando a relação entre os valores observados e esperados para -log 10(p) com intervalo de confiança de 95%.

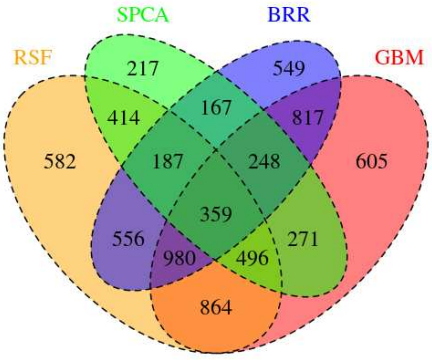


Fonte: Autor.

Na Figura 2 foi exibido o diagrama de Venn, mostrando as interseções dos conjuntos de SNPs que apresentaram VIMP (RSF), influência relativa (GBM), escore de importância (SPCA) e efeitos (BRR) maiores que zero. Foram observados 4438 (RSF), 4640 (GBM), 2359 (SPCA) e 3863 (BRR) SNPs com efeitos positivos. A interseção entre os quatro métodos

contém um total de 359 SNPs. O número de SNPs nas interseções duplas foi de 1374 (GBM x SPCA), 961 (SPCA x BRR), 2404 (GBM x BRR), 2699 (RSF x GBM), 2082 (RSF x BRR) e 1456 (RSF x SPCA). Quanto às interseções triplas, foram observados: 607 (GBM x SPCA x BRR), 546 (RSF x SPCA x BRR), 855 (RSF x GBM x SPCA) e 1339 (RSF x GBM x BRR).

Figura 2: Diagrama de Venn considerando os conjuntos de SNPs com VIMP, influência relativa, escore de importância e efeitos maiores que zero, segundo os métodos RSF, GBM, SPCA e BRR, e todas as possíveis interseções entre os conjuntos.



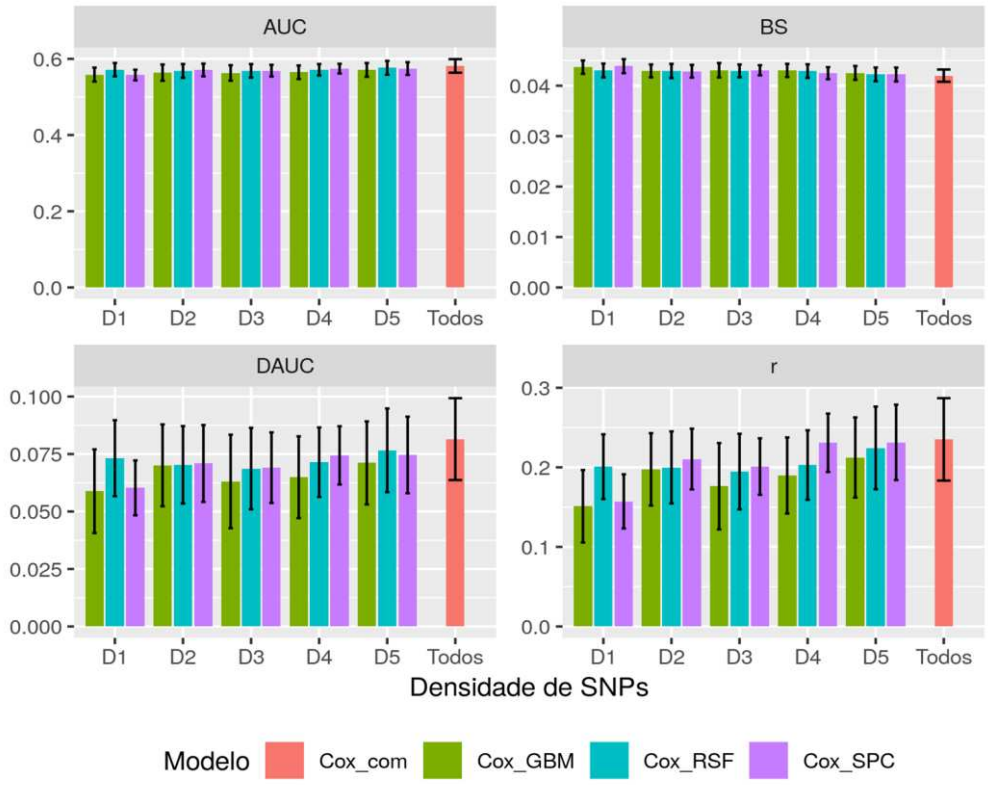
Fonte: Autor.

3.3 Utilização de subconjuntos de SNPs no modelo misto de Cox

Com intuito de verificar o impacto da seleção de variáveis utilizando os métodos RSF, GBM e SPCA na predição de valores genéticos, subconjuntos contendo 1%, 2,5%, 5% e 10% dos SNPs mais relevantes identificados por cada método forma utilizados na predição genômica via modelo misto de Cox. Na Figura 3 são apresentados os valores médios estimados para AUC, BS, DAUC e habilidade preditiva (r) do modelo misto de Cox, via validação cruzada 7 - fold, com seus respectivos desvios-padrão, considerando os cinco subconjuntos de SNPs selecionados pelos métodos RSF, GBM e SPCA. Não foram encontradas diferenças significativas entre os métodos para as diferentes densidades de marcadores em termos de habilidade preditiva, pelo teste t pareado com 5% de probabilidade, ajustado pela correção Bonferroni. Com o aumento da densidade dos SNPs, foi observada uma tendência crescente nos valores estimados pelas medidas r e DAUC, levemente crescente para a AUC, e uma tendência levemente decrescente nos valores estimados pela medida BS. As duas primeiras apresentaram um padrão semelhante, nitidamente diferente do apresentado pelas medidas AUC e BS. Li et al. (2019) comentam que a utilização de muitos SNPs em modelos genômicos nem sempre conduzem a uma melhor acurácia. Segundo os autores, um pequeno número de SNPs

pode ser capaz de capturar efeitos principais, interações e relações não lineares e pode gerar erros de predições menores que os gerados por um modelo com todos os SNPs.

Figura 3: Gráfico de barras para os valores médios estimados para AUC, BS, DAU e $r_{(y,\widehat{GEBV})}$, para o modelo misto de Cox com todos os SNPs (Cox_com), e considerando os SNPs selecionados pelos métodos RSF (Cox_RSF), GBM (Cox_GBM) e SPCA (Cox_SPCA). D1 - 1% dos SNPs, D2 – 2,5% dos SNPs, D3 - 5% dos SNPs, D4 - 10% dos SNPs e D5 - SNPs com VIMP positivo.



Fonte: Autor.

A Tabela 4 apresenta os valores médios estimados para a variância genética aditiva e para a herdabilidade da característica resistência à doença Pasteurellosis, considerando os subconjuntos com 1%, 2,5%, 5% e 10% dos SNPs, e com os SNPs que apresentaram importância positiva, selecionados pelos métodos RSF, GBM e SPCA. É apresentada também, a razão entre a variância genética dos modelos com subconjuntos e com todos os SNPs. Pelo modelo misto de Cox, utilizando todos os SNPs foi estimada uma herdabilidade de 0,2474± 0,0514 e uma variância genética aditiva de 0,3507±0,0937. Sendo que, essa estimativa de herdabilidade não difere da estimada por (PALAIOKOSTAS et al., 2016), via modelos bayesianos, no valor de 0,28 com intervalo de 95% de maior densidade *a posteriori* (0,17-0,4).

Tabela 4: Estimativas dos componentes de variância e proporção da variância genética explicada para o tempo de sobrevivência estimados via modelo misto de Cox para os subconjuntos selecionados pelos métodos RSF (*Random Survival Forest*), GBM (*Gradient Boosting Machine*) e SPCA (Análise de Componentes Principais Supervisionados).

	RSF			GBM			SPC		
	h^2	Vg	P	h^2	Vg	P	h^2	Vg	P
D1	0,19 (0,03)	0,25 (0,04)	71%	0,15 (0,02)	0,19 (0,04)	54%	0,08 (0,04)	0,09 (0,06)	26%
D2	0,27 (0,03)	0,38 (0,05)	108%	0,23 (0,03)	0,32 (0,06)	91%	0,18 (0,07)	0,24 (0,10)	69%
D3	0,33 (0,04)	0,52 (0,09)	148%	0,31 (0,03)	0,47 (0,06)	134%	0,24 (0,09)	0,35 (0,15)	100%
D4	0,38 (0,03)	0,65 (0,09)	185%	0,36 (0,04)	0,59 (0,11)	168%	0,32 (0,06)	0,51 (0,12)	146%
D5	0,32 (0,04)	0,50 (0,10)	142%	0,38 (0,05)	0,66 (0,13)	188%	0,39 (0,02)	0,66 (0,07)	189%

Fonte: Autor. h^2: Herdabilidade; Vg: Variância genética aditiva; p: Razão da variância genética estimada pelos subconjuntos, pela estimada pelo conjunto total de SNPs; D1: 1% dos SNPs; D2: 2.5% dos SNPs, D3: 5% dos SNPs; D4: 10% dos SNPs; D5: SNPs com VIMP positivo.

Quando o método RSF foi utilizado para a seleção dos SNPs empregados para o cômputo da matriz **G**, notou-se que a herdabilidade e a variância genética apresentaram um comportamento crescente entre 1 e 10%. Aparentemente, estes componentes começam a decrescer com utilização dos SNPs com importância positiva. Quando utilizados os SNPs selecionados pelos métodos GBM e SPCA, percebeu-se um comportamento crescente até o conjunto composto por todos os SNPs com importância positiva. Os valores estimados para herdabilidade e para variância genética, na maioria das vezes, foram superiores aos estimados com a utilização de todos os SNPs.

As herdabilidades estimadas com a utilização dos subconjuntos, foram próximas da estimada com todos os SNPs, apenas para as densidades D2 (RSF – 0,27±0,03 e GBM – 0,23±0,03) e D3 (SPCA – 0,24±0,09). A variância genética explicada por 1%, 2,5% e 5% do conjunto total de SNPs, variou entre 26% (SPCA) e 148% (RSF) do total da variância explicada por todos os SNPs. De modo geral, o modelo RSF mostrou um desempenho superior ao dos demais métodos, com 1 e 2,5% dos SNPs com importância positiva, as herdabilidades estimadas foram de, respectivamente, 0,19 e 0,27, estes valores, dentre todos os valores de herdabilidade,

foram os mais próximos do obtido por Palaiokostas et al. (2016) no valor de 0,28. As proporções de variância explicada foram de 71 e 108%, respectivamente, para 1 e 2,5% dos SNPs. Li et al. (2019) ao considerar os tops 400 e 1000 SNPs, de um total de 38,082 SNPs, mostrou que estas quantidades eram suficientes para a obtenção de proporções altas de variância genética, com 400 dos SNPs – 48,47% (RF) e 53,29% (GBM), e com 1000 dos SNPs – 61,29% (RF) e 82,45% (GBM).

4 CONCLUSÃO

Neste estudo avaliamos o desempenho de dois modelos de aprendizado de máquina (RSF e GBM), e de um modelo de redução de dimensionalidade (SPCA), quando aplicados em estudos de seleção genômica ampla para a característica resistência à doença pasteurellosis em douradas - gilthead sea bream (*Sparus aurata*). Comparando estes modelos com o modelo BRR, nota-se que, a RSF apresenta um desempenho preditivo (AUC e BS) muito semelhante ao do BRR, sendo o BRR claramente superior aos demais métodos. Os métodos de aprendizado de máquina (RSF e GBM) destacaram vários SNPs, espalhados pelo genoma, como preditores importantes do tempo de sobrevivência. Embora nenhuma associação significativa tenha sido encontrada pelo modelo misto de Cox. O padrão de disposição da importância dos SNPs para os métodos RSF, GBM e misto de Cox são muito semelhantes. A utilização de subconjuntos de SNPs com pelo menos 2,5% do total de SNPs proporciona estimativas de herdabilidade e de habilidade preditiva próximas das obtidas com todos os SNPs. Por fim, conclui-se que dentre os métodos de sobrevivência considerados, a RSF é a que apresenta o maior potencial para aplicação em estudos de seleção genômica ampla com dados censurados.

REFERÊNCIAS

ANTONELLO, J.; MASSAULT, C.; FRANCH, R.; HALEY, C.; PELLIZZARI, C.; BOVO, G.; PATARNELLO, T.; KONING, D. J. DE; BARGELLONI, L. Estimates of heritability and genetic correlation for body length and resistance to fish pasteurellosis in the gilthead sea bream (Sparus aurata L.). **Aquaculture,** v. 298, p. 29-35, 2009.

BAIR, E.; TIBSHIRANI, R. Semi-supervised methods to predict patient survival from gene expression data. **PLoS Biology,** v. 2, n. 4, p. 511-522, 2004.

BAIR, E.; HASTIE, T.; PAUL, D.; TIBSHIRANI, R. Prediction by supervised principal components. **American Statistical Association,** v. 101, n. 473, p. 119-137, 2006.

BAIR, E.; TIBSHIRANI, R. superpc: Supervised principal components, R package version 1.09., 2015. Disponível em: https://cran.r-project.org/web/packages/superpc/superpc.pdf. Acesso em: 2 fev. 2020.

BARRÍA, A.; CHRISTENSEN, K. A.; YOSHIDA, G. M.; CORREA, K.; JEDLICKI, A.; LHORENTE, J. P.; DAVIDSON, W. S.; YÁÑEZ, J. M. Genomic Predictions and Genome-Wide Association Study of Resistance Against Piscirickettsia salmonis in Coho Salmon (Oncorhynchus kisutch) Using ddRAD Sequencing. **G3,** v. 8, p. 1183-1194, 2018.

BREIMAN, L. Random forest. **Machine Learning,** v. 45, n. 1, p. 5-32, 2001.

CHEN, Y.; JIA, Z.; MERCOLA, D.; XIE, X. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. **Computation and Mathematical Methods in Medicine,** p. 1-8, 2013.

DU, C.; WEI, J.; WANG, S.; JIA, Z. Genomic selection using principal component regression. **Heredity,** v. 121, p. 12-23, 2018.

ENDELMAN, J. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. **The Plant Genome,** v. 4, n. 3, p. 250-255, 2011.

FRIEDMAN, J. H.; TIBSHIRANI, R. Additive logistic regression: a statistical view of boosting. **The Annals of Statistics**, v. 28, p. 337-407, 2000.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **Annals of Statistics,** v. 29, n. 5, p. 1189-1232, 2001.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational Statistics and Data Analysis**, v. 38, n. 4, p. 367-378, 2002.

GRAF, E.; SCHMOOR, C.; SAUERBREI, W.; SCHUMACHER, M. Assessment and comparison of prognostic classification schemes for survival data. **Statistics in medicine**, v. 18, n. 17-18, p.2529-2545, 1999.

GRINBERG, N. F.; ORHOBOR, O. I.; KING, R. D. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. **Machine Learning,** p. 1-27, 2019.

GONZÁLEZ-RECIO, O.; FORNI, S. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. **Genetics Selection Evolution,** v. 43, n. 7, p. 1-12, 2011.

HARRELL JR., F. E.; LEE, K. L.; MARK, D. B. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. **Statistics in Medicine**, v. 15, p. 361-387, 1996.

HE, L. coxmeg: Cox Mixed-Effects Models for Genome-Wide Association Studies. R package version 1.0.11, 2019. Disponível em: https://cran.r-project.org/web/packages/coxmeg/index.html. Acesso em: 22 jul. 2020.

HOUSTON, R. D. Future directions in breeding for disease resistance in aquaculture species. **Revista Brasileira de Zootecnia,** v. 46, n. 6, p. 545-551, 2017

ISHWARAN, H. Variable importance in binary regression trees and forests. **Electron. J. Statist.,** v. 1, p. 519-537, 2007.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. Random survival forests. **The annals of applied statistics**, v. 2, n. 3, 841-860, 2008.

ISHWARAN, H.; KOGALUR, U. B. Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 2.6.1., 2018. Disponível em: https://cran.r-project.org/web/packages/randomForestSRC/index.html. Acesso em: 2 jul. 2018.

LAFFERTY, K. D.; HARVELL, C. D.; CONRAD, J. M.; FRIEDMAN, C. S.; KENT, M. L.; KURIS, A. M.; POWELL, E. N.; RONDEAU, D.; SAKSIDA, S. M. Infectious diseases affect marine fisheries and aquaculture economics. **Annual Review of Marine Science,** v. 7, p. 471-496, 2015.

LÁZARO, S. F.; VARONA, L.; SILVA, F. F.; VENTURA, H. T.; VERONEZEA, R.; BRITO, L. C.; COSTA, E. V.; LOPES, P. S. Genetic evaluation for days to calving in Nellore heifers using Exponential and Gaussian Censored Bayesian models. **Livestock Science,** v. 230, p. 1-4, 2019.

LI, B.; ZHANG, N.; WANG, Y. G.; GEORGE, A. W.; REVERTER, A.; LI, Y. Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods. **Frontiers in Genetics,** v. 9, p. 1-20, 2018.

MÉSZÁROS, G.; SOLKNER, J.; DUCROCQ, V. The survival kit: software to analyze survival data including possibly correlated random effects. **Computer Methods and Programs in Biomedicine,** v. 110, p. 503–510, 2013.

MISZTAL, I.; TSURUTA, S.; LOURENCO, D.; MASUDA, Y.; AGUILAR, I.; LEGARRA, A.; VITEZICA, Z. Manual for BLUPF90 Family of Programs. University of Georgia, Athens, GA., 2018. Disponível em: http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf. Acesso em: 29 abr. 2020.

NADERI, S.; YIN, T.; KÖNIG, S. Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. **Journal of Dairy Science,** v. 99, n. 9, p. 7261-7273, 2016.

ODEGARD, J.; BARANSKI, M.; GJERDE, B.; GJEDREM, T. Methodology for genetic evaluation of disease resistance in aquaculture species: challenges and future prospects. **Aquaculture Research,** v. 42, p. 103-114, 2011.

OGUTU, J. O.; PIEPHO, H. P.; SCHULZ-STREECK, T. A comparison of random forests, boosting and support vector machines for genomic selection. **BMC Proceedings,** v. 5, p. 1-5, 2011.

PALAIOKOSTAS, C.; FERRARESSO, S.; FRANCH, R.; HOUSTON, R. D.; BARGELLONI, L. Genomic Prediction of Resistance to Pasteurellosis in Gilthead Sea Bream (*Sparus aurata*) Using 2b-RAD Sequencing. **G3,** v.6, p.3693-3700, 2016. Disponível em: https://www.g3journal.org/content/6/11/3693.supplemental. Acesso em: 9 dez. 2019.

PALAIOKOSTAS, C.; CARIOU, S.; BESTIN, A.; BRUANT, J. S.; HAFFRAY, P.; MORIN, T.; CABON, J.; ALLAL, F.; VANDEPUTTE, M.; HOUSTON, R. D. Genome-wide association and genomic prediction of resistance to viral nervous necrosis in European sea bass (*Dicentrarchus labrax*) using RAD sequencing. **Genetics Selection Evolution,** v. 50, n. 30, p. 1-11, 2018.

PÉREZ, P.; CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483–495, 2014.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Disponível em: https://www.R-project.org/. Acesso em: 2 jul. 2018.

RIDGEWAY, G. The state of boosting. **Computing Science and Statistics**, v. 31, p. 172–181, 1999.

RIDGEWAY, G. gbm: Generalized Boosted Regression Models. **R package**, 2017. Disponível em: https://cran.r-project.org/web/packages/gbm/gbm.pdf. Acesso em: 3 out. 2018.

RIDGEWAY, G. Generalized Boosted Models: A guide to the gbm package, 2019. Disponível em: https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf. Acesso em: 31 jan. 2020.

ROBLEDO, D.; MATIKA, O.; HAMILTON, A.; HOUSTON, R. D. Genome-Wide Association and Genomic Selection for Resistance to Amoebic Gill Disease in Atlantic Salmon. **G3,** v. 8, p. 1195-1203, 2018.

RIPATTI, S.; PALMGREN, J. Estimation of multivariate frailty models using penalized partial likelihood. **Biometrics,** v. 56, p. 1016-1022, 2000.

SAURA, M.; CARABAÑO, M. J.; FERNÁNDEZ, A.; CABALEIRO, S.; DOESCHL-WILSON, A. B.; ANACLETO, O.; MAROSO, F.; MILLÁN, A.; HERMIDA, M.; FERNÁNDEZ, C.; MARTÍNEZ, P.; VILLANUEVA, B. Disentangling Genetic Variation for Resistance and Endurance to Scuticociliatosis in Turbot Using Pedigree and Genomic Information. **Frontiers in Genetics,** v. 10, p. 1-13, 2019.

SMITH, B. J. boa: An R Package for MCMC Output Convergence Assessment and Posterior Inference. **Journal of Statistical Software,** v. 21, p. 1-37, 2007.

THERNEAU, T. M.; GRAMBSCH, P. M.; PANKRATZ, V. S. Penalized survival models and frailty. **Journal of Computational and Graphical Statistics,** v. 12, p. 156-175, 2003.

THERNEAU, T. M. coxme: Mixed Effects Cox Models. R package version 2.2-16, 2020. Disponível em: https://cran.r-project.org/web/packages/coxme/index.html. Acesso em: 22 jul. 2020.

YAZDI, M. H.; VISSCHER, P. M.; DUCROCQ, V.; THOMPSON, R. Heritability, reliability of genetic evaluations and response to selection in proportional hazard models. **Journal of Dairy Science,** v. 85, p. 1563-1577, 2002.

CONCLUSÕES GERAIS

No estudo de simulação com dados censurados, as medidas de correlação propostas, a correlação maximal e a correlação de Pearson para dados censurados, mostraram-se mais adequadas para a avaliação da acurácia da predição genômica para fenótipos com observações censurados, do que a correlação de Pearson, principalmente para a característica C3.

O coeficiente angular estimado pela regressão linear simples, e o coeficiente associado ao efeito marginal obtido via regressão Tobit para dados censurados e não censurados, apresentam valores semelhantes na maioria dos cenários considerados. Em situações onde os pressupostos da normalidade e da homogeneidade dos resíduos forem aceitos, a regressão Tobit pode apresentar melhores estimativas.

No estudo com dados reais do tempo de sobrevivência de juvenis de douradas (*Sparus aurata*), submetidos à doença Pasteurellosis, o método Random Survival Forest desponta como uma alternativa interessante para aplicação em dados genômicos, apresentando desempenho preditivo semelhante ao obtido para o método Regressão Ridge Bayesiana.

A seleção de subconjuntos de marcadores com base nos métodos de aprendizado de máquina, *Random Survival Forest* e *Gradient Boosting Machine*, e de redução de dimensionalidade, Análise de Componentes Principais Supervisionados, para os dados utilizados, não implica em diferenças significativas na predição genômica via modelo misto de Cox.

A estimação dos componentes de variância e da herdabilidade com base nos subconjuntos de SNPs, implica em estimativas superestimadas dos mesmos, sendo 2,5% do total de SNPs suficientes para a explicação de uma porcentagem satisfatória da variância genética.