Nativa, Sinop, v. 10, n. 4, p. 466-471, 2022.
Pesquisas Agrárias e Ambientais
DOI: https://doi.org/10.31413/nativa.v10i4.13448
ISSN: 2318-7670

Desempenho dos métodos de estimação genômicos na identificação da resistência do arroz à brusone

Zeferino Gomes da SILVA NETO[1*], Sebastião MARTINS FILHO[1], Lucas Souza da SILVEIRA[1], Antônio Policarpo Souza CARNEIRO[1], Vinicius Silva dos SANTOS[2]

[1]Programa de Pós-Graduação em Estatística Aplicada e Biometria, Universidade Federal de Viçosa, Viçosa, MG, Brasil.
[2]Universidade Federal do Acre, Rio Branco, AC, Brasil.
*E-mail: zeferino.neto@ufv.br
(ORCID: 0000-0002-8982-7375; 0000-0002-8317-4318; 0000-0003-4356-751X; 0000-0002-9043-3242; 0000-0002-8387-2917)

Submetido em 17/02/2022; Aceito em 17/10/2022; Publicado em 01/11/2022.

RESUMO: Nos últimos anos, a perda de safras de arroz vem aumentando devido a estresses bióticos e abióticos, dentre os quais se destaca a brusone, que pode resultar em perdas de 100% em cultivares de arroz susceptíveis. Portanto, torna-se estratégico identificar metodologias que selecionem cultivares mais resistentes à doença. Neste trabalho, objetivou-se utilizar a análise de curva ROC (*Receiver operator characteristic*) e medidas tradicionais para a avaliação do desempenho de modelos de estimação genômicos (RR-BLUP, BLASSO e Bayes Cπ) na identificação da resistência do arroz à brusone. Os modelos RR-BLUP e Bayes Cπ foram mais acurados para a predição de resistência à brusone, enquanto o menor tempo de execução foi obtido pelo RR-BLUP. A área abaixo da curva ROC foi equivalente às medidas tradicionais para avaliar a acurácia dos modelos, com a vantagem de permitir a avaliação gráfica. Pela análise gráfica, o BLASSO obteve menor desempenho em altos níveis de especificidade (>0,75). Em menores níveis de especificidade, a sensibilidade dos modelos foi similar. A metodologia ROC mostrou-se uma boa alternativa para avaliação de modelos de predição genômica, podendo ser utilizada para a seleção de cultivares de arroz resistentes à brusone.
Palavras-chave: brusone do arroz; modelagem estatística; seleção genômica ampla; análise ROC; acurácia.

Performance of genomic estimation methods in the identification of rice resistance to brusone

ABSTRACT: In recent years, rice crop losses have increased due to biotic and abiotic stresses, among which brusone, which can result in 100% losses in susceptible rice cultivars. Therefore, it becomes strategic to identify methodologies that select resistant cultivars. In this work, we aimed to use ROC (Receiver operator characteristic) curve analysis and traditional measures to evaluate the performance of genomic estimation models (RR-BLUP, BLASSO and Bayes Cπ) in identifying rice resistance to brusone. The RR-BLUP and Bayes Cπ models were most accurate for the prediction of brusone resistance, while the best runtime was obtained by RR-BLUP. The area under the ROC curve was equivalent to traditional measures to evaluate the accuracy of the models, with the advantage of allowing graphical evaluation. By graphical analysis, BLASSO performed worst at high levels of specificity (>0.75). At lower levels of specificity, the sensitivity of the models was similar. The ROC methodology proved to be a good alternative for the evaluation of genomic prediction models, and can be used for the selection of rice cultivars resistant to brusone.
Keywords: rice blast; statistical modelling; genomics wide selection; ROC analysis; accuracy.

1. INTRODUÇÃO

O arroz (*Oryza sativa*) é uma cultura alimentar consumida por mais da metade da população mundial, fornecendo 20% das calorias diárias dos seres humanos (SOUZA et al., 2020). Em sua composição, encontram-se, principalmente: carboidratos, cálcio, ferro, tiamina e vitamina E (RATHNA et al., 2019), além de compostos bioativos com propriedades antioxidante, anticancerígena, antidiabética e anti-inflamatória (VERMA; SRIVASTAV, 2020).

No cenário mundial, os países do continente asiático se destacam na produção de arroz, principalmente a Índia, que possui a maior área de cultivo de arroz, e a China, que tem a maior produção do mundo (GADAL et al., 2019). Mundialmente, o aumento da demanda por arroz se dá em função do crescimento constante da população. O consumo per capita global deste alimento é estimado em cerca de 57 kg/ano (NASCENTE et al., 2019) e estima-se que a demanda de arroz passará de 496 milhões de toneladas em 2020 para 555 milhões de toneladas em 2035 (MARTINS et al., 2020).

Para atender a essa demanda crescente, é preciso buscar estratégias para o aumento da produtividade e qualidade dos grãos, como o desenvolvimento de variedades com maior rendimento a fim de facilitar a reprodução de plantas ideais.

O conhecimento da base genética das variações fisiológicas, de desenvolvimento e morfológicas do arroz é essencial para melhorar a qualidade, o valor nutricional, a confiabilidade e a sustentabilidade deste suprimento alimentar mundial. Um dos fatores genéticos mais importantes que está relacionado à produtividade é a resistência a doenças (RAMALIGAM et al., 2020; NIZOLLI et al., 2021).

Nos últimos anos, a perda de safras de arroz tem aumentado devido à ocorrência de estresses bióticos e abióticos (DIXIT et al., 2020). Dentre os estresses bióticos mais comuns que infectam a cultura de arroz, destaca-se a brusone. Causada por um fungo hemibiotrófico (*Magnaporthe oryzae*), a doença pode resultar em perdas de 100% em cultivares de arroz suscetíveis, limitando sua produtividade (MARTINS et al., 2020). De acordo com Jiang et al. 2020, nas áreas tropicais e temperadas do planeta a brusone é avaliada como uma das doenças mais graves do arroz. Métodos de controle de brusone são utilizados, mas alguns são muito prejudiciais, como o controle químico, que afeta de maneira negativa o meio ambiente e a saúde humana (WIRASWATI et al., 2019).

Tendo em vista a identificação da resistência de *Oryza sativa* a brusone, um método inovador e benéfico é a utilização de Seleção Genômica Ampla (*Genome Wide Selection* – GWS). A GWS visa a predição de valores genéticos genômicos (GEBV - *Genomic Estimated Breeding Value*) dos indivíduos por meio de marcadores moleculares (MEUWISSEN et al., 2001) e permite a seleção precoce de indivíduos geneticamente superiores. Ela é uma alternativa para estimar a probabilidade de a planta ser resistente ou suscetível. Essa característica de classificação binária permite a utilização de vários métodos de estimação, no que diz respeito à acurácia, além de avaliar a sensibilidade e especificidade na predição genômica.

Uma boa ferramenta que pode ser utilizada para avaliar o desempenho de modelos de predição para duas classes (uma suscetível e outra resistente), classificando-as o mais corretamente possível, é a análise da curva ROC (*Receiver Operator Characteristic*). No entanto, os procedimentos adotados ainda são pouco difundidos na área de ciências agrárias e predição genômica.

Neste contexto, objetivou-se utilizar medidas tradicionais para avaliação de acurácia na predição genômica, bem como utilizar a metodologia de curva ROC, para a avaliação do desempenho dos métodos na identificação da resistência do arroz à brusone.

2. MATERIAL E MÉTODOS

2.1. Dados utilizados no estudo

Para a condução do estudo, foi utilizado um banco de dados coletado e disponibilizado por Zhao et al. (2011). O banco dispõe de uma população com 413 plantas de arroz *Oryza sativa* coletadas em 82 países, caracterizando as principais regiões produtoras de arroz no mundo, e contém informações referentes à genotipagem e fenotipagem, permitindo o estudo desses indivíduos.

O banco de dados genotípicos possui um painel com 44.100 marcadores SNPs que, após o controle de qualidade com *call rate* > 70% e MAF > 1% (MAF- *Minor Allele Frequency*), levou à utilização de 36.901 SNPs. Após o *call rate*, os marcadores com dados faltantes (4,33% do total de genótipos) foram imputados de acordo com as respectivas frequências alélicas.

A avaliação fenotípica do arroz foi realizada em Stuttgart (Arkansas, EUA) entre os meses de maio a outubro nos anos de 2006 e 2007. Duas repetições por ano foram cultivadas em um delineamento de blocos inteiramente casualizados em parcelas de 5 m com espaçamento de 25 cm entre as plantas e 50 cm entre as fileiras. No total, o conjunto de dados fenotípicos dispõe de 34 características relacionadas à morfologia das plantas, qualidade dos grãos, desenvolvimento das plantas, qualidade nutricional e ao grau de suscetibilidade do genótipo à doença.

Para este trabalho, foi utilizada a variável do banco de dados referente à doença brusone na folha do arroz. A gravidade da doença foi inicialmente pontuada em uma escala de "0" (sem lesões da doença) a "9" (morte total da planta) quando as plantas tinham entre três e quatro semanas de idade. Esta escala foi convertida para tipos de reação (resistente e suscetível) de acordo com o tamanho e as características das lesões descritas em Mackill; Bonman (1992). Dessa forma, plantas pertencentes às classes 0, 1 e 2, foram classificadas como resistentes (classe 1) e plantas pertencentes às classes 3, 4, 5, 6, 7, 8 e 9 foram classificadas como suscetíveis (classe 0). Foram excluídas 28 plantas não avaliadas quanto à resistência, restando 385 plantas, 282 resistentes e 103 suscetíveis à doença. Mais informações sobre a fenotipagem, genotipagem e controle de qualidade dos dados podem ser consultadas em Zhao et al. (2011).

2.2. Modelagem estatística dos dados

Os tipos de reação (reagente ou susceptível) foram ajustados utilizando o modelo *threshold*. Esse modelo visa estimar a probabilidade de a planta pertencer a uma das duas reações. Essa probabilidade é estimada por uma variável latente dada por:

$$\ell = 1\mu + Zm + e \qquad (01)$$

em que: ℓ = vetor de variáveis latentes em escala gaussiana cujo valor está ligado a uma variável categórica por meio da função de ligação *probit*; Z = relaciona efeitos de SNP's aos valores da resistência à brusone; *m* = vetor de efeito aditivo dos marcadores com matriz de incidência; *e* = vetor de erro associado ao modelo.

Além disso, a matriz de incidência Z foi parametrizada conforme Vitezica et al. (2013), a fim de se enquadrar na teoria de genética quantitativa. Assim:

$$Z = \begin{cases} se\ AA, então\ 2 - 2p_j, \\ se\ Aa, então\ 1 - 2p_j, \\ se\ aa, então\ 0 - 2p_j. \end{cases} \qquad (02)$$

em que: p_j = frequência alélica do marcador j de *AA*, *Aa* e *aa*, que correspondem ao genótipo da planta *i* no marcador *j*, que pode ser homozigoto dominante, heterozigoto ou homozigoto recessivo, respectivamente.

Esse modelo foi ajustado por metodologias bayesianas (RR-BLUP Bayes, Bayes Cπ e BLASSO) para predizer e selecionar indivíduos, utilizando os 36.901 marcadores SNPs.

No método RR-BLUP Bayes (*Bayesian Ridge Regression*) é pressuposta a homogeneidade das variâncias dos SNPs, de modo que apenas um valor de σ^2 é assumido. Os parâmetros de efeito de SNPs (β_i), variância dos marcadores (σ^2) e variância aditiva (σ_u^2) seguem respectivamente: $\beta_i|\sigma^2 \sim N(0, \sigma^2), \sigma^2 \sim \chi^{-2}(\ , S^2)$, em que "*v*" representa os graus de liberdade, S^2 é o parâmetro de escala da distribuição e p_i denota as frequências alélicas. Meuwissen et al. (2001) consideram os valores 4,012 ou 4,2 para "*v*" e 0,002 e 0,0429 para S^2.

O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*), foi proposto por De Los Campos et al. (2009), a partir de uma interpretação bayesiana baseada no LASSO.

As distribuições a priori dada aos parâmetros dos efeitos de marcadores e componentes de variância deste modelo é dada por Pérez; De Los Campos (2014):

$$m_j|\sigma_\epsilon^2, \tau_j, \lambda^2 \sim N(0, \tau_j^2 \cdot \sigma_\epsilon^2); \quad (03)$$
$$\tau_j \sim Exp\left(\frac{\lambda^2}{2}\right); \ \lambda^2 \sim gama(r, s) \quad (04)$$

Por sua vez, o método Bayes Cπ é semelhante ao método BLASSO, mas se difere na distribuição a priori dada ao método, que segundo Pérez; De Los Campos (2014), foi implementada como:

$$m_j|\sigma_m^2, \pi \sim [\pi \cdot N(0, \sigma_m^2) + (1-\pi) \cdot (m_j = 0)]; \quad (05)$$
$$\sigma_m^2 \sim \chi^{-2}(df_m, S_m); \pi \sim beta(p_0, \pi_0) \quad (06)$$

Essa diferença de distribuições *a priori* propicia a seleção de marcadores, uma vez que conduz uma quantidade (1-π) de marcadores a zero.

Para avaliar a qualidade do ajuste, e para que os efeitos dos marcadores não fossem superestimados devido à estimação e validação na mesma amostra (CRUZ et al., 2013), uma técnica de validação cruzada foi adotada. O método de validação cruzada consistiu em dividir a população em k grupos (k = 5). Em seguida, um grupo foi utilizado como população de validação e k-1 grupos foram utilizados como população de estimação. Na população de estimação, os efeitos dos marcadores foram estimados e utilizados na população de validação a fim de obter as estimativas dos valores genéticos genômicos (GEBVs). Esse procedimento foi executado até que cada um dos k grupos fosse utilizado uma vez como população de validação.

Dessa forma, os modelos foram validados pelo procedimento de validação cruzada via Jacknife (5-fold). Como haviam 385 plantas, a melhor maneira de fazer essa validação de forma a obter grupos com o mesmo número de indivíduos foi dividindo-os em 5 grupos de 77 plantas.

2.3. Avaliação dos modelos estatísticos

Ao longo de toda a análise, que envolveu os passos descritos acima, a avaliação dos classificadores ocorreu a partir de medidas como: taxa de erro na validação, correlação de Spearman, viés, tempo de execução computacional e área abaixo da curva ROC (*Area Under the Curve* – AUC).

A taxa de erro na validação cruzada consistiu em somar os indivíduos que não foram reclassificados corretamente em cada população de validação do procedimento de validação cruzada e dividir pela quantidade de populações. A classificação incorreta dos indivíduos se deu a partir da observação dos valores preditos ($\hat{y}_i$) e observados (y_i), onde a predição foi errônea se $y_i \neq \hat{y}_i$.

Após a validação cruzada os GEBVs também foram correlacionados com os valores observados utilizando o coeficiente de correlação de Spearman. Este coeficiente é uma derivação do coeficiente de correlação de Pearson (medida importante na seleção genômica devido ao fato de informar a capacidade preditiva de um modelo), no entanto, não necessita da pressuposição de normalidade das variáveis correlacionadas, pois, seu cálculo é baseado em postos e é calculado da seguinte forma:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \quad (07)$$

em que: d_i = diferença de postos entre $\hat{y}_i$ e y_i; n = número de pares ($\hat{y}_i, y_i$).

Se o posto de $\hat{y}_i$ for igual ao posto de y_i, então o numerador da equação acima será zero e, consequentemente, ρ será igual a 1, indicando uma máxima correlação entre as duas classificações.

As estimativas de viés foram obtidas a partir do coeficiente de regressão das variáveis respostas observadas em função dos GEBVs. Segundo Resende et al. (2014), ter uma estimativa não viesada na seleção genômica é importante quando a seleção envolve indivíduos de muitas gerações utilizando os efeitos dos marcadores obtidos em apenas uma geração. O viés é uma medida de tendência que indica se as estimativas dos parâmetros estão sendo superestimadas, subestimadas ou não possui viés.

O tempo computacional foi obtido com o intuito de saber qual dos métodos possuem maior eficiência computacional combinado ao melhor valor de acurácia. O computador utilizado possui um processador Intel ® Core™ i7-4510U 2.60 GHz com 8,00 GB de memória RAM.

Por fim, os métodos também foram comparados por meio de análise ROC. A análise ROC foi realizada observando qual curva aparenta alcançar um melhor resultado, ou seja, a mais afastada da linha $x = y$.

2.4. Obtenção das estimativas dos parâmetros dos modelos

As estimativas dos parâmetros foram obtidas utilizando o pacote BGLR (PÉREZ; DE LOS CAMPOS, 2014) e a análise ROC foi realizada pelos pacotes GGplot2 e pROC do software R. Nesta análise, foram realizadas 100.000 iterações com burn-in de 20.000 para eliminar o período de aquecimento da cadeia e o thin de 10 para remover o efeito da autocorrelação. Os parâmetros foram preditos por meio de processo iterativo MCMC e a convergência das cadeias de Markov foram avaliadas por meio do critério de Geweke. Este critério indica se a diferença entre a média das nA (10% das n iterações) primeiras iterações com a média das nB (50% das n iterações) últimas iterações seguem distribuição normal com média zero. Mais informações acerca dos métodos utilizados podem ser consultadas em Pérez; De Los Campos (2014).

3. RESULTADOS

A escolha do melhor modelo de predição da resistência à brusone pode ser realizada através da avaliação de suas medidas de ajuste. Na Tabela 1, estão apresentados os valores das medidas de ajuste tradicionais (taxa de erro na validação, correlação de Spearman, viés, tempo de execução da análise), bem como a área sob a curva ROC (AUC) calculadas para cada modelo testado.

Tabela 1. Valores de taxa de erro na validação, correlação de Spearman, viés, tempo de execução da análise e área abaixo da curva ROC (AUC) para os diferentes modelos testados.
Table 1. Values of validation error rate, Spearman correlation, bias, analysis runtime and area under the ROC curve (AUC) for the different models tested.

Medida de ajuste	RR-BLUP	Bayes Cπ	BLASSO
Taxa de erro - validação	0,218	0,218	0,221
Correl. de Spearman	0,476	0,478	0,432
Viés	0,495	0,490	0,471
Tempo de Execução	4h52min	6h1min	6h25min
AUC	0,823	0,822	0,808

O modelo BLASSO teve a maior taxa de erro na validação cruzada (22,1%). Por sua vez, os modelos RR-BLUP e Bayes Cπ obtiveram taxas de erro próximas à do BLASSO e iguais entre si (21,8%). A estimativa da correlação de Spearman entre os valores genéticos reais e preditos pelo modelo RR-BLUP foi semelhante à obtida pelo modelo Bayes Cπ (0,476 e 0,478, respectivamente). Entretanto, em comparação com a estimativa do modelo BLASSO, houve uma diminuição na correlação de Spearman (0,432).

Os modelos RR-BLUP e Bayes Cπ obtiveram valores próximos de viés, ao passo que o BLASSO obteve o menor valor. Em relação ao tempo de execução, esta variável foi menor para o RR-BLUP, seguido dos modelos Bayes Cπ e BLASSO, respectivamente.

Também foram obtidos os valores de AUC para cada modelo. O BLASSO resultou em menor AUC, enquanto os modelos RR-BLUP e Bayes Cπ resultaram em AUC maiores e similares. Para melhor compreensão dos resultados obtidos pela análise ROC, as curvas dos diferentes modelos estudados estão apresentadas na Figura 1.

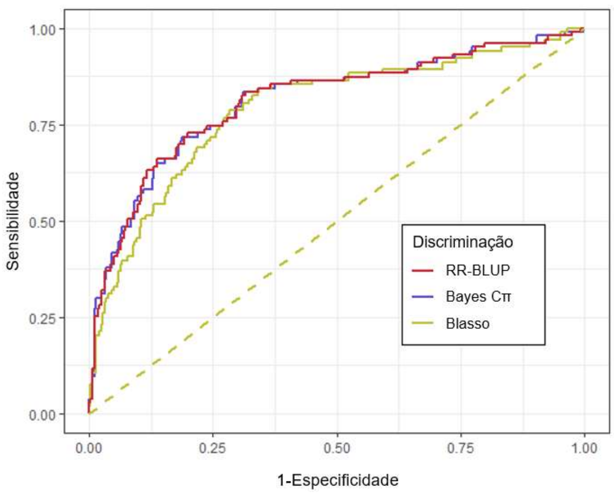

Figura 1. Curvas ROC comparando o desempenho dos modelos RR-BLUP, Bayes Cπ e BLASSO para predizer a resistência à doença brusone do arroz (*Oryza sativa*). Fonte: O autor.
Figure 1. ROC curves comparing the performance of RR-BLUP, Bayes Cπ and BLASSO models for predicting resistance to brusone disease of rice (*Oryza sativa*). Source: The author.

Observa-se que em valores menores de 1-Especificidade o modelo BLASSO apresenta menor especificidade, enquanto os demais apresentam comportamento similar. Acima de 0,25 de 1-Especificidade todos os modelos se igualam em relação à sensibilidade.

4. DISCUSSÃO

Biscarini et al. (2014) também utilizaram a taxa de erro na validação cruzada para avaliar o vigor da raiz de beterraba açucareira com dois níveis (alto ou baixo) utilizando o modelo *threshold* sob o método G-BLUP. A taxa de erro na validação cruzada encontrada por estes autores foi de 0,073%.

Quanto mais próxima de zero for a taxa de erro de validação, melhor será a acurácia do modelo. Porém, para que essa medida tenha uma interpretação igual à acurácia na seleção genômica, deve ser calculada a taxa de acerto, dada por: Taxa de acerto = 1 - taxa de erro. Os valores da taxa de acerto no RR-BLUP e Bayes Cπ foram 0,782 (78,2%) e no método BLASSO de 0,779 (ou 77,9%). Assim, por meio da análise desta medida, o método BLASSO obteve menor acurácia em relação aos demais modelos. Rutkoski et al. (2014) utilizaram os modelos G-BLUP, BLASSO e Bayes Cπ para predizer a resistência de hastes de trigo à ferrugem. Os autores encontraram taxas de acerto na validação cruzada semelhantes entre BLASSO (0,579) e Bayes Cπ (0,578), enquanto o G-BLUP foi o menos acurado dentre os três modelos (0,568). Gianola (2013) também discute que os diversos modelos bayesianos possuem acurácias similares para fins preditivos.

Em relação à estimativa da correlação de Spearman, os maiores valores encontrados para os modelos RR-BLUP e Bayes Cπ indicam adequabilidade (em termos de comparação de métodos) com a taxa de erro na validação, tendo em vista que, quanto mais próximo de 1, melhor será a correlação entre os valores preditos e os valores genéticos reais. Esses resultados estão de acordo com Bhering et al. (2015) que, ao avaliar métodos para selecionar indivíduos superiores em melhoramento de plantas, encontraram que o RR-BLUP e G-BLUP foram equivalentes e obtiveram maiores valores de correlação de Spearman. Por sua vez, o BLASSO obteve os valores mais baixos.

Os valores do viés das estimativas em cada modelo foram obtidos em cada grupo da população de validação no procedimento de validação cruzada e posteriormente, foi obtida a média das populações. O viés de um modelo indica sua tendenciosidade e para que um modelo não seja tendencioso esse valor deve ser igual a 1. Quando o valor de viés é maior que 1, há indício de que o modelo apresenta subestimação à variável resposta e, quando for menor que 1, tem-se evidencia que o modelo está superestimando a variável resposta. Foi possível observar que todos os modelos superestimam a resistência a doença brusone. Entretanto, nos modelos RR-BLUP e Bayes Cπ, o viés foi mais próximo de 1, indicando que, quando a resistência à brusone foi avaliada pelo modelo BLASSO, as estimativas produzidas foram mais tendenciosas. Os resultados encontrados estão de acordo com Azevedo et al. (2015), que discutem que a análise do viés é uma das principais formas de diferenciar a capacidade de predição de métodos bayesianos.

Por sua vez, o tempo de execução refere-se ao tempo gasto para os métodos estimarem os parâmetros do modelo. Em relação ao tempo de execução, foi obtido que RR-BLUP (4h52min) < Bayes Cπ (6h1min) < BLASSO (6h25min). Essa diferença no tempo de execução é explicada pelo número de parâmetros que foram estimados em cada modelo. Como os métodos RRBLUP, Bayes Cπ e BLASSO são a nível de marcador, o vetor de parâmetros a ser estimado foi $\boldsymbol{m}$ (36.901×1) com 36.901 parâmetros de efeitos de marcadores mais os parâmetros de variância. No entanto, o RR-BLUP é um método que pressupõe homogeneidade das variâncias, ao contrário dos outros dois métodos que supõem uma variância para cada marcador, assim, quanto menos parâmetros precisam ser estimados, menor é o tempo gasto para estimá-los. Esse resultado está de acordo com o encontrado por Bhering et al. (2015), que observaram menor tempo de execução para o RR-BLUP, quando comparado aos modelos BLASSO e G-BLUP.

Os valores de AUC, assim como a taxa de erro, são variáveis aleatórias estimadas por validação cruzada. Foi possível observar por meio da AUC que todos os modelos têm elevado poder de discriminação (>0,80) (HOSMER et al., 2013). Os modelos RR-BLUP e Bayes Cπ apresentam melhor performance (AUC 0,823 e 0,822, respectivamente) quando comparado ao BLASSO (AUC = 0,808).

De acordo com Ling et al. (2003), a AUC é mais precisa do que a taxa de erro. Mas sempre que possível, é recomendável plotar e analisar a curva dos modelos. Por meio das curvas ROC apresentadas na Figura 1, a performance dos modelos pode ser observada e avaliada de maneira mais segura, uma vez que a comparação não é feita apenas a partir de uma medida. Pode-se perceber que, em níveis de 1-especificidade abaixo de 0,25, o modelo derivado do método BLASSO obteve pior desempenho em relação aos modelos derivados dos métodos RR-BLUP e Bayes Cπ, por apresentar a curva mais próxima da diagonal principal. Isso significa que o BLASSO gera mais falsos positivos em menores níveis de 1-especificidade. Esse resultado pode ser explicado em termos dos QTLs (*quantitative trait loci*) de efeito zero. É desejável que os modelos de predição genômica sejam capazes de reduzir efetivamente os QTLs de efeito zero para muito perto de zero. Entretanto, o BLASSO não é capaz de reduzir efetivamente tal efeito (FANG et al., 2012), o que explica sua menor sensibilidade.

Em níveis de 1-especificidade acima de 0,25, os valores de sensibilidade dos três modelos se igualam, o que condiz com as demais medidas avaliadas. Destaca-se, portanto, a importância de avaliar toda a curva ROC, ao invés da observação isolada dos valores de AUC. Deste modo, mais informações acerca da performance dos modelos podem ser obtidas.

5. CONCLUSÕES

A área abaixo da curva ROC se mostrou equivalente às medidas de ajuste tradicionais para avaliar os modelos. Dentre eles, o RR-BLUP e Bayes Cπ foram os mais acurados para a predição de resistência de arroz à brusone. Além disso, o modelo RR-BLUP obteve menor tempo de execução.

Pela análise gráfica das curvas ROC, percebeu-se que o modelo BLASSO obteve menor desempenho que os demais em altos níveis de especificidade (>0,75). Em contrapartida, para valores de especificidade abaixo de 0,75, os modelos apresentaram valores de sensibilidade similares.

6. AGRADECIMENTOS

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – Brasil) – Código de Financiamento 001.

7. REFERÊNCIAS

AZEVEDO, C. F.; DE RESENDE, M. D. V.; E SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S. F.; RESENDE JR., M. F. R.; MUÑOZ, P. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics**, v. 16, n. 1, p. 1-13, 2015. https://doi.org/10.1186/s12863-015-0264-2

BHERING, L. L.; JUNQUEIRA, V. S.; PEIXOTO, L. A.; CRUZ, C. D.; LAVIOLA, B. G. Comparison of methods used to identify superior individuals in genomic selection in plant breeding. **Genetics and Molecular Research**, v. 14, n. 3, p. 10888-10896, 2015. https://doi.org/10.4238/2015.september.9.26

BISCARINI, F.; STEVANATO, P.; BROCCANELLO, C.; STELLA, A.; SACCOMANI, M. Genome-enabled predictions for binomial traits in sugar beet populations. **BMC Genetics**, v. 15, n. 1, p. 87, 2014. https://doi.org/10.1186/1471-2156-15-87

CRUZ, C. D.; SALGADO, C. S.; BHERING, L. L. **Genômica aplicada**. Visconde de Rio Branco: Suprema Gráfica Editora, 2013. 424p.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. **Genetics**, v. 182, n. 1, p. 375-385, 2009. https://doi.org/10.1534/genetics.109.101501

DIXIT, S.; SINGH, U. M.; SINGH, A. K.; ALAM, S.; VENKATESHWARLU, C.; NACHIMUTHU, V. V.; YADAV, S.; ABBAI, R.; SELVARAJ, R.; DEVI, M. N.; RAMAYYA, P. J.; BADRI, J.; LAKSHMI, T. R. J.; LAKSHMIDEVI, G.; VIDHYA J. L. R. K.; PADMAKUMARI, A. P.; LAHA, G. S.; PRASAD, M. S.; SEETALAM, M.; SINGH, V. K.; KUMAR, A. Marker Assisted Forward Breeding to Combine Multiple Biotic-Abiotic Stress Resistance/Tolerance in Rice. **Rice**, v. 13, n. 29, p. 1-15, 2020. https://doi.org/10.1186/s12284-020-00391-7

FANG, M.; JIANG, D.; LI, D.; YANG, R.; FU, W.; PU, L.; GAO, H.; WANG, G.; YU, L. Improved LASSO priors for shrinkage quantitative trait loci mapping. **Theoretical and Applied Genetics**, v. 124, n. 7, p. 1315-1324, 2012. https://doi.org/10.1007/s00122-012-1789-7

GADAL, N.; SHERSTHA, J.; POUDEL, M. N.; POKHAREL, B.; A review on production status and growing environments of rice in Nepal and in the world. **Archives of Agriculture and Environmental Science**, v. 4, n. 1, p. 83-87, 2019. https://doi.org/10.26832/24566632.2019.0401013

GIANOLA, D. Priors in whole-genome regression: the bayesian alphabet returns. **Genetics**, v. 194, p. 573-596, 2013. DOI: 10.1534%2Fgenetics.113.151753

HOSMER JR., D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3º ed. Hoboken: John Wiley & Sons, Inc., 2013. 510p.

JIANG, H.; FENG, Y.; QIU, L.; GAO, G.; ZHANG, Q.; HE, Y. Identification of Blast Resistance QTLs Based on Two Advanced Backcross Populations in Rice. **Rice**, v. 13, n. 31, 1-15, 2020. https://doi.org/10.1186/s12284-020-00392-6

LING, C. X.; HUANG, J.; ZHANG, H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In: In: Xiang Y., Chaib-draa B. (Ed.) **Advances in Artificial Intelligence**. Canadá, 2003. p. 329-341. https://doi.org/10.1007/3-540-44886-1_25

MACKILL, A. O.; BONMAN, J. M. Inheritance of blast resistance in near-isogenic lines of rice. **Phytopathology**, v. 82, p. 746-749, 1992. https://doi.org/10.1590/S0100-84551996000400012

MARTINS, B. E. M.; CHAIBUB, A. A.; CORTÊS, M. V. C. B.; SILVA LOBO, V. L.; FILIPPI, M. C. C. Characterization of bacterial isolates for sustainable rice blast control. **Revista Caatinga**, v. 33, n. 3, p. 702-712, 2020. https://doi.org/10.1590/1983-21252020v33n313rc

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. https://doi.org/10.1093/genetics/157.4.1819

NASCENTE, A. S.; FILIPPI, M. C. C.; SOUSA, T. P.; CHAIBUB, A. A.; SOUZA, A. C. A.; LANNA, A. C. Upland rice gas exchange, nutrient uptake and grain yield as affected by potassium fertilization and inoculation of the diazotrophic bacteria Serratia spp. **Australian Journal of Crop Science**, v. 13, n. 6, p. 944-953, 2019. https://doi.org/10.21475/ajcs.19.13.06.p1689

NIZOLLI, V. O.; PEGORARO, C.; DE OLIVEIRA, A. C. Rice blast: strategies and challenges for improving genetic resistance**. Crop Breeding Applied Biotechnology**, v. 21, e387721S9, 2021. http://dx.doi.org/10.1590/1984-70332021v21Sa22

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483-495, 2014. https://doi.org/10.1534/genetics.114.164442

RAMALINGAM, J.; RAVEENDRA, C.; SAVITHA, P.; VIDYA, V.; CHAITHRA, T. L.; VELPRABAKARAN, S.; SARASWATHI, R.; RAMANATHAN, A.; PILLAI, M. P. A.; ARUMUGACHAMY, S.; VANNIARAJAN, C. Gene Pyramiding for Achieving Enhanced Resistance to Bacterial Blight, Blast, and Sheath Blight Diseases in Rice. **Frontiers in Plant Science**, v. 11, 2020. https://doi.org/10.3389/fpls.2020.591457

RATHNA, P. T. S.; NELSON, A. R. L. E.; RAVICHANDRAN, K.; ANTONY, R. Nutritional and functional properties of coloured rice varieties of South India: a review. **Journal of Ethnic Foods**, v. 6, p. 7-11, 2019. https://doi.org/10.1186/s42779-019-0017-3

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. 1. ed. [s.l.] : Suprema Gráfica Editora, 2014. 881p.

RUTKOSKI, J. E.; POLAND, J. A.; SINGH, R. P.; et al. Genomic selection for quantitative adult plant stem rust resistance in wheat. **The Plant Genome**, v. 7, n. 3, p. 1-10, 2014. DOI: 10.3835/plantgenome2014.02.0006

SOUZA, D. C.; RUSSINI, A.; VARGAS, R. R.; BOTTEGA, E. L.; SCHLOSSER, J. F.; FARIAS, M. S. Determinação de área mínima para aquisição de colhedora para a colheita do arroz irrigado na fronteira oeste do Rio Grande do Sul. **Tecno-Lógica**, v. 24, n. 2, p. 208-214, 2020. http://dx.doi.org/10.17058/tecnolog.v24i2.14904

VERMA, D. K.; SRIVASTAV, P. P. Bioactive compounds of rice (Oryza sativa L.): Review on paradigm and its potential benefit in human health. **Trends in Food Science & Technology**, v. 97, p. 355-365, 2020. https://doi.org/10.1016/j.tifs.2020.01.007

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, v. 195, n. 4, p. 1223-1230, 2013. https://doi.org/10.1534/genetics.113.155176

WIRASWATI, S. M.; RUSMANA, I.; NAWANGSIH, A. A.; WAHYUDI, A. T. Antifungal activities of bacteria producing bioactive compounds isolated from rice phyllosphere against Pyricularia oryzae. **Journal of Plant Protection Research**, v. 59, n. 1, p. 86-94, 2019. https://doi.org/10.24425/jppr.2019.126047

ZHAO, K.; TUNG, C.-W.; EIZENGA, G. C.; et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature Communications**, v. 2, n. 1, p. 467, 2011. https://doi.org/10.1038/ncomms1467.