GMR
The Original by FUNPEC-RP

Regression trees in genomic selection for carcass traits in pigs

L.S. Silveira[1], L.P. Lima[1], M. Nascimento[1], A.C.C. Nascimento[1] and F.F. Silva[2]

1 Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
2 Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil

Corresponding author: L.S. Silveira
E-mail: lucasvrb@hotmail.com

Genet. Mol. Res. 19 (1): gmr18498
Received October 05, 2019
Accepted December 07, 2019
Published January 31, 2020
DOI http://dx.doi.org/10.4238/gmr18498

ABSTRACT. Genome-Wide Selection (GWS) uses molecular markers to predict the genetic merit of animals and plants. Usually, a high density of molecular markers to predict this genetic merit is used. Thus, statistical methods need to deal with problems of high dimensionality, multicollinearity and computational efficiency. We examined a set of machine learning methods, in particular the tree-based regression methods (Regression Tree, Bagging, Random Forest and Boosting) and evaluated them in relation to predictive ability and bias. Moreover, these methods were compared with the Bayesian Least Absolute Shrinkage and Selection Operator (BLASSO) method, which is routinely used in GWS. For this, we used information of 10 carcass traits in Piau x Commercial pigs. The tree-based regression methods were superior to the BLASSO method, presenting shorter computational times to predict the genetic values of the analyses, especially, the Random Forest and Bagging methods. Furthermore, the predictive abilities of tree-based regression methods were competitive with BLASSO. In terms of bias, the BLASSO underestimated the predictions while tree-based regression methods overestimated the predictions. In addition, among the methods, the Random Forest was the one that obtained the bias value closest to ideal for most of the traits, demonstrating the superiority of this method.

Key words: Crop breeding; Genetic improvement; Tree-based regression methods

INTRODUCTION

The Genomic-Wide Selection (GWS), proposed by Meuwissen et al. (2001), consists in the use of molecular markers to predict the genetic merit of animals and plants using the subsequent selection of individuals. Among the advantages of GWS, the genetic gain per unit of time, low cost, high efficiency in selection of genetically superior individual and reduction of generation interval in the selection of best individuals are preeminent (Meuwissen et al., 2001; Goddard and Hayes, 2008).

Many statistical methods have been proposed, tested and used to improve the prediction of genetic values. Among these methodologies there are those based on Bayesian inference (Meuwissen et al., 2001; Gianola, 2013), dimensional reduction (Azevedo et al., 2015a), nonlinear regression (González-Camacho et al., 2012) and tree-based regression and its refinements, such as Bagging, Random Forest and Boosting (Ogutu et al., 2011; Ho et al., 2019). Tree-based regression methods are still little used in GWS, though they have useful features, such as easy interpretation, they deal with quality variables without the need to create dummy variables and can be used for both regression and classification (James et al., 2013). Moreover, this approach does not require assumptions concerning the distribution of model parameters and variable response.

Tree-based methods are algorithms that partition the predictor space in subspaces based on some specifications. These subspaces are also divided until they reach an established stopping criterion. From this, the predicted value for a new individual is obtained by the average of trained individuals in the region in which the new individual belongs (James et al., 2013).

González-Recio and Forni (2011) compared Boosting and Random Forest methods with other Bayesian methods (Bayes A and Bayesian LASSO - BLASSO) to predict categorical traits. They concluded that the best method might depend on the genetic architecture of the population and found higher prediction accuracy in the adjustment by Boosting and BLASSO and the best performance in the correct classification of individuals using Random Forest. Bayes A and Boosting methods had the best accuracy for hybrids. However, for quantitative traits in GWS, there are few studies using tree-based regressions. Therefore these methods should be analyzed with this group of variables, since most important agronomic traits are of this nature (Yang et al., 2010).

We evaluated the methods used in tree-based regressions (Regression Tree, Bagging, Random Forest and Boosting) and compared them with the BLASSO method, which is the most widely used in GWS studies. For this, a set with 10 quantitative traits of economic importance in pigs were used. These methods were compared by predictive ability, bias in prediction and computational time after the cross-validation procedure.

MATERIAL AND METHODS

Dataset

The dataset used in this study come from the Pig Breeding Farm of the Department of Zootechnics of the Federal University of Viçosa, Viçosa, Minas Gerais, collected from November 1998 to July 2001. These records came from an F2 population that consisted of 345 pigs originated from the crossing of two boars of the Brazilian Piau naturalized breed, with 18 females of the UFV strain, by the mating of commercial animals (Landrance x Larga White x Pietrain). DNA extraction was performed according to Peixoto et al. (2006). We used 237 SNPs markers that are distributed as follows on the *Sus scrofa* chromosomes: SSC1 (56), SSC4 (54), SSC7 (59), SSC8 (31), SSC17 (25).

Carcass traits are of great interest in the development of pig farming, seeking to increasingly meet the demands of the consumer market regarding lower fat deposition and higher yield and carcass length (Bertol et al., 2010). Thus, in this study, the following traits were considered: carcass yield (CY); backfat thickness (BFT); Midline backfat thickness immediately after the last rib (LR); backfat thickness on the shoulder region (SBT); backfat thickness after the last rib, 6.5 cm from the midline (ETO); midline backfat thickness between last and next to last but one lumbar vertebra (LL); midline lower backfat thickness above the last lumbar vertebra (L); carcass length by the Brazilian carcass classification method (MBCC); carcass length by the American carcass classification method (MLC); diameter of the longissimus dorsi muscle in the region of the last rib at 6.5 cm from the dorsolumbar line, from a transverse section in the carriage (PROLOM).

In order to predict the genetic merit of unrelated individuals, the data used in the analyses were corrected for fixed parent effects. According to Resende et al. (2012) without this correction, molecular markers may be capturing only kinship between individuals, which could reduce the accuracy of validation in individuals from independent populations or from other generations.

Blasso

The basic linear model proposed by Meuwissen et al. (2001) to estimate genetic merit is given by:

$$y = 1\mu + Wm_a + e \qquad \text{(Eq. 1)}$$

where y is the phenotype vector, μ is the general mean of the trait, m_a is the additive genetic effects vector with incidence matrix of markers W and e is the residual vector. The Bayesian Least Absolute Shrinkage and Selection Operator (BLASSO) includes a common variance term for the genetic effects of markers and residual effects. In this method, the effects of markers follow a double exponential distribution, as follows: $m_i|\lambda^2 \sim ExpDupla\left(0,\frac{\sigma}{\lambda}\right)$ the additive genetic variance of each marker is given respectively by $\sigma_{mi}^2 = \tau_i^2\sigma^2$ with $i = 1,2,\dots,m$. Additive genomic values are estimated from expression: $\hat{a} = W\hat{m}$. The complete conditional distributions a posteriori for the BLASSO parameters are presented in detail by de los Campos et al. (2009). In this work, 41,000 iterations were used for MCMC (Markov chain Monte Carlo) algorithms, which 1,000 were burn-in to ensure chain heating and one in 10 iterations were selected (thin). Convergence analysis was performed using the criterion proposed by Raftery and Lewis (1992), using the Boa software package R (Smith, 2007).

Regression Tree, Bagging, Random Forest and Boosting

A Regression Tree is constructed by a process known as recursive binary division, which is an iterative procedure that divides training data into partitions or branches and then remains dividing each partition into smaller groups. Initially, the predictor W_j and the division point s are selected. Its separates the predictor space into two regions $\{W|W_j < s\}$ and $\{W|W_j \geq s\}$, which leads to the largest possible reduction in the Residual Square Sum (RSS). Thus, all predictors are considered $W_1,\dots,W_p$ and all possible split point values s for each predictor, then the predictor and split point is chosen in a way that the resulting tree has the smallest RSS. Therefore, for any j and s, the regions are defined as $R_1(j,s) = \{W|W_j < s\}$ and $R_2\,(j,s) = \{W|W_j \geq s\}$ and we seek the value of j and s that minimize the equation:

$$\sum_{i:\, w_i\in R_1(j,s)}(y_i-\hat{y}_{R_1})^2+\sum_{i:\, w_i\in R_2(j,s)}(y_i-\hat{y}_{R_2})^2 \tag{Eq. 2}$$

where $\hat{y}_{R_1}$ is the average response for training observations in $R_1(j,s)$ and $\hat{y}_{R_2}$ is the average response for training observations in $R_2(j,s)$.

This process continues until achieve a stopping criterion, for example, until any region contains no more than five observations. Once regions $R_1,\dots,R_J$ are created, the response is predicted for a given test set observation, using the average of the training observations in the region to which this test observation belongs.

Bagging also known as Bootstrap Aggregation has been proposed to reduce the variance of the Regression Tree method (James et al., 2013). For this, using bootstrap sampling in the training population, B subsets are generated with replenishment of the available sampling, obtaining B models $\hat{f}^1(\cdot),\dots,\hat{f}^B(\cdot)$which will be used in the construction of B trees. According to James et al. (2013), predictions by this method are obtained by:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b*}(x) \tag{Eq. 3}$$

where x is a recorded set of unused predictors in the training population, and $\hat{f}^{b*}(\cdot)$ is the predictor function of the previously trained b-th tree.

Each tree of this method has high variance and low bias because it is built deep (with many divisions) and not pruned (James et al., 2013). However, by averaging the results obtained on B trees, the variance is reduced. Moreover, in practice, increasing B reduces the error without leading to overfitting, which is estimated by OOB (Out-of-Bag). That is, in bootstrap sampling, the remaining unsampled training population data (one third of each sample) is used as a test population to obtain the error.

Random Forest, on the other hand, was proposed by Breiman (2001) and uses a set of bootstrap samples in the training population to build several trees, which one tree is built at each sampling. Each tree is constructed by randomly selecting a subset of predictors as candidates for region division. A variable in the validation population containing the x records for predictors is classified according to the mean of the B trees built, so that:

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x, \Psi_{\mathrm{b}}) \qquad \text{(Eq. 4)}$$

where $T(\cdot)$ is the prediction function according to variable x and Ψ_b , where Ψ_b is a set of parameters that characterize tree b in terms of division variable, cutoff point in each region, and predicted values for the terminal regions. According to James et al. (2013), the main difference between Random Forest and Bagging is that the Random Forest considers a p number of variables (p <m) to predict the averages, in order to reduce the variance of the predictions obtained in the Regression Tree method. If Random Forest is built using p = m then it is equivalent to Bagging.

In addition, Boosting, as described by James et al. (2013), uses regression trees adjusting the residual of an initial model. The residual is updated in each tree, that grows sequentially from the residual of the previous tree and, just like in Bagging, the response variable of Boosting involves a combination of a large number of trees, so that $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x)$. The function $\hat{f}(\cdot)$ refers to the final tree combined with the sequentially adjusted B trees and λ is the shirinkage parameter that controls the learning rate of the method. In addition, this method needs to be adjusted with a d number of divisions in each tree. This parameter controls the complexity of boosting and is known as depth.

Cross validation, predictive ability and bias

The methods were compared using a cross-validation study. This study is based on population division (345 observations) into 5 groups with 69 observations each. One of the groups was used as the validation population and the remaining groups as the training population. In the training population, the model parameters were estimated and in the validation population, predictive ability and bias values were obtained. All groups were used as the validation population and this provided a mean and, also, a standard error for the predictive ability and bias values. Predictive ability was calculated as Pearson's correlation between observed and predicted values. And, the bias was obtained from the linear regression coefficient between the observed value and the predicted value. In this measure, the target value is 1, so values less than 1 indicate that genetic values are being overestimated and values greater than 1 indicate that they are being underestimated.

Computational Resources

The analyses were implemented in R software (R Development Core Team, 2018) using the packages presented in Table 1.

Table 1. R software packages and functions used in the analysis.

Methodologies	Packages	Functions
BLASSO	BGLR	BGLR
Regression Tree	tree	Tree
Bagging	randomForest	randomForest
Random Forest	randomForest	randomForest
Boosting	gbm	Gbm

RESULTS AND DISCUSSION

The results of mean of the predictive ability and standard errors obtained in cross validation procedure to Regression Tree, Bagging, Random Forest, Boosting and BLASSO methods for each carcass trait are presented in Table 2.

Table 2. Mean of predictive ability values (Pearson correlation) and standard errors in cross validation procedure of the BLASSO, Regression Tree, Bagging, Random Forest and Boosting methods for each carcass trait.

Traits	BLASSO	Regression Tree	Bagging	Random Forest	Boosting
CY	0.16 ± 0.07	0.08 ± 0.05	0.07 ± 0.06	0.10 ± 0.06	0.06 ± 0.06
MBCC	0.25 ± 0.05	0.10 ± 0.05	0.23 ± 0.02	0.25 ± 0.03	0.19 ± 0.03
MLC	0.26 ± 0.02	0.10 ± 0.05	0.23 ± 0.03	0.24 ± 0.04	0.15 ± 0.06
SBT	0.16 ± 0.04	0.12 ± 0.06	0.18 ± 0.04	0.18 ± 0.03	0.12 ± 0.04
LR	0.26 ± 0.04	0.15 ± 0.03	0.27 ± 0.02	0.30 ± 0.03	0.26 ± 0.05
LL	0.29 ± 0.04	0.13 ± 0.05	0.22 ± 0.03	0.22 ± 0.04	0.17 ± 0.04
L	0.23 ± 0.04	0.06 ± 0.04	0.23 ± 0.05	0.23 ± 0.05	0.12 ± 0.05
ETO	0.25 ± 0.04	0.03 ± 0.06	0.21 ± 0.03	0.22 ± 0.04	0.12 ± 0.02
BFT	0.27 ± 0.05	0.05 ± 0.07	0.23 ± 0.03	0.25 ± 0.02	0.17 ± 0.01
PROLOM	0.15 ± 0.07	0.04 ± 0.06	0.19 ± 0.05	0.20 ± 0.06	0.12 ± 0.08

CY = Carcass yield (%); BFT = Backfat thickness (mm); LR = Midline backfat thickness immediately after the last rib (mm); SBT = higher backfat thickness on the shoulder region (mm); ETO = backfat thickness after the last rib, 6.5 cm from the midline (mm); LL = midline backfat thickness between last and next to last but one lumbar vertebrae (mm); L = midline lower backfat thickness above the last lumbar vertebrae; MBCC = carcass length by the Brazilian carcass classification method; MLC = carcass length by the American carcass classification method; PROLOM = diameter of the longissimus dorsi muscle in the region of the last rib at 6.5 cm from the dorsolumbar line, from a transverse section in the carriage.

For most traits analyzed, the BLASSO, Bagging and Random Forest presented the highest values of predictive ability and these values were very similar among the three methodologies. The BLASSO highlighted for the CY and LL traits. However, for PROLOM, the Random Forest provided the highest predictive ability followed by Bagging. The Boosting method did not show superiority in any trait analyzed here, although it presented predictive ability values similar to the BLASSO, Bagging and Random Forest for the LR. The Bagging method obtained estimates very close to the Random Forest, but did not exceed it by any estimate. The Regression Tree method was always inferior to the other methods, presenting lower estimates of predictive ability in most traits.

The BLASSO was chosen to be compared with the tree-based regression methods due to the good results that has been obtaining in the GWS. We can cite Teixeira et al.

(2016), who proposed and evaluated the use of factor analysis in the same dataset and found that the accuracy in the selection (predictive ability divided by the square root of the heritability of the trait) obtained by BLASSO outperformed the other methods considered by them. Also, Santos et al. (2018) evaluated three asymmetric traits of this same dataset using quantile regression and compared it with the BLASSO, presenting the BLASSO with accuracy in the selection similar to those obtained by quantile regression. Additionally, the good performance of the BLASSO was expected, according with the results observed in the study conducted by de los Campos et al. (2009).

Additionally, the Bagging and Random Forest methods were competitive with each other for all traits, which can be explained by the fact that Random Forest is a particular case of Bagging. According to James et al. (2013) an improvement in prediction using these methods can be obtained by increasing the number of tree combinations. However, the disadvantage of this improvement is the loss of interpretation in the results. While in the Bagging 237 trees were used, in the Random Forest only 79 were used, and, how more trees are used less interpretation we have.

The inferiority of the Regression Tree method can also be explained by James et al. (2013), who say that Regression Tree results are complex which may lead to good predictions in the training population, but not so good results in the validation population. According to James et al. (2013), an alternative to improve this method is obtained by using a reduction in tree construction. This reduction can occur during its construction, limiting itself to a region where the division no longer causes a significant reduction in the SSR, or alternatively, building a larger tree and, from its, obtain smaller trees leading to a lower error rate in the validation.

Table 3 presents the mean of bias obtained with the cross-validation procedure for each trait. The standard error of the mean is also presented.

Table 3. Mean of Bias values (regression coefficient) and standard errors in cross validation procedure of the BLASSO, Regression Tree, Bagging, Random Forest and Boosting methods for each carcass trait.

Traits	**BLASSO**	**Regression Tree**	**Bagging**	**Random Forest**	**Boosting**
CY	1.51 ± 0.65	0.10 ± 0.06	0.19 ± 0.11	0.33 ± 0.18	0.04 ± 0.08
MBCC	1.11 ± 0.30	0.15 ± 0.09	0.72 ± 0.11	0.90 ± 0.17	0.27 ± 0.04
MLC	0.94 ± 0.12	0.12 ± 0.06	0.77 ± 0.15	0.86 ± 0.15	0.19 ± 0.08
SBT	1.43 ± 0.45	0.17 ± 0.08	0.74 ± 0.17	0.81 ± 0.17	0.17 ± 0.05
LR	1.27 ± 0.25	0.18 ± 0.03	0.99 ± 0.16	1.17 ± 0.14	0.35 ± 0.06
LL	1.23 ± 0.32	0.19 ± 0.10	0.75 ± 0.20	0.81 ± 0.26	0.24 ± 0.08
L	0.90 ± 0.20	0.06 ± 0.03	0.73 ± 0.19	0.76 ± 0.21	0.15 ± 0.07
ETO	1.30 ± 0.25	0.05 ± 0.08	0.70 ± 0.13	0.81 ± 0.16	0.16 ± 0.02
BFT	1.23 ± 0.32	0.06 ± 0.03	0.74 ± 0.17	0.84 ± 0.06	0.22 ± 0.02
PROLOM	1.93 ± 1.06	0.06 ± 0.09	0.59 ± 0.16	0.73 ± 0.22	0.16 ± 0.11

CY = Carcass yield (%); BFT = Backfat thickness (mm); LR = Midline backfat thickness immediately after the last rib (mm); SBT = higher backfat thickness on the shoulder region (mm); ETO = backfat thickness after the last rib, 6.5 cm from the midline (mm); LL = midline backfat thickness between last and next to last but one lumbar vertebrae (mm); L = midline lower backfat thickness above the last lumbar vertebrae; MBCC = carcass length by the Brazilian carcass classification method; MLC = carcass length by the American carcass classification method; PROLOM = diameter of the longissimus dorsi muscle in the region of the last rib at 6.5 cm from the dorsolumbar line, from a transverse section in the carriage.

The tree-based regression methods overestimated genetic values, with values less than 1 (one) for almost all traits except the Random Forest, which underestimated this value for the LR trait. The BLASSO method gave, in most traits, bias values greater than 1 (one), underestimating these values except for the MLC and L traits where the method obtains values smaller than one. Among the tree-based regression methods, Boosting and Regression Tree again can be considered inferior because they are the methods that presented more biased values for all traits obtaining values farthest from 1 (one). For CY, MLC and L traits, BLASSO and Random Forest were the methods that presented the closest estimates from 1 (one), which is considered the ideal value in the GWS studies (Azevedo et al. 2015a). For the LR trait, the Bagging method presented the closest value from 1 (one). In addition, for the other six remaining traits, the Random Forest method was the one that presented values closest to the ideal.

Prediction bias is of great importance in genomic selection especially when it comes to quantitative traits since selecting individuals with overestimated genetic values can lead to large economic losses. The same is true when considering individuals with underestimated genetic values. Thus, it is usual to consider bias in GWS studies for genetic improvement as in the studies by Azevedo et al. (2015b) in animal breeding and Sousa et al. (2019) in plant breeding.

For quantitative traits Ogutu et al. (2011) compared the Random Forest, Boosting and Support Vector Machines methods applied to a simulated dataset and concluded that the accuracy of the Boosting method was better, however, these methods had little differences between them. Gonzalez-Camacho et al. (2018) make a brief discussion about machine learning methods, taking Random Forest as one of the methods and conclude that the flexibility of these methodologies allows them to become a good alternative to parametric methods like BLASSO. Also, according to Gonzalez-Camacho et al. (2018), the Random Forest captures complex feature interactions and is robust to overfitting. Gonzalez-Recio et al. (2014) suggest the use of Support Vector Machine (SVM) and Random Forest for classification problems and Reproducing Kernel Hilbert Space (RKHS) and Boosting for regression problems. However, they comment that each study requires a particular study about the traits analyzed.

In addition, in other work using the same dataset, Costa et al. (2015) who combined genomic data with pedigree data in the same individuals with carcass and growth phenotypic traits and studied the relative importance of additive and dominance genetic variation using the G-BLUP model. These authors concluded that dividing genetic variance into additive and dominance variance improves knowledge about the genetic control of the trait. So far, in our studies we have seen that the Random Forest and Bagging methods have obtained better estimates among the tree-based regression methods, however, the effects used in the models are only additive.

A major advantage of tree-based regression methods was their computational agility (Table 4). In our study, the computational time of BLASSO was 45% greater than Boosting performed with 10,000 trees. However, the computational time for Boosting may increase according to the number of trees used. In addition, BLASSO execution time was almost five times greater than the Bagging method and more than

11 times greater than in Random Forest. The Regression Tree method was the one with the shortest computational time, spending less than 4 seconds to analyze all traits. These differences in big datasets can become a major problem in studies that need to be done quickly. In addition, the number of iterations required for convergence of Markov Chains can increase, and consequently, the time to run the BLASSO. In this case, tree-based regression methods can be more appropriate.

With respect to GWS in pigs, Samorè and Fontanesi (2016) made a study about the advances and difficulties found in recent times. Among the difficulties are the high cost of genotyping compared to the individual value of an animal, little time for genetic evaluation and pyramidal population structures that affect the number of genotyped and phenotyped animals. However, as advantages, the consanguinity control, the selection among full siblings, the way of storing biological material and reducing the generation interval are improved. In this study it was further emphasized that carcass traits such as those studied here have low accuracy and heritability with low or moderate magnitudes. This low accuracy may be affected by the size of the training population and its proximity to the validation population.

Table 4. Computational time values in seconds to perform the analysis of the BLASSO, Regression Tree, Bagging, Random Forest and Boosting methods for each trait.

Traits	BLASSO	Regression Tree	Bagging	Random Forest	Boosting
CY	168.1	0.4	56.7	23.6	75.7
MBCC	161.6	0.4	54.5	22.9	72.7
MLC	171.2	0.4	57.1	23.6	77.1
SBT	162.4	0.4	54.2	22.9	73.1
LR	162.5	0.4	55.8	23.4	73.1
LL	163.3	0.4	55.7	22.9	73.5
L	149.5	0.4	56.6	24.6	67.3
ETO	145.8	0.3	56.9	24.7	65.6
BFT	161.8	0.3	56.3	24.9	72.8
PROLOM	145.4	0.4	59.7	25.6	65.4

CY = Carcass yield (%); BFT = Backfat thickness (mm); LR = Midline backfat thickness immediately after the last rib (mm); SBT = higher backfat thickness on the shoulder region (mm); ETO = backfat thickness after the last rib, 6.5 cm from the midline (mm); LL = midline backfat thickness between last and next to last but one lumbar vertebrae (mm); L = midline lower backfat thickness above the last lumbar vertebrae; MBCC = carcass length by the Brazilian carcass classification method; MLC = carcass length by the American carcass classification method; PROLOM = diameter of the longissimus dorsi muscle in the region of the last rib at 6.5 cm from the dorsolumbar line, from a transverse section in the carriage.

Also, in relation to genetic improvement for carcass traits, studies aiming to find associations between molecular markers and traits of interest have been proposed and used by several authors. Guo et al. (2017) identified 13 suggestive loci in nine chromosomes of two commercial breeds that are associated with growth and fatness, which the most significant was associated with backfat thickness. In addition, Blaj et al. (2018) used association studies to detect genes associated with backfat thickness, met to fat ratio and carcass length traits. The study did not identify any more genes than those previously identified in the literature, but contributed to improve the genetic map resolution of three pig populations.

CONCLUSIONS

Overall in this study, the best predictive ability values were presented by the Random Forest and BLASSO methods. The Random Forest method obtained better bias values for most traits. This method also presented computational time approximately 11 times shorter than the time needed with BLASSO. Thus, we consider the Random Forest method as an appropriate alternative for working with genomic datasets.

ACKNOWLEDGMENTS

The authors thank the Brazilian funding organization Coordination for the Improvement of Higher Education Personnel (CAPES).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Azevedo CF, Nascimento M, Silva FF, Resende MDV, et al. (2015a). Comparison of dimensionality reduction methods to predict genomic breeding values for carcass traits in pigs. *Genet. Mol. Res.* 14: 12217-12227. Doi: 10.4238/2015.October.9.10.

Azevedo CF, Resende MDV, Silva FF, Viana JMS, et al. (2015b). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC genetics.* 16: 105-117. Doi: 10.1186/s12863-015-0264-2.

Bertol TM, Campos RD, Coldebella A, Santos Filho JD, et al. (2010). Qualidade da carne e desempenho de genótipos de suínos alimentados com dois níveis de aminoácidos. *Pesq. Agropec. Bras.* 45: 621-629. Available: http://www.scielo.br/pdf/pab/v45n6/a12v45n6 (accessed August 25, 2019).

Blaj I, Tetens J, Preuß S, Bennewitz J, et al. (2018). Genome-wide association studies and meta-analysis uncovers new candidate genes for growth and carcass traits in pigs. *PloS one.* 13: e0205576. Doi: 10.1371/journal.pone.0205576.

Breiman L (2001). Random forests. *Mach learn.* 45: 5-32.

Costa EV, Diniz DB, Veroneze R, Resende MDV, et al. (2015). Estimating additive and dominance variances for complex traits in pigs combining genomic and pedigree information. *Genet. Mol. Res.* 14: 6303-6311. Doi: 10.4238/2015.June.11.4.

De los Campos G, Naya H, Gianola D, Crossa J, et al. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics.* 182: 375. Doi: 10.1534/genetics.109.101501.

Gianola D (2013). Priors in Whole-Genome Regression: The Bayesian Alphabet Returns. *Genetics.* 194: 573. Doi: 10.1534/genetics.113.151753.

González-Camacho JM, de los Campos G, Pérez P, Gianola D, et al. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. TAG. *Theor. Appl. Genet.* 125: 759-771. Doi: 10.1007/s00122-012-1868-9.

González-Camacho JM, Ornella L, Pérez-Rodríguez P, Gianola D, et al. (2018). Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance. *Plant Genome.* 11: 170104. Doi: 10.3835/plantgenome2017.11.0104.

González-Recio O and Forni S (2011). Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genet. Sel. Evol.* 43: 7. Doi: 10.1186/1297-9686-43-7.

González-Recio O, Rosa GJ and Gianola D (2014). Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* 166: 217-231. Doi: 10.1016/j.livsci.2014.05.036.

Guo Y, Qiu H, Xiao S, Wu Z, et al. (2017). A genome-wide association study identifies genomic loci associated with backfat thickness, carcass weight, and body weight in two commercial pig populations. *J. Appl. Genet.* 58: 499-508. Doi: 10.1007/s13353-017-0405-6.

Ho DSW, Schierding W, Wake M, Saffery R, et al. (2019). Machine Learning SNP Based Prediction for Precision Medicine. *Front Genet.* 10: 267. Doi: 10.3389/fgene.2019.00267.

James G, Witten D, Hastie T and Tibshirani R (2013). *An introduction to statistical learning* (Vol. 112, p. 18). Springer, New York.

Meuwissen TH, Hayes BJ and Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* 157: 1819 -1829.

Ogutu JO, Piepho HP and Schulz-Streeck T (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC proc.* 5(3): 11. Doi: 10.1186/1753-6561-5-s3-s11. BMC Proc. 2011 May 27;5 Suppl 3:S11. doi: 10.1186/1753-6561-5-S3-S11.

Peixoto J, Guimarães SE, Lopes PS, Soares MA, et al. (2006). Associations of leptin gene polymorphisms with production traits in pigs. *J. anim. breed genet.* 123: 378-383. Doi: 10.1111/j.1439-0388.2006.00611.x.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raftery AE and Lewis SM (1992). [Practical Markov Chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. *Stat. Sci.* 7: 493-497.

Resende MDV, Resende Junior MFR, Aguiar AM, Abad JIM, et al. (2010). *Computação da seleção genômica ampla (GWS)*. Colombo: Embrapa Florestas, 79p, Série Documentos, 210.

Resende MDV, Silva FF, Lopes PS and Azevedo CF (2012). Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial. Viçosa: Ed. UFV.

Samore AB and Fontanesi L (2016). Genomic selection in pigs: state of the art and perspectives. *Ital. J. Anim. Sci.* 15: 211-232. Doi: 10.1080/1828051X.2016.1172034.

Santos PMD, Nascimento ACC, Nascimento M, Silva FF, et al. (2018). Use of regularized quantile regression to predict the genetic merit of pigs for asymmetric carcass traits. *Pesq. Agropec. Bras.* 53: 1011-1017. Doi: 10.1590/s0100-204x2018000900004.

Smith BJ (2007). boa: an R package for MCMC output convergence assessment and posterior inference. *J. Stat. Softw.* 21: 1-37.

Sousa TV, Caixeta ET, Alkimim ER, Oliveira ACB, et al. (2018). Early Selection Enabled by the Implementation of Genomic Selection in *Coffea arabica* Breeding. *Front plant. Sci.* 9: 1934. Doi: 10.1111/nph.13322.

Teixeira FRF, Nascimento M, Nascimento ACC, Paixão DM, et al. (2015). Determinação de fatores em características de suínos. *Rev. Bras. Biomet.* 33: 130-138.