SELECTION OF SNP MARKERS: ANALYZING GAW17 DATA USING DIFFERENT METHODOLOGIES

Mariana Pavan IÓCA[1]
Daiane Aparecida ZUANETTI[1]

▪ ABSTRACT: The quantity and complexity of generated data due to advances in genetic sequencing technologies has made statistical analysis an essential tool for their correct study and interpretation. However, there is still no agreement about which methodologies are more appropriate for those data, especially for the selection of genetic features that influence a specific phenotype. Genetic data are usually characterized by having a number of variables which is much greater than the number of observations. These variables exhibit little variability and high correlation. These characteristics hinder the application of traditional methodologies for variable selection. In this work *(i.)* we present different methodologies for selecting variables - Random Forest, LASSO and the traditional Stepwise method; *(ii.)* we apply them to genetic data to select SNP markers that characterize the presence or absence of a disease and *(iii.)* we compare their performances. Random Forest and Lasso show similar prediction performance, however none of them correctly select the relevant SNPs.

▪ KEYWORDS: LASSO; Random Forest; SNP markers; variable selection.

1 Introduction

The science of Genetics studies the presence, variation and transmission of features through generations. Some of the first studies date back to 1860, in which, the Austrian monk Gregor Mendel unveiled the principles of heredity by crossing pea plants. However, it was only in the middle of the 1950s that Francis Crick, James Watson and Maurice Wilkins discovered DNA (deoxyribonucleic acid) and how the genetic information is stored. DNA is composed of two strands of polynucleotides (a

[1] Universidade Federal de São Carlos - UFSCar, Departamento de Estatística, CEP: 13565-905, São Carlos, SP, Brasil. E-mail: *mariana.p.ioca@gmail.com; dzuanetti@ufscar.br*

double helix) that contain the nitrogenous bases adenine (A), thymine (T), cytosine (C) and guanine (G). A bonds with T and C bonds with G, via hydrogen bonds.

Advances in sequencing technology have decreased the cost and increased the speed of obtaining data, allowing the first investigations on molecular markers (DNA sequences capable of presenting the polymorphism of the individuals being studied) to emerge. The first strategy for the study of molecular markers is introduced at the beginning of the 1980s, with the characterization of RFLP (Restriction Fragment Length Polymorphism) markers in swine (CHARDON *et al.*, 1985) and cattle (BECKMANN *et al.*, 1986).

Most of the first investigations used limited and laborious methodologies to analyze data. However, technological development promoted the improvement and evolution of methods with greater precision. Currently, RFLP is no longer used and has been replaced by the SNP (Single Nucleotide Polymorphism) technique. SNP markers are composed of the variation of only one nucleotide (A,T,C or G) in a determined gene and they may or may not cause a change in the phenotype (observed physical feature). Sometimes, a SNP is not the cause of a disease, however, its identification helps establish the location (in the genome) of genetic factors that contribute to the phenotype's variability, that is, in the presence of the disease.

Despite technological advances, there are few statistical studies on genetic data, more specifically those that associate SNP markers with the presence of diseases, as this situation presents problems that hinder the use of traditional statistical tools. Generally, data present a greater number of variables than observations, highly correlated variables and rare mutations. The most used methods for variable selection are based on the estimation of a simple linear regression model between the genotype of each SNP and the studied phenotype. The most significant SNPs are chosen by hypothesis tests. This approach presents benefits, such as: low computational processing time, ease of use and interpretation of results. However, Zeng *et al.* (2015), Oliveira (2015) and Feng *et al.* (2012) warn of the deficiencies of this approach that does not consider the association structure between SNPs, does not allow the interaction between the effects of two or more SNPs on the phenotype and usually has low power.

With the goal to overcome the introduced deficiencies, some methodologies have been proposed, some of them classified as machine learning methodologies. Here, we highlight: Random Forest (MOKRY *et al.*, 2013; OLIVEIRA, 2015; BREIMAN, 2001), Principal Component Analysis (LEWIS *et al.*, 2011), Allelic Frequency Analysis (SUEKAWA *et al.*, 2010; SASAZAKI *et al.*, 2011), Genetic Algorithms (GOLDBERG, 1989; OLIVEIRA, 2015), Sparse Partial Least Squares (CHUN and KELES, 2010) and LASSO (PARK and CASELLA, 2008; OLIVEIRA, 2015).

Comparing the predictive accuracy of the LASSO and the combination of the likelihood estimation method with the traditional Stepwise variable selection method, in a linear regression model, Kumar *et al.* (2019) and Hastie and Tibshirani (2017) show that LASSO outperforms the latter methodology in several different scenarios. Considering logistic models, that are the focus of this study, Alcântara Junior (2020) show similar results. Analyzing genetic data, in particular, Ogutu *et al.* (2011) evaluate the predictive accuracy of Random Forest (RF), Boosting and Support Vector Machines (SVMs) for predicting genomic breeding values using dense SNP markers and show better performance for Boosting than for SVMs and RF.

Based on these methodologies, the objectives of this work are to study and apply Random Forest and LASSO in genetic data to select SNP markers that characterize the presence or absence of a disease and to compare their performance with the Stepwise. Proposals and suggestions for adaptation and better use of these methodologies in the analysis of genetic data are also carried out. The data set used is the GAW17 (ALMASY *et al.*, 2011), which is a well-known data set for evaluating the performance of methodologies in the selection of SNPs.

The manuscript is organized as follows. Sections 2 and 3 present the LASSO and Random Forest methodologies, respectively. Section 4 discusses the GAW17 data. In Section 5, we describe how the analysis is carried out and how the results are obtained. Finally, Section 6 shows final remarks and a discussion.

2 LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator), proposed by Tibshirani (1996), is a variable selection methodology whose goal is to find a more parsimonious Least Squares estimator for a regression model. Consider the linear model,

$$y_k = \sum_{i=0}^{d} (\beta_i x_{ki}) + \epsilon_k, \tag{1}$$

for k =1,2,. . . ,N, where y_k is the observed value of the response variable (phenotype) of the k-th individual, β_is are the unknown parameters (regression coefficients), x_{ki} is the observed value of the i-th covariate (genotype of the i-th SNP) for the k-th individual and ϵ_k is the random error for the k-th individual.

According to Izbicki and Santos (2018), the main idea of LASSO is to add the restriction $\sum_{i=1}^{d} |\beta_i| \leq c$, where $c = c(\lambda)$, to the traditional Least Squares formula used to estimate the value of β_is. With this addition, the estimate of some β_is are approximately zero and, therefore, the method selects those covariates that present $\beta_i \neq 0$ as relevant variables.

Such methodology is then used to define the complexity of the model, since a greater number of covariates in the model means a more complex model. Usually, lower values of c represent models with a lower degree of complexity, as they generally lead to the selection of smaller amounts of covariates. Thus, the LASSO methodology looks for

$$\hat{\boldsymbol{\beta}} = arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \sum_{k=1}^{n} \left(y_k - \beta_0 - \sum_{i=1}^{d} \beta_i x_{ki} \right)^2 + \lambda \sum_{i=1}^{d} |\beta_i| \,, \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter. It is worth noting that the penalty used in LASSO is quickly implemented and often has good computational performance.

LASSO can be extended to a wide variety of objective functions in addition to linear regression, such as Generalized Linear Models, which will be better described in Section 2.3. In the application presented in this work, we use the LASSO version that predicts a probability of success (such as a Logistic Regression) and not a value on the real line.

2.1 Choosing λ by cross-validation

Note that for each value of λ in Eq. (2), a different set of β_is is obtained and, if $\lambda = 0$, the LASSO becomes identical to the Least Squares estimator, that is, with all non-zero regression coefficients. One of the main advantage of the methodology is that we can choose the value of λ which provides a final model with good predictions and which also selects the most important covariates.

In order to find the best model that LASSO can estimate for the data being analyzed, it is essential to choose a good value for λ. In general, according to Izbicki and Santos (2018), the choice of λ is made from a cross-validation of different fitted models. It is worth mentioning that LASSO does not perform cross-validation and it only fits a model for the fixed λ.

Cross-validation is a technique that evaluates the prediction performance of an estimated model in a set of independent data. The most frequently used method of cross-validation is the k-fold. Hastie *et al.* (2008) defines the k-fold algorithm as follows:

- the database used to estimate the model is randomly divided into k mutually exclusive subsets of approximately the same size;
- then, all possible combinations (called Training bases) of $k-1$ of these k groups are made and, for each combination, we estimate models for different values of λ;
- the Prediction Error (EP) of each estimated model is calculated on the subset that is not used in the model estimation (called the Validation base). In this case, the EP is defined as the sum of the squared residuals of the base that is separated for validation; and
- we calculate the average EP for each λ. The value with the lowest average EP is chosen as the λ that provides the best model among those estimated.

2.2 Implementation

During the studies, some problems were observed. As it is known, the LASSO estimate for β_is is biased and can change significantly depending on the Training and Validation bases used in their estimation. In some situations, β_is that have an estimate that is close to zero in one sample, may have a high absolute value in another.

An alternative to solve this problem in the application included in this study is to use Generalized Linear Models (GLM) to calculate the β_is estimate. In this case, LASSO is used for variable selection and the final model and its estimates are obtained from the fitting of a GLM only considering the variables selected by LASSO.

LASSO is already implemented in R packages. One of which is *glmnet* (FRIEDMAN *et al.*, 2010). Built by Jerome Friedman, Trevor Hastie and Rob Tibshirani, this package not only performs estimation of β_is but also chooses the best value of λ from cross-validation.

2.3 Generalized Linear Models

Since we use Generalized Linear Models for the final estimation of the regression coefficients for the LASSO methodology and also along with the Stepwise method to compare the performance of different procedures, we briefly describe these models. GLM were proposed by Nelder and Wedderburn (1972). Until their proposal, attempts were made to fit normal linear models for almost any type of random phenomenon, but in many cases it was necessary to transform the variables involved in these phenomena for these models to be considered adequate.

With the limitations of the normal model, GLM were created based on the idea of opening a range of options for the distribution of the response variable (PAULA, 2013), because these models allow the distribution of $y|x$ to belong to the exponential family and not only to the normal family. Poisson, Inverse Normal, Normal, Gamma and Binomial are examples of distributions that belong to the exponential family. The estimation of GLM using maximum likelihood methodology ensures the uniqueness of the $\boldsymbol{\beta}$'s estimator.

Because the data used in this investigation present a binary response variable, we use a Binomial distribution with $n = 1$ to fit the GLM. Our interest is to estimate the probability of occurrence of a specific disease for each individual. The canonical link function of the Binomial distribution is the logit function, given by

$$g(p_k) = \log\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \sum_{i=1}^{d} \beta_i x_{ki},$$

where p_k is the $k-$th individual's probability of success. In this case, we have a Logistic Regression.

As already mentioned in Section 2, we can use LASSO to select relevant variables and GLM to estimate the regression coefficients of the final model. We follow this method for the application of the LASSO in this study.

3 Random Forest

Decision Tree, Boosting and Bagging are essential methodologies for the Random Forest development. Therefore, we briefly discuss these three techniques to better understand the methodology we use here.

3.1 Decision Tree

Decision Tree or Regression Tree is a supervised machine learning methodology. In summary, a tree is built from the significance of each covariate in relation to the response variable and a node is created from each significant covariate. Thus, we look for the variable that has the greatest effect on the response variable and from it, a node is created, usually composed of two branches or two classes of the predictor variable. Then, we try to identify which of these two branches should be partitioned from a new covariate.

According to Izbicki and Santos (2018), the Decision Tree algorithm includes two steps:

- ***I.*** *Creating a Decision Tree:* it seeks to find the partitions in which the response variable appears as homogeneously as possible on the leaves; and

- ***II.*** *Tree Pruning:* each node is removed, one at a time, and the effect caused by that removal is observed in the Prediction Error for the Validation data. From these values, we decide which nodes will remain in the Tree. This is an essential step to avoid model overfitting to the Training data and to reduce the complexity of the Tree.

3.2 Bagging and Boosting

Bagging (BREIMAN, 1996) is a technique whose objective is to reduce the variance of a forecasting model, in our case a Decision Tree. This methodology consists in fitting a Decision Tree for several independent samples. In the end, each Tree has the weight of a vote and the decision is made by the simple majority of them, that is, $50\% + 1$ of the total votes. For data with great variability and low bias, Bagging is a methodology that usually presents good results, that is, good individual forecasts.

Unlike Bagging, the final vote of Boosting (SCHAPIRE *et al.*, 1998) is defined in a weighted way since, while they are estimated, the Trees consider greater weight for the observations that were previously erroneously predicted. Another difference between the two methodologies is that while Bagging is concerned with keeping the prediction variance small, Boosting focuses on reducing bias with better predictions. In general, Boosting outperforms Bagging and, in most cases, Boosting is preferable (HASTIE *et al.*, 2008).

3.3 Random Forest

Random Forest methodology (BREIMAN, 2001) is a combination of different and uncorrelated Decision Trees for decision making. Its performance is similar to Boosting (HASTIE *et al.*, 2008).

Random Forest's main idea is to reduce the correlation between the Trees without increasing the variance of the prediction. In other words, the methodology seeks to find a balance between Boosting and Bagging. We can define the algorithm for the growth of a Forest as follows:

1. resample η bootstrap samples - Training bases;
2. for each of the η samples, grow a tree following the next steps:
   i. randomly select m covariates from the set of d available covariates. Breiman and his contributor Adele Cutler recommend that the number of variables to be used in each Tree is $m = \frac{d}{3}$ with at least 5 nodes. In practice, the best value for these parameters can be defined from the data;
   ii. choose the most significant variable among the m drawn variables;
   iii. divide the node into two "child nodes" so that the leaves are as homogeneous as possible in the response variable;
   iv. repeat steps ii. and iii. until the entire tree grows. Then, proceed to prune the tree; and
3. take the prediction of new observations.

In the case of Random Forest regression, the final prediction for an observation is the average of its Tree predictions, defined as

$$\hat{f}^{\eta} = \frac{1}{\eta} \sum_{\ell=1}^{\eta} T_{\ell}(\mathbf{x}) \tag{3}$$

where $T_{\ell}(\mathbf{x})$ is the predicted value for $\mathbf{x}$ in the $\ell-$th fitted Tree.

3.3.1 Out-of-bag Samples and Cross-validation

One of the characteristics of Random Forest is the use of Out-of-Bag (OOB) samples, a cross-validation methodology, to grow and prune Trees and fit the best Forest. Hastie *et al.* (2008) defines an OOB sample, for a specific bootstrap Training sample, as being the base composed of all the observations that are not part of it. In the estimation via Random Forest, the training phase ends when the OOB Prediction Error is stabilized.

The final predicted value for each observation is calculated using only the Trees of which the observation is not part of the bootstrap Training base. We calculate the total Prediction Error in a similar way to the k-fold method explained in Section 2.1.

3.3.2 Variable Importance and Implementation

The OOB samples are also used to calculate the importance of the variables, which is given by the change in the Prediction Error when the $i-$th covariate is excluded from the Trees, while keeping the other variables in Trees. Therefore, to measure the importance of each variable we have the following procedure:

1. the OOB Prediction Error of the η Trees is calculated;
2. for each one of the m covariates of each Tree, remove the nodes relative to it and calculate the new Error;

3. the Partial Importance (IP) of the i-th variable in the $\ell-$th Tree ($IP_{i\ell}$) is calculated as the relative difference (in percentage) between the Error of the $\ell-$th Tree without the $i-$th variable and the Error of the $\ell-$th full Tree, given by Equation (4) as:

$$IP_{i\ell} = \left(\frac{EP_{i\ell} - EP_{c\ell}}{EP_{c\ell}}\right) * 100; \tag{4}$$

4. the importance of the $i-$th variable is the average of its η Partial Importances.

Random Forest is implemented in an R package called randomForest (LIAW and WIENER, 2002) with the methodology created by Breiman. The default values of the parameters set in the package, such as number of covariates m and depth of the trees, among others, are the authors' recommendations. From it, we can fit a Random Forest for classification or regression, with the first case being the interest of this study.

4 GAW17 data

The database analyzed in this study is called Genetic Analysis Workshop 17 (GAW17) and was built from simulated and real data for 697 unrelated individuals, 327 men and 370 women. The simulation of a complex disease and risk factors was made based on the real data contained in the 1000 Genomes Project. In this simulation, 24,487 SNPs divided in 22 chromosomes were obtained.

The simulation was made for a common and complex disease that has a prevalence of 30% in the population. Along with the disease, three other continuous quantitative phenotypes were simulated: Q1, Q2 and Q4 that are not explored in this study, in addition to smoking status. The simulated SNP markers are autosomal, that is, there are no markers present on the sex chromosome. For more details on the simulation of these data see Almasy *et al.* (2011). Although the disease risk is also a function of Q1, Q2 and Q4 and, consequently, is influenced by their influential SNPs, this study focuses mainly on identifying the SNPs that are relevant to the disease liability.

Originally, the information in the database for each SNP is based on nitrogen bases A, T, C or G, with 16 possible pairs: A/A, T/T, C/C, G/G, A/C, A/G, A/T, C/A, C/G, C/T, G/A, G/C, G/T, T/A, T/C, T/G. For this study, we classify the observations as follows:

- A/A or T/T = 1: dominant homozygote;
- C/C or G/G = -1: recessive homozygote; and
- A/C, A/G, A/T, C/A, C/G, C/T, G/A, G/C, G/T, T/A, T/C or T/G = 0: heterozygous.

Almasy *et al.* (2011) highlights 51 SNPs that are used in determining the presence or absence of the disease in the individuals being analyzed. They are distributed as follows: 30 SNPs on chromosome 1, 3 on chromosome 2, 5 on chromosome 8, 6 on chromosome 14, 1 on chromosome 16 and 2 on chromosomes 17, 18 and 19. Table 1 shows the number of individuals in each category (-1, 0 or 1) for each of these 51 relevant SNPs. Almasy *et al.* (2011) also presents the value for the regression coefficients used for the phenotype simulation.

Table 1 - Descriptive analysis of significant SNPs in the entire database, where the number after the letter C identifies the chromosome from which it comes

chromosome 1				**chromosome 2**			
SNP	*-1*	*0*	*1*	*SNP*	*-1*	*0*	*1*
C1S9391	0	1	696	C2S2286	696	1	0
C1S9423	696	1	0	C2S2288	693	4	0
C1S9432	683	13	1	C2S2307	0	1	696
C1S9445	696	1	0	**chromosome 8**			
C1S9446	696	1	0	C8S4825	696	1	0
C1S9449	696	1	0	C8S4839	696	1	0
C1S9455	693	4	0	C8S886	696	1	0
C1S9457	696	1	0	C8S900	695	2	0
C1S7061	689	7	1	C8S909	695	2	0
C1S11396	696	1	0	**chromosome 14**			
C1S3181	696	1	0	C14S1381	696	1	0
C1S3182	696	1	0	C14S1382	0	5	692
C1S5748	0	1	696	C14S3630	0	1	696
C1S9164	695	2	0	C14S3695	696	1	0
C1S9165	0	1	696	C14S3704	0	5	692
C1S9172	691	6	0	C14S3706	0	246	451
C1S9173	0	2	695	**chromosome 16**			
C1S9174	696	1	0	C16S1894	0	1	696
C1S9189	688	9	0	**chromosome 17**			
C1S9200	696	1	0	C17S4578	39	154	504
C1S9222	0	1	696	C17S4581	0	1	696
C1S9250	695	2	0	**chromosome 18**			
C1S9266	693	4	0	C18S2475	696	1	0
C1S9267	694	3	0	C18S2492	0	24	673
C1S9306	696	1	0	**chromosome 19**			
C1S9320	696	1	0	C19S4929	695	2	0
C1S9333	696	1	0	C19S4937	695	2	0
C1S9346	696	1	0	**Disease**			
C1S9373	696	1	0	*0*		*1*	
C1S2919	696	1	0	488		209	

We observe in Table 1 that approximately 30% of the individuals exhibit the disease being studied, which shows that we are not dealing with a rare disease. We also note that 30 of the relevant SNPs present only one observation in a different category. These results show that the SNPs that determine the disease are uncommon, which makes our study a case of very low variability in covariates, which consequently makes these difficult to select as importante variables.

In our study, the presence of each one of the 51 important SNPs represents an increase in the probability of the individual presenting the disease. Thus, a person who has a variation in the 51 SNPs is the one most likely to have the disease. As SNPs are rare, we do not have a large number of individuals that present a considerable proportion of these 51 SNPs, therefore we do not have a large difference between the disease's probability for sick and non-sick individuals, which increases the chance of error.

Another factor of complexity of these data is that when we split the full base into Training (containing 70% of the observations) and Testing (with the remaining 30% of the observations), the importance of SNPs that have only 1 different observation is too complex to be evaluated and modelled since this only observation is either in Training or Testing base.

The full data set and the R codes used for the analysis are available on GitHub at the link `https://github.com/Mariana3112/TCC`.

5 Analysis and Results

According to Mendel's Second Law, the information contained in a chromosome is independent of the others. Thus, for selecting SNPs that influence the presence of the disease, the methodologies are applied separately to each of the 22 autosomal chromosomes.

The database is randomly divided in two parts: the first with 70% of the base (489 observations) is called Training and the second with remaining 30% (208 observations) is called Testing. For the models' estimation, including determining the LASSO regularization parameter (λ), we use the Training base. The obtained models are applied to the Testing base only to compare their prediction performance. This method allows to compare the performance of the models on a different base than the one used in their fit, showing us how they behave in completely independent data. Then, in addition to analyze the SNPs selected by each methodology we evaluate the predictive accuracy of each model in the independent data.

In Sections 2 and 3 we notice that both methodologies use a method of cross-validation for the complete models' estimation and, for that, they split the Training base into two bases: Effective Training and Validation. The way the base are divided can influence the variable selection and estimation of the final models. Taking this into account, we fit 21 different models for each one (with different seeds for the division of the Training base), instead of performing a single fitting for each methodology. Then, we record the covariates selected in each run for LASSO and Random Forest for each chromosome. For LASSO, we select all SNPs that have a non-zero regression coefficient and, for Random Forest, the 30 most relevant

SNPs.

5.1 Stability in the Variable Selection

We notice a variation, which is more evident in LASSO, in the variable selection in both methodologies among the 21 fits. For this reason, in order to finally select the important SNPs in each methodology, we analyze the frequency with which each SNP appeared in the 21 fits. The results are shown in Table 2.

Table 2 - Number of SNPs that appear in each scenario

Number of fits							
21		**≥17**		**≥14**		**≥12**	
LASSO	Forest	LASSO	Forest	LASSO	Forest	LASSO	Forest
0	123	0	257	1	337	3	401
	≥11		**≥7**		**≥1**		
	LASSO	Forest	LASSO	Forest	LASSO	Forest	
	5	444	66	654	961	2,935	

When analyzing Table 2, if we define important SNPs as those that are selected in at least 17 runs (80% out of total runs), LASSO does not select any covariate. If we consider important SNPs as those that are selected in at least one fit, both methodologies select many SNPs. These results highlight that for the data used in this work, LASSO presents great instability in the variable selection and Random Forest is not constant, but is more stable than LASSO.

Considering the values presented in Table 2, we define important SNPs as those that are selected in at least 11 runs of considered methods.

5.2 LASSO

In Section 5.1 we observe that 5 covariates are selected in at least 11 of LASSO's fits. Table 3 shows the selected SNPs and in which categories the observations are allocated for the whole and Training bases.

Table 3 - Descriptive analysis of the SNPs selected by LASSO, where the number after the letter C in the name of the SNP identifies which chromosome it comes from

	Full base			**Training**		
SNPs	-1	0	1	-1	0	1
C3S5389	685	12	0	482	7	0
C3S5742	683	12	2	479	9	1
C3S4611	693	4	0	486	3	0
C15S774	0	4	693	0	4	485
C18S2320	693	4	0	485	4	0

First of all, we note that only C18S2320, in chromosome 18, is on any of the chromosomes that contain relevant SNPs for the latent liability and is close to two relevant SNPs in gene PIK3C3. However, it is not an important SNP and LASSO is not successful in selecting any of the 51 relevant covariates.

We also highlight that the heterozygous observations of SNPs C15S774 and C18S2320 are all in the Training base and consequently, none are in the Testing base. Therefore, the real importance of these covariates is not analyzed and confirmed in the independent database.

Despite not selecting any of the expected covariates, it is possible to note that LASSO identifies SNPs with marginal distributions similar to those used in the simulation. Looking at Tables 1 and 3, we observe that C3S5389 has similar characteristics to C1S9189. However, when we analyze their contingency table, we realize that the SNPs' mutations do not occur in the same individuals. Therefore, one should not replace the other. The same happens with the SNP C3S5742, which could be seen as a substitute for the C1S9432, but as the mutations do not occur in same individuals, they are not substitutes. For SNPs identified on chromosome 3, we also highlight that they are close to 13 SNPs of BCHE gene that are relevant for the continuous phenotype Q2 which impacts the disease risk. The marginal distribution of SNP C15S774 resembles that of C14S1382 and C14S3704, but it is not the same individuals that suffer these mutations.

Not unlike other selected variables, SNP C18S2320 has a marginal distribution similar to an important SNP on another chromosome. SNP C18S2320 along with C3S4611 could have been selected in place of C1S9455, C1S9266 or C2S2288 but the mutations in these SNPs happen in different individuals.

In Section 2, we discuss the fact that the estimate for β_is can be biased when calculated by LASSO. Therefore, we use it for selecting the variables and estimate the final model by a Logistic Regression. Thus, we obtained the model defined in Equation (5) as

$$\begin{gathered}\log\left(\frac{\hat{p_k}}{1-\hat{p_k}}\right) = 41.71 + (2.91 * \mathrm{C3S5389}_k) + (1.87 * \mathrm{C3S5742}_k) \\ +(12.68 * \mathrm{C3S4611}_k) - (12.68 * \mathrm{C15S774}_k) + (12.68 * \mathrm{C18S2320}_k).\end{gathered} \tag{5}$$

We finally apply the fitted model on the Testing base and the Prediction Error is calculated as $EP_{LASSO} = 51.64$. This performance indicator is important later on, when we compare the models resulting from each tested methodology.

5.3 Random Forest

In Section 5.1, we observe that 444 SNPs are selected in at least 11 of the 21 estimated Random Forests. Due to the large number of variables, we do not show a detailed analysis of their characteristics as we did for LASSO.

Despite being more stable than the LASSO in the variable selection, we notice that Random Forest also selects different variables depending on the OOB samples. Therefore, we run the methodology 5 more times, with different seeds, only considering the 444 most important SNPs. The final model is chosen by the lowest Prediction Error and presents $EP_{RF} = 50.08$ in the Testing base.

Figure 1 shows the variables that are most important in the final model. We note that almost all the chromosomes presented in Table 1 are represented, except chromosomes 8 and 19. However, no real relevant SNP is selected. In addition to SNPs C18S2320 and C3S5742 that are also identified by LASSO, the Random Forest selects C17S4432 and C17S4431 as two of the most relevant SNPs. Although they are not true important SNPs, they are close to two true relevant SNPs, C17S4578 and C17S4581, in PRKCA gene.

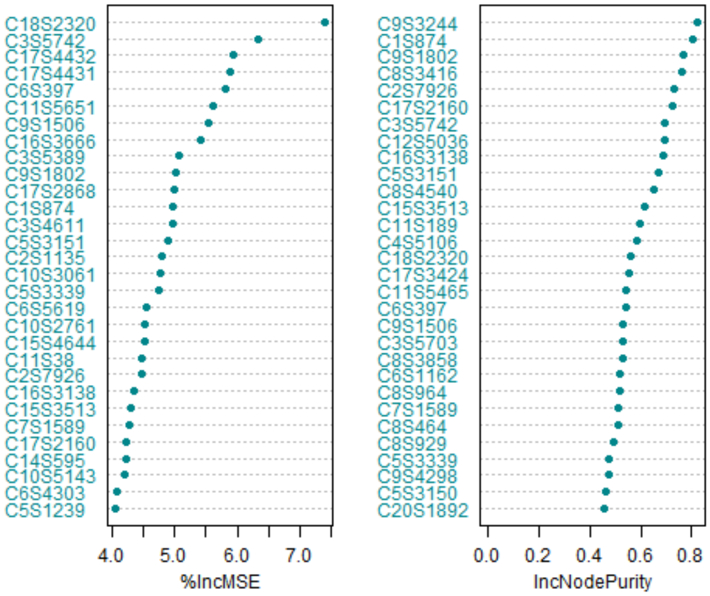

Figure 1 - Importance of variables via Random Forest, where the number after the letter C in the name of the SNP represents which chromosome it comes from.

5.4 Random Forest and Logistic Regression

Given that the Random Forest method still has a large number of selected SNPs, much larger than the actual number of significant SNPs, we also estimate a third model combining the Random Forest and Logistic Regression methods. The idea of this third method is to use the Random Forest for a pre-selection of variables and then estimate the final model through Logistic regression. This final model is estimated in two ways:

1. using the 444 variables selected by the Random Forest; and

2. applying a second variable selection criterion, the Stepwise method.

A Logistic regression model is fitted with the 444 previously selected variables. As it is a very extensive model, we only show the EP for comparison with other methods, $EP_{RFLG} = 79.38$.

In search of a more parsimonious model, we also select variables via Stepwise along with Logistic Regression. Stepwise is a methodology in which variables are added and removed from the model in each step based on a pre-established criterion. The criterion used here is the AIC (Akaike Information Criterion). Lower AIC values generally indicate better fitted models and the variables are included or excluded from the model if the value of the AIC decreases with the action that is tested.

After performing 1,000 iterations, we find a model with 75 covariates. We only include the performance measure because it is an extensive model. The resulting EP in the Testing base is given by $EP_{RFLGS} = 76.92$.

5.5 Performance's Comparison

We observe that cases where we combine the Random Forest and GLM methodologies display the worst prediction results, as they are the models with the greatest errors. Applying Stpewise to select variables in large dimensional data is also almost impracticable computationally. The models that combine the selection via LASSO with fitting via GLM have similar prediction errors to the models in which fitting and selection is done via Random Forest. However, the model using only Random Forest shows the lowest Prediction Error.

Regarding complexity, the model fitted by LASSO is much simpler, being composed of only 5 covariates, while the Random Forest selects 444. From the results observed in Section 5.1 and considering the data set being studied, the Random Forest methodology is more stable in selecting variables compared to the selection made by LASSO. None of the studied methodologies show good performance in selecting the 51 true relevant SNPs. This is expected considering the complexity of the analyzed data. However, they identify other SNPs with similar marginal distribution or close, sometimes in the same gene, to true significant SNPs.

Considering other studies that also analyze the GAW17 unrelated data, Wang *et al.* (2011) propose a supervised coalescence of SNPs in a specific region (a gene, for instance) that collapses multiple common and rare SNPs into a gene-level marker and treats them as a single predictor in the model. The variable selection is carried out using an empirical Bayes method which assumes a mixture prior distribution for the regression coefficients. They select SNP C8S890, that is not a true relevant SNP but is close to three important SNPs on PTK2B gene. They also identify two other SNPs in important genes for Q1 and Q2 and several other false-positive SNPs.

Other collapsing methods of SNP information are discussed by Agne *et al.* (2011), Saad *et al.* (2011), among others. The first authors calculate the significance of each collapsed region by permutation test and select two or three regions among several false-positive areas that actually contain true relevant SNPs.

6 Discussion

In this manuscript, we verify the efficiency of two machine learning methodologies, LASSO and Random Forest, to select SNPs markers that impact the probability of the presence or absence of a disease in unrelated individuals. We also verified the performance of the Stepwise method combined with Logistic Regression after a pre-selection of variables via Random Forest. This type of genetic data usually show low variability (rare events), large correlation and contain greater number of variables than observations in the sample. These features make it difficult to use traditional methods of variable selection and new methodologies need to be proposed and tested.

When exploring the GAW17 data, we compare the stability of LASSO and Random Forest in selecting variables when the Effective Training base is modified and observe whether or not the markers actually used in the data simulation are selected by the methodologies. Proposals and suggestions for adaptation and better use of the tested methodologies in the analysis of genetic data are also carried out.

With the considered data, the most appropriate methodology to reduce the Prediction Error is the Random Forest. However, we emphasize that the model using LASSO for selecting variables and Logistic Regression for model fitting is more parsimonious and presents a similar Prediciton Error.

None of the studied methodologies are able to correctly select the relevant SNPs and this may be due to the fact that most of them present very low variability in the considered sample. This is a very common situation in genetic data, as mutations occur in a very small number of people. Despite the predictive capacity of the fitted models, the associations found between some SNPs and the presence or absence of the disease appear to be spurious, at first, since they are not true relevant SNPs in the response variable. However, the identified SNPs usually have similar marginal distribution or are closely located, sometimes in the same gene, to true important SNPs. These results are not very different from the results of other authors cited in this manuscript.

For future studies, methodologies that consider selection of rare variables should be explored or proposed to identify significant SNPs markers. Zeng *et al.* (2015) use Principal Component Analysis and Mixed Models, in addition to the suggestion of two-stage modeling with different approaches to select them.

IÓCA, M. P.; ZUANETTI, D. A. Seleção de marcadores SNPs: analisando os dados GAW17 com diferentes metodologias. *Rev. Bras. Biom.,* Lavras, v.39, n.1, p.71-88, 2021.

- *RESUMO: A quantidade e a complexidade dos dados gerados devido ao avanço nas tecnologias de sequenciamento genético fez da análise estatística uma ferramenta essencial para o estudo e interpretação correta deles. No entanto, ainda não há um consenso sobre quais metodologias são mais adequadas para esses dados, especialmente para a seleção de características genéticas que influenciam um específico fenótipo. Os dados genéticos geralmente apresentam características, tais como: número de variáveis muito maior que o número de observações, variáveis com pouca variabilidade e muito correlacionadas entre si, que dificultam a aplicação de metodologias tradicionais de seleção de variáveis. Nesse trabalho (i.) apresentamos diferentes metodologias de seleção de variáveis - Florestas Aleatórias, LASSO e o método tradicional Stepwise; (ii.) aplicamo-as em dados genéticos para selecionar marcadores SNP (do inglês Single Nucleotide Polymorphism) que caracterizam a presença ou não de uma doença e (iii.) comparamos suas performances. As Florestas Aleatórias e o LASSO apresentam performance de predição parecidas, mas nenhuma delas seleciona corretamente os SNPs importantes.*

- *PALAVRAS-CHAVE: LASSO; Florestas Aleatórias; marcadores SNP; seleção de variáveis.*

References

AGNE, M.; HUANG, C.; HU, I.; WANG, H.; ZHENG, T.; LO, S. Identifying influential regions in extremely rare variants using a fixed-bin approach. In: BMC. 2011 *Proceedings*, v.5, n.S9, p.S3, 2011.

ALCÂNTARA JUNIOR, G. P. *Avaliação de métodos de estimação e seleção de variáveis em modelos de regressão logística*, 2020. 116p. (Master) - Universidade Federal de São Carlos, São Carlos, 2020.

ALMASY, L.; DYER, T. D.; PERALTA, J. M.; KENT, J. W.; CHARLESWORTH, J. C.; CURRAN, J. E.; BLANGERO, J. Genetic analysis workshop 17 mini-exome simulation. In: BMC. 2011 *Proceedings*, v.5, n.S9, p.S2, 2011.

BECKMANN, J.; KASHI, Y.; HALLERMAN, E.; NAVE, A.; SOLLER, M. Restriction fragment length polymorphism among Israeli Holstein-Friesian dairy bulls. *Animal Genetics*, v.17, n.1, p.25-38, 1986.

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

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

CHARDON, P.; KIRSZENBAUM, M.; CULLEN, P. R.; GEFFROTIN, C.; AUFFRAY, C.; STROMINGER, J. L.; COHEN, D.; DAIMAN, M. Analysis of the sheep MHC using HLA class I, II, and C4 cDNA probes. *Immunogenetics*, v.22, n.4, p.349-358, 1985.

CHUN, H.; KELES, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v.72, n.1, p.3-25, 2010.

FENG, Z. Z.; YANG, X.; SUBEDI, S.; MCNICHOLAS, P. D. The LASSO and sparse least squares regression methods for SNP selection in predicting quantitative traits. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, v.9, n.2, p.629-636, 2012.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, v.33, n.1, p.1-22, 2010.

GOLDBERG, D. E. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Publishing Company, 1989. 412p.

HASTIE, T.; TIBSHIRANI, R. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv: Methodology*, 2017.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. 2ed. Springer, 2008. 745p.

IZBICKI, R.; SANTOS, T. *Machine learning sob a ótica estatística: Uma abordagem preditivista para estatística com exemplos em R*. Notes, 2018. 225p.

KUMAR, S.; ATTRI, S.; SINGH, K. Comparison of lasso and stepwise regression technique for wheat yield prediction. *Journal of Agrometeorology*, v.21, n.2, p.188-192, 2019.

LEWIS, J.; ABAS, Z.; DADOUSIS, C.; LYKIDIS, D.; PASCHOU, P.; DRINEAS, P. Tracing cattle breeds with principal components analysis ancestry informative SNPs. *PloS One*, v.6, n.4, e18007, 2011.

LIAW, A.; WIENER, M. Classification and regression by Random Forest. *R News*, v.2, n.3, p.18-22, 2002.

MOKRY, F. B.; HIGA, R. H.; MUDADU, M. de A.; LIMA, A. O. de; MEIRELLES, S. L. C.; SILVA, M. V. G. B. da; CARDOSO, F. F.; OLIVEIRA, M. M. de; URBINATI, I.; NICIURA, S. C. M.; TULLIO, R. R.; ALENCAR, M. M. de; REGITANO, L. C. de A. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. *BMC genetics*, v.14, n.1, p.47, 2013.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, v.135, n.3, p.370-384, 1972.

OGUTU, J. O.; PIEPHO, H. ; SCHULZ-STREECK, T. A comparison of random forests, boosting and support vector machines for genomic selection. In: BMC. 2011 *Proceedings*, v.5, n.S3, p.S11, 2011.

OLIVEIRA, F. C. D. *Um método para seleção de atributos em dados genômicos*, 2015. 273p. Thesis (Ph.D.) - Universidade Federal de Juiz de Fora, Juiz de Fora, 2015.

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

PAULA, G. A. *Modelos de regressão: com apoio computacional.* IME-USP, 2013. 434p.

SAAD, M.; SAINT PIERRE, A.; BOHOSSIAN, N. and MACÉ, M.; MARTINEZ, M. Comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. In: BMC. 2011 *Proceedings*, v.5, n.S9, p.S33, 2011.

SASAZAKI, S.; HOSOKAWA, D.; ISHIHARA, R.; AIHARA, H.; AYAMA, K.; MANNEN, H. Development of discrimination markers between japanese domestic and imported beef. *Animal Science Journal*, v.82, n.1, p.67-72, 2011.

SCHAPIRE, R. E.; FREUND, Y.; BARTLETT, O.; LEE, W. S. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, v.26, n.5, p.1651-1686, 1998.

SUEKAWA, Y.; AIHARA, H.; ARAKI, M.; HOSOKAWA, D.; MANNEN, H.; SASAZAKI, S. Development of breed identification markers based on a bovine 50K SNP array. *Meat Science*, v.85, n.2, p.285-288, 2010.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v.58, n.1, p.267-288, 1996.

WANG, L.; PUNGPAPONG, V.; LIN, Y.; ZHANG, M.; ZHANG, D. Genome-wide case-control study in GAW17 using coalesced rare variants. In: BMC. 2011 *Proceedings*, v.5, n.S9, p.S110, 2011.

ZENG, P.; ZHAO, Y.; QIAN, C.; ZHANG, L.; ZHANG, R.; GOU, J.; LIU, J.; LIU, L.; CHEN, F. Statistical analysis for genome-wide association study. *Journal of Biomedical Research*, v.29, p.97-285, 2015.

Received on 28.08.2020.

Approved after revised on 28.01.2021.