Ciência Rural, Santa Maria, v.52:5, e20201072, 2022
http://doi.org/10.1590/0103-8478cr20201072
ISSNe 1678-4596
CROP PRODUCTION

Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data

Roberta de Amorim Ferreira[1,2] **Gabriely Teixeira**[1] **Luiz Alexandre Peternelli**[1*]

[1]Departamento de Estatística, Universidade Federal de Viçosa (UFV), 36570-900, Viçosa, MG, Brasil. E-mail: peternelli@ufv.br. *Corresponding author.
[2]Instituto Federal de Minas Gerais (IFMG), Governador Valadares, MG, Brasil.

ABSTRACT: *Splitting the whole dataset into training and testing subsets is a crucial part of optimizing models. This study evaluated the influence of the choice of the training subset in the construction of predictive models, as well as on their validation. For this purpose we assessed the Kennard-Stone (KS) and the Random Sampling (RS) methods in near-infrared spectroscopy data (NIR) and marker data SNPs (Single Nucleotide Polymorphisms). It is worth noting that in SNPs data, there is no knowledge of reports in the literature regarding the use of the KS method. For the construction and validation of the models, the partial least squares (PLS) estimation method and the Bayesian Lasso (BLASSO) proved to be more efficient for NIR data and for marker data SNPs, respectively. The evaluation of the predictive capacity of the models obtained after the data partition occurred through the correlation between the predicted and the observed values, and the corresponding square root of the mean squared error of prediction. For both datasets, results indicated that the results from KS and RS methods differ statistically from each other by the F test (P-value < 0.01). The KS method showed to be more efficient than RS in practically all repetitions. Also, KS method has the advantage of being easy and fast to be applied and also to select the same samples, which provides excellent benefits in the following analyses.*

Método Kennard-Stone supera a Amostragem Aleatória na seleção de amostras de calibração em dados SNPs e NIR

RESUMO: *A divisão de subconjuntos de treinamento e teste é parte fundamental da otimização de modelos. O objetivo deste trabalho foi avaliar a influência da escolha do subconjunto de treinamento na construção dos modelos, bem como sua validação. Os métodos Kennard-Stone (KS) e a amostragem aleatória (AA) foram avaliados em dados de espectroscopia no infravermelho próximo (NIR) e em dados de marcadores SNPs (Single Nucleotide Polymorphisms). Vale destacar, que em dados SNPs, não há conhecimento de relatos na literatura a respeito da utilização do método KS. Para a construção e validação dos modelos, o método de estimação dos mínimos quadrados parciais (PLS) e Lasso bayesiano (BLASSO) mostraram-se mais eficientes para os dados NIR e para os dados SNPs, respetivamente. A avaliação da capacidade preditiva dos modelos obtidos após a partição dos dados ocorreu por meio da correlação entre os valores preditos e os valores reais, e da raiz quadrada do erro quadrático médio de predição. Para ambos os conjuntos de dados, os resultados indicam que os métodos KS e AA diferem estatisticamente entre si pelo teste F (valor P < 0.01), com o KS mais eficiente do que o AA em praticamente todas as repetições. Além disso, o método KS possui a vantagem de ser fácil e rápido de ser aplicado e também de selecionar sempre as mesmas amostras, o que proporciona grandes benefícios em futuras análises.*
Palavras-chave: *divisão de dados, regressão PLS, BLASSO, quimiometria, poder preditivo.*

INTRODUCTION

Under the framework of prediction modeling, splitting the data set into parts that will be used to build the models, as well as their validation, is a crucial step in their development. In particular, it is paramount to efficiently select the training set, as it will be used in the construction of the model (RAJER-KANDUČ et al., 2003). A proper model construction will guarantee more accurate predicting results in practice.

In general, data set samples can be separated into three complementary and disjoint subsets: calibration, validation, and prediction. The calibration and validation subsets are used for model construction, while the prediction subset is used to test its predictive ability (GALVÃO et al., 2005; LEE et al., 2018).

Different ways to obtain calibration and validation subsets are available in the literature. Among these, we can mention the Leave-one-out Cross-Validation and the *k*-Fold Cross-Validation (JAMES et al., 2014), the Kennard-Stone (KS) algorithm (DASZYKOWSKI et al., 2002; KENNARD & STONE, 1969; SOUSA et al., 2011),

Received 12.11.20 Approved 06.30.21 Returned by the author 10.01.21
CR-2020-1072.R1
Editors: Leandro Souza da Silva Alberto Cargnelutti Filho

and the Random Sampling (RS) (RAJER-KANDUČ et al., 2003).

The KS algorithm is often used in the area of chemometrics (HONORATO et al., 2007; ROQUE et al., 2019; SOUSA et al., 2011; WU et al., 1996) for presenting good results, and for being easy to be applied and understood. The RS method, also simple and easy to understand, is popular in statistical applications in several areas (AKDEMIR et al., 2015; HE,. et al., 2015; LONG et al., 2016). For these reasons, in this research, we will use the RS method and the KS algorithm, which here in will be referred to as the KS method.

After partitioning the dataset, the definition of the appropriate modeling method is of paramount importance in different data types. In Genomic Selection (GS), which is based on the study and analysis of data from SNP *(Single Nucleotide Polymorphisms)* markers (AZEVEDO, C. et al., 2013), DE LOS CAMPOS et al. (2009) suggested for modeling some statistical methods under Bayesian frameworks, such as the Bayesian Lasso (BLASSO), to perform the modeling.

In chemometrics, one way to obtain multivariate information of a specific sample is to use near-infrared (NIR) spectroscopy (MORGANO et al., 2008; TREVISAN & POPPI, 2006). For this kind of data, the partial least squares (PLS) regression method has shown to be useful because of its efficiency in dealing with experimental noises (TEÓFILO et al., 2009) and is still very popular in this area (PASQUINI, 2018).

In the literature, we can find some studies which compared the KS method with other methods of data partitioning (GALVÃO et al., 2005; SAPTORO et al., 2012; SIANO & GOICOECHEA, 2007). However, for the best of our knowledge, there is no literature showing the application of the KS method in the area of Genomic selection. The significant issue of this study is to evaluate the performance of the KS method in SNP data.

This research evaluated the RS and KS methods for the selection of calibration samples and to compare their corresponding predictive ability from the PLS and the BLASSO modeling in NIR data and SNP markers, respectively.

MATERIALS AND METHODS

This study was performed on two datasets, one for NIR and the other for SNP data. In summary, we first partition the dataset into three parts: calibration, validation, and prediction subsets. We used two methods for defining the samples belonging to this calibration subset: the KS and RS methods.

The way the calibration subset was created and how it affects the prediction ability of the fitted model is the subject of our comparison in this paper. Our interest was to find a calibration set that had greater representativeness in the complete data set, that is, to select samples that are uniformly distributed and capable of capturing the existing variability. Also, over the calibration and validation subsets, we fit prediction models using the PLS approach (on a NIR dataset) and the BLASSO approach (on an SNP dataset). The datasets, the partition methods, and the modeling approaches are described hereafter.

In this paper, we showed and compared the results of the prediction ability of each fitting model as a function of the type of partition (KS or RS) used to create the calibration subsamples. Therefore, to ease the understanding and to facilitate the writing, for short, herein in this paper, we refered to the outputs and discussion relating directly to the "KS or RS methods". It should be understood as "output and discussion referring to the model fitting when the calibration set is created using KS or RS partition methods".

Calibration subset construction

Kennard-Stone method (KS)

The KS method (KENNARD & STONE, 1969) performs the split of the original dataset into two subsets (calibration and validation) in such a way that each one contains samples that can capture the maximum variability of the original set (SOUSA et al., 2011).

This method uses the Euclidean distance for each pair (p, q) of samples to select the samples that will belong to the calibration subset:

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J}[x_p(j) - x_q(j)]^2} \quad p,q \in [1,N]$$

where J represents the number of covariates, and N corresponds to the sample size. In the case of NIR data, J represents the number of wavelengths, while in the SNP data, J is the number of markers. The remaining samples will compose the validation set.

When using the Euclidean distance, the samples that are most distant from each other are selected, resulting in a more uniform and comprehensive distribution of the calibration subset (SOUSA et al., 2011). A positive point of the KS method is that, for a particular dataset, the sample selected for calibration is unique.

Random Sampling method (RS)

The RS is a simple method for choosing the elements that will constitute the sample subsets. In general, it consists of selecting, without replacement, n < N elements from a list with N elementary units (BOLFARINE & BUSSAB, 2005).

RS method is a handy and straightforward technique, as it ensures that the selected subset follows the statistical distribution of the entire set (GALVÃO et al., 2005). This partitioning method is very efficient and has already been used in SNPs data sets (ZHOU & WANG, 2007) and also in NIR data (FERRAGINA et al., 2015; LEE et al., 2018; VAZQUEZ et al., 2012).

Model fitting

Partial Least Squares (PLS)

The PLS method originated in the area of econometrics (WOLD 1982), and was later published in three distinct studies (WOLD, S. et al., 1984a; WOLD,; et al., 1984b; WOLD et al., 1987). According to BROWN (1995), PLS began to be successfully applied also in the area of chemometrics.

PLS is a method based on the compression of data into latent variables, which are linear combinations of the original variables. These latent variables are orthogonal components, which eliminates the problem of multicollinearity (GOKTAS & YENAY, 2020; RESENDE et al., 2014).

For the execution of PLS, one can use the SIMPLS algorithm, based on the singular value decomposition. Detailed information about this method can be obtained in DE JONG (1993).

Bayesian Lasso Method (BLASSO)

The BLASSO method, a Bayesian version of the LASSO regression (TIBSHIRANI, 1996), was proposed by DE LOS CAMPOS et al. (2009) as an alternative to performing the Whole Genomic Selection analysis. In general, BLASSO is preferred, compared to LASSO, as it has no restrictions on the number of covariates, besides being more stable when it has high dimensionality (RESENDE et al., 2012).

The general linear model for predicting the effects is given by:

$y = 1\mu + X\beta + e$

where y is the vector of phenotypic observations (N×1), N is the number of genotyped and phenotyped individuals; 1 is the column vector of 1's (N×1); μ is the mean of the variable y; X is the incidence matrix (N×p); β is the vector that contains the regression coefficients, of dimension p×1; e is the residuals vector (N×1).

According to De Los Campos et al. (2009) when using BLASSO the following distributions are assumed:

$e|\sigma^2 \sim MVN(0, I\sigma^2)$

$b_i|\lambda, \sigma^2 \sim \prod_i \left(\frac{\lambda}{2\sigma}\right) e^{\left[\frac{-\lambda|b_i|}{\sigma}\right]}$

where MNV refers to the multivariate normal distribution; λ is the smoothing parameter; σ^2 is the variance component that has a scaled inverted chi-squared distribution a *priori*.

Using a formulation in terms of an increased hierarchical model, we have:

$b_i|\tau \sim N(0, D\sigma^2)$ where $D = diag(\tau_1^2, \tau_2^2, \ldots, \tau_p^2)$

$p(\tau^2|\lambda^2) = \prod_i \left(\frac{\lambda^2}{2}\right) e^{\left[\frac{-\lambda^2\tau_i^2}{2}\right]}$

According to Park and Casella (2008), this formulation leads to a double exponential distribution for the coefficients, that is:

$b_i|\lambda^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda}\right)$

The λ smoothing parameter causes some effects to be approximately zero, but not effectively zero. Further detailed explanation about BLASSO can be obtained in DE LOS CAMPOS et al. (2009), RESENDE et al. (2014), and PARK & CASELLA (2008).

Datasets

NIR dataset

The Genetic Program for the Improvement of Sugarcane (PMGCA) of the Federal University of Viçosa (UFV), Viçosa, Minas Gerais, Brazil, provided the NIR dataset to be considered in our study (ASSIS et al., 2017).

A total of N = 256 samples were analyzed to predict the sugarcane lignin content. The NIR spectra were obtained directly from the middle third of the +3 leaf of each genotype, without any special sample preparation. We arranged the spectra data in a 256×1038 matrix, where the rows correspond to the samples and the columns to the NIR wavelengths.

SNP dataset

In this case, we considered a corn yield dataset in an irrigated condition presented by CROSSA et al., (2010). The grain yield was evaluated as a response variable. The number of rows (individuals) in the grain yield dataset is N = 264. There were 1135 SNP markers available for analysis. Therefore, in this case, the matrix has a dimension of 264×1135. Therefore, in this case, the matrix has a dimension of 264 x times1135.

Model Construction

For the construction of the proposed models and subsequent evaluation, it was necessary to partition each dataset into three parts: calibration, validation, and prediction with sample sizes $n1$, $n2$ and n, respectively. Initially, n = 36 samples (~ 14%) were taken, randomly, to constitute the prediction subset in each dataset. The remaining N – n samples were used to form the calibration and validation subsets. Subsequently, the KS and RS methods were applied so that the calibration and validation subsets consisted, respectively, of $n1$ = 198 (77%) and $n2$= 22 (9%) samples for NIR dataset, and of $n1$ = 205 (77%) and $n2$ = 23 (9%) samples for the SNP dataset.

The efficiency of the constructed models was verified on the prediction subset using the following evaluation indexes (FERREIRA, M., 2015): the Pearson's correlation coefficient (r) and the square root of the mean squared error (RMSE) between the original observed values and the corresponding predicted values obtained by a particular model, i.e.

$$r = \frac{[\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})]}{\sqrt{[\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}} \quad \text{and} \quad \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

where: y_i and $\hat{y}_i$ are the observed and predicted values, respectively; $\bar{\hat{y}}$ is the predicted mean value; $\bar{y}$ is the observed mean value and n is the number of samples belonging to each prediction subset.

To improve the statistical significance of the RS method results, we repeated this procedure k times in both datasets (Figure 1). In the NIR dataset, the RS method was repeated k = 40 times; that is, 40 different partitions were constructed, resulting in one KS model and 40 RS models. In the SNPs dataset, the RS method was repeated k = 10 times, thus obtaining one KS model and 10 RS models.

The calibration and validation subsets were used for the construction of the corresponding models, while the prediction subset was used to evaluate their predictive ability.

The entire process, as illustrated in figure 1, was repeated 15 times for both datasets to compensate for any sampling problems and to improve the statistical significance of the comparison between the RS and KS methods. To check whether the average value of the RMSE of the RS method differs statistically from the RMSE of the KS method, the F test (ANOVA) was used (GALVÃO et al., 2005).

In order to minimize the systematic effect existing in the random selection and the heterogeneity present between the repetitions, a randomized block design (in which each repetition represents a block, totaling 15 blocks) was applied, at the 1% significance level. Thus, 15 pairs of RMSE and r values were obtained for in both datasets, and 600 pairs for RS in the NIR dataset and 150 pairs for RS in the SNPs dataset.

Computational resources

All computational routines employed were implemented in the R software (R DEVELOPMENT

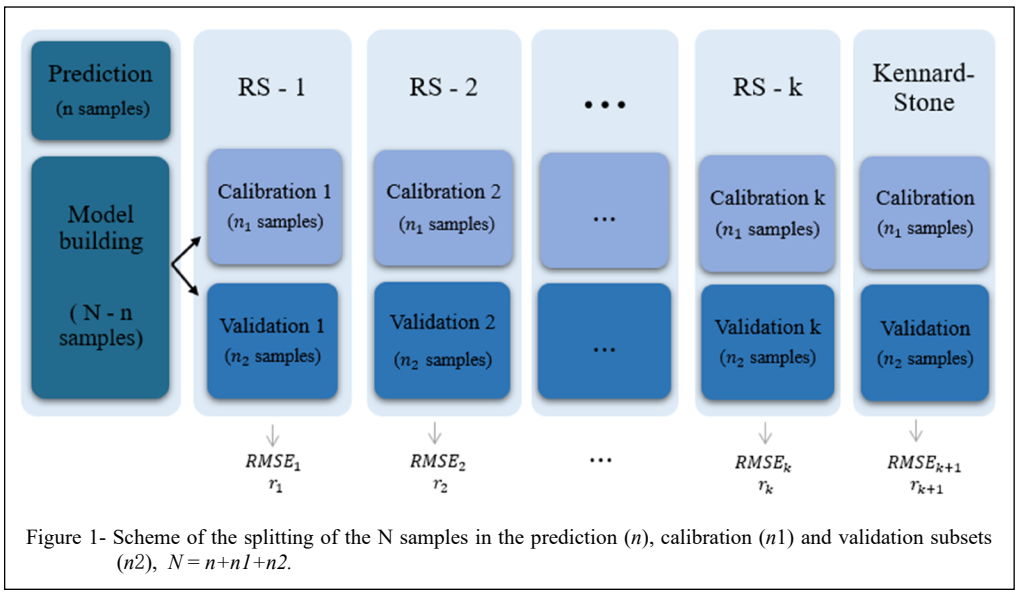

Figure 1- Scheme of the splitting of the N samples in the prediction (n), calibration ($n1$) and validation subsets ($n2$), $N = n+n1+n2$.

CORE TEAM, 2019). For the KS method, we used the kenStone function of the prospectr package (STEVENS; RAMIREZ-LOPEZ, 2013), in the RS method, the sample function of the base package was used. To adjust the model with the BLASSO method, we used the bglr function of the BGLR package (PÉREZ; LOS CAMPOS, DE, 2014). After an evaluation phase, it was decided to use 25,000 iterations, with *burn-in* of 10,000, and *thin* of 3. To calibrate the models with the PLS method, we considered the plsr function of the pls package with 20 principal components, the SIMPLS multivariate regression method, and a leave-one-out internal cross-validation (MEVIK; WEHRENS, 2007).

RESULTS AND DISCUSSION

NIR dataset

The original NIR spectra of the 256 sugarcane leave samples are shown in figure 2.

For the analysis and modeling of NIR spectroscopy data, we should apply some pretreatments on the data matrix as a way to remove or minimize the sources of spectral variability (including noise) and to improve selectively (PASQUINI, 2018). In our study, the best pretreatments, in most situations, were: mean centering, SG smoothing (polynomial degree = 2, and window size = 5), and multiplicative signal correction (MSC). Details about these pretreatments can be reported elsewhere (FERREIRA, M., 2015).

Figure 3 contains the model performance in terms of RMSE and r for the 15 different partitions of the original dataset. It was observed that the values from the KS partition method were, in general, better than those obtained from RS within all repetitions. The variability of RMSE values is similar within each repetition (Figure 3-a), indicating homogeneity. However, the variation among repetitions is considerably high. For example, the lowest RMSE value in repetition 8 is 2.14, while in repetition 12 the minimum is 0.85. . The difference among repetitions should be expected since we used a different sample split that generated each repetition, as illustrated in figure 1. The same behavior could be seen in figure 3-b for the correlations. To ease the comparisons, we averaged the values of RMSE and r from the RS method (Table 1).

In all repetitions, the RMSE values of the KS method were lower than the average of the RS method (Table 1). Also, the correlation values of the KS method were higher than the corresponding average correlations of the RS method in all repetitions. For the RMSE statistic, the superiority of KS over RS as a base for creating the prediction model could be verified by the F test ($F = 14.87$, $P < 0.01$, with 1 and 599 degrees of freedom). This result showed that the prediction ability of the PLS model is better, on average, when the KS method was used to create the calibration subset.

It was observed in repetition 12 (Table 1) the best results, both for KS ($r = 0.92$ and RMSE = 0.93) and for RS ($r = 0.87$ and RMSE = 1.15). In average terms, we have a correlation of 0.74 for RS and 0.80 for KS, which represents an improvement of approximately 8% in the results obtained. To visualize the partitions and to have an insight on the reasons the outputs differ, we picked up one of the repetitions (repetition 12) among the 15 available, and also one of the random choices from the RS method in this repetition. Figure 4 shows how the calibration and

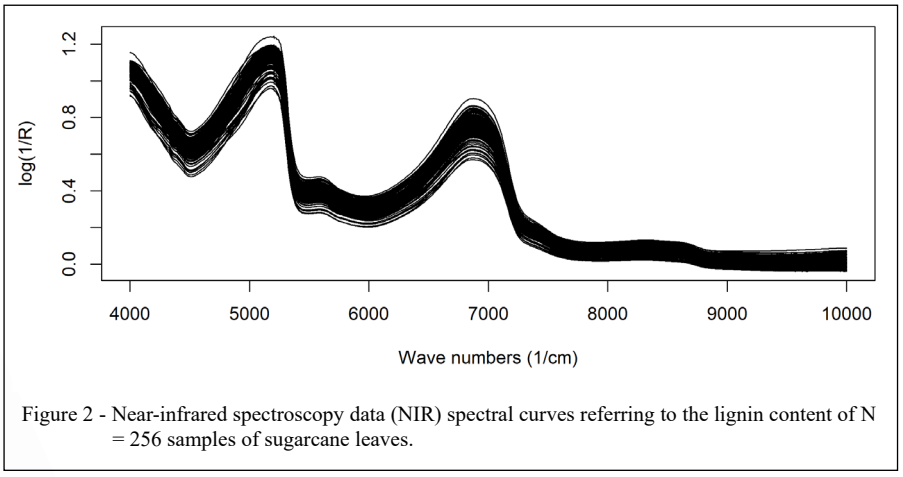

Figure 2 - Near-infrared spectroscopy data (NIR) spectral curves referring to the lignin content of N = 256 samples of sugarcane leaves.

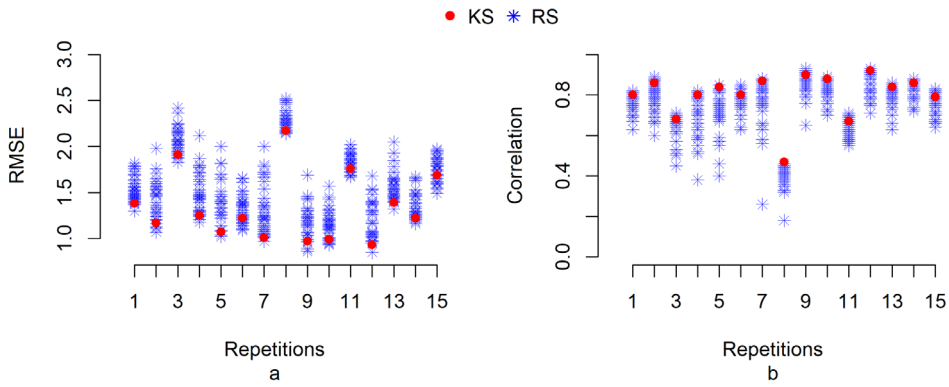

Figure 3 - Observed values of (a) square root of the mean squared error (RMSE) and (b) Pearson's correlation (Correlation), between the predicted and observed values on the prediction subsets, in the 15 repetitions performed on the NIR data set. In each repetition, the red dot represents the value obtained by the fitting model on the dataset partition following the Kennard-Stone method (KS), and the blue asterisks denote the corresponding values on the k = 40 dataset partitions provided by the Random Sampling (RS) method.

validation samples were distributed in this choice. Since the prediction accuracy is calculated on the prediction subsample, we added, on the next plot, the corresponding prediction subset (Figure 5).

Table 1 - Correlation coefficient (r) and the square root of the mean squared error (RMSE) between the predicted ($\hat{y}_p$) and the observed (y) values belonging to the prediction subset, evaluated by the PLS method.

Repetition	----------RS----------		----------KS----------	
	r	RMSE	r	RMSE
1	0.75	1.54	0.80	1.38
2	0.78	1.42	0.86	1.17
3	0.63	2.04	0.68	1.91
4	0.70	1.46	0.80	1.25
5	0.70	1.39	0.84	1.07
6	0.76	1.33	0.80	1.22
7	0.77	1.32	0.87	1.01
8	0.40	2.26	0.47	2.17
9	0.85	1.15	0.90	0.97
10	0.83	1.15	0.88	0.99
11	0.63	1.84	0.67	1.76
12	0.87	1.15	0.92	0.93
13	0.79	1.57	0.84	1.39
14	0.82	1.34	0.86	1.22
15	0.75	1.76	0.79	1.69

RS: Random Sampling partition method; KS: Kennard-Stone partition method. For KS, the values are single. For RS, the average is over 40 repetitions, as shown in figure 3.

Although very important, RS does not guarantee the representativeness of the complete data set (RAJER-KANDUČ et al., 2003). Also, unlike KS, it does not ensure that the most distant samples, that is, of more significant dissimilarity, are included in the sampled set.

In figures 4-a and 5-a you can have an idea that the samples obtained in the KS capture some more extreme points to compose the calibration set. This resulted is expected since this method is based on the Euclidean distance between the samples. Thus it selects samples that are more distant from each other and; consequently, captures more data variability (SOUSA et al., 2011). These samples could be considered more representative in the calibration set (WU et al., 1996).

|Conversely, because the RS method is selecting samples at random, it may not be able to capture all the variability present in the data set (GALVÃO et al., 2005). As much as the method may select more distant samples, due to randomness, this is less likely (Figure 4-b). In this case, the prediction samples may not belong to the region where the calibration samples were selected (Figure 5-b), which is possibly causing a lower prediction accuracy when compared to the KS method. The visualization may not be as evident as the two-dimensional graph shows only the first two main components. MORAIS et al., (2019), working with spectroscopy data, also found that KS had better prediction accuracy than RS in several data sets. Likewise, WU et al., (1996) also

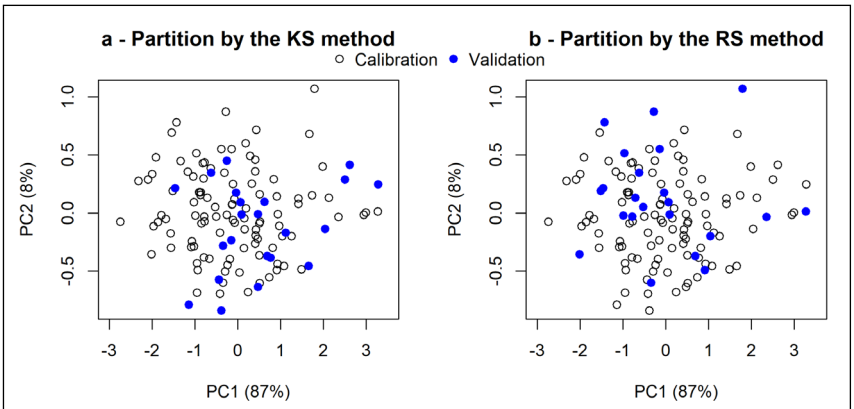

Figure 4 - Data partition using the Kennard-Stone (KS) and random sampling (RS) methods. PC1 and PC2 correspond to the first and second principal components used to facilitate two-dimensional graphic visualization. The percentage values refer to the proportion of the variance explained by each component.

compared the use of KS with other methods of data partition, including RS, and concluded that KS was superior. These authors also point out that the sample size was not large enough to generalize such a result, but they affirmed the use of KS as a useful approach to partition the data. However, we performed the comparison process 15 times (15 replications) over a dataset with 256 samples, in order to make our results even more robust.

In the NIR data approach, the KS method is widely used for data partitioning (GOGÉ et al., 2012; NASCIMENTO et al., 2016; ROQUE et al., 2019). Some studies affirmed the use of the KS method for partitioning data sets compared to other methods (SAPTORO et al., 2012; SIANO; GOICOECHEA, 2007; WU et al., 1996). However, to the best of our knowledge, there are no reports in the literature regarding the use of the KS method applied to SNPs data sets.

SNPs dataset

The RMSE and correlations obtained from the BLASSO method on the SNP dataset are shown

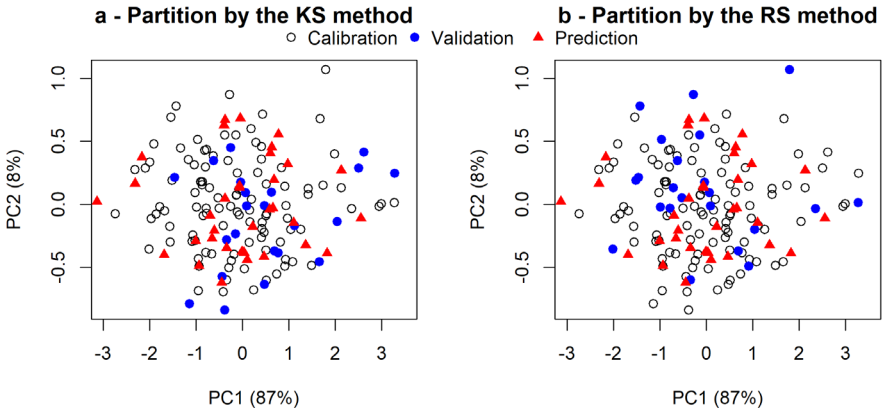

Figure 5 - Data partition using the Kennard-Stone (KS) and random sampling (RS) methods. PC1 and PC2 correspond to the first and second principal components used to facilitate two-dimensional graphic visualization. Percentage values refer to the proportion of variance explained by each component. Notice that the prediction data is identical in both graphs.

in figure 6. The RMSE values obtained from the KS method were, in general, lower than those from the RS method in almost all repetitions. We also noticed the same kind of behavior, compared to the NIR dataset, for the RMSE and correlation values within and among the repetitions. Within repetition, the variability was very similar, but the location parameter varied among them. This difference in the location parameter among repetition would be expected since we used different seeds to perform the split for each of these repetitions. To circumvent this systematic effect when comparing the KS versus RS methods, we considered each replication as a block in a randomized block design (DBC) analysis. The average values of RMSE and r from the RS method and the RMSE and r from the KS method are presented in table 2.

It was observed in repetition 4 (Table 2), that in terms of RMSE, we have the best results, both for KS (RMSE = 0.64) and for RS (RMSE = 0.69). The average RMSE was higher for the RS method than for the KS method in 14 out of the 15 repetitions. In general, considering all repetitions, the mean RMSE of KS and RS differ from each other by the F test ($F = 30.54$, $P < 0.01$, with 1 and 149 degrees of freedom). Thus, when using the SNPs data set, we verified that there is a significant difference between the RMSE means obtained by the two methods, at the level of 1% significance, which indicates the superiority of the KS approach aiming to split the data set with the purpose of prediction. We can also observe that the KS correlation was higher than the mean correlation of RS, except in repetitions 5 and 13. In average terms, we have a correlation of 0.41 for RS and 0.54 for KS, which represents an improvement of approximately 32% in the results obtained.

Considering the results from both datasets analysis, the mean prediction values (either for RMSE or for R) obtained when using the KS method for defining the calibration set are statistically better than those from the RS method. Better, in this context, means smaller for RMSE, and higher for R. Furthermore, a great advantage of using KS is the ease and speed the results are obtained. While KS is done only once, the RS method must be repeated several times, thus generating higher computational effort. We should also highlight that, in the same population, KS always selects the same samples, which eventually may facilitate other further analysis.

A disadvantage of using the KS method is the impossibility of obtaining a measure of the sample variability directly. Conversely, when performing the RS method, it is possible to get such a quantity, which even allows us to define confidence intervals if it is of the researcher's interest.

CONCLUSION

The Random Sampling (RS) and the Kennard-Stone (KS) methods for partitioning a data

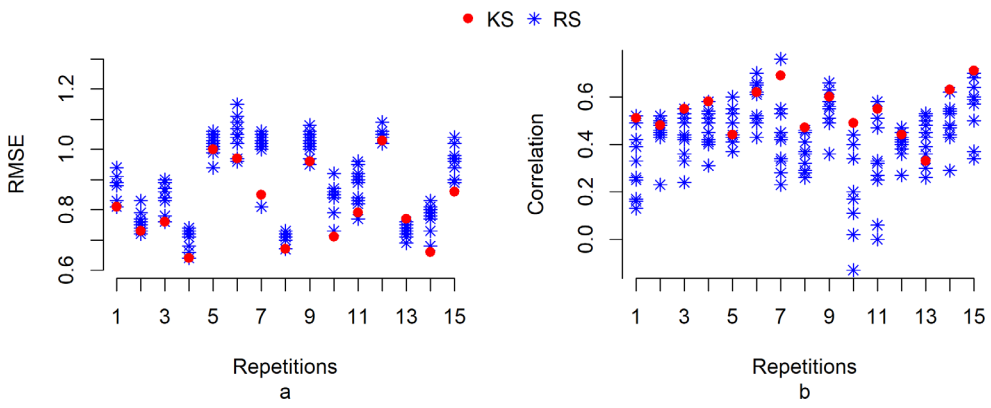

Figure 6 - Observed values of (a) square root of the mean squared error (RMSE) and (b) Pearson's correlation (Correlation) between the predicted and observed values on the prediction subsets in the 15 repetitions performed on the SNPs data set. In each repetition, the red dot represents the value obtained by the fitting model on the dataset partition following the Kennard-Stone method (KS), and the blue asterisks denote the corresponding values on the k = 10 dataset partitions provided by the Random Sampling (RS) method.

Table 2 - Mean values of the correlation coefficient (r) and the square root of the mean quadratic error (RMSE) between the predicted value ($\hat{y}_p$) and the real (y) belonging to the prediction subset, evaluated by the BLASSO method.

Repetition	----------RS----------		----------KS----------	
	r	RMSE	r	RMSE
1	0.31	0.90	0.51	0.81
2	0.45	0.76	0.48	0.73
3	0.43	0.84	0.55	0.76
4	0.46	0.69	0.58	0.64
5	0.49	1.02	0.44	1.00
6	0.58	1.06	0.62	0.97
7	0.43	1.01	0.69	0.85
8	0.36	0.71	0.47	0.67
9	0.54	1.02	0.6	0.96
10	0.18	0.84	0.49	0.71
11	0.39	0.88	0.55	0.79
12	0.39	1.05	0.44	1.03
13	0.38	0.72	0.33	0.77
14	0.37	0.78	0.63	0.66
15	0.37	0.97	0.71	0.86

RS represents random sampling.
KS indicates the Kennard-Stone method.

set in training and testing subsamples allow for predicting models that are statistically different. The KS method, when compared to the RS method, presents lower mean RMSE values and mean correlations, between the predicted and observed values, higher in the prediction subset.The KS method has a better predictive ability, besides requiring shorter execution time and computational effort. The KS method may be a good alternative to commonly used partition methods also for SNP data.

ACKNOWLEDGMENTS

This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil - Finance code 001. We also thank the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG) for the financial support of research projects and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the research scholarships. Finally, we thank RIDESA, the Inter-University Network for the Development of the Sugarcane Industry in Brazil, for providing the dataset.

DECLARATION OF CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS' CONTRIBUTIONS

Conceptualization: LAP. Data acquisition: LAP, GT and RAF. Design of methodology and data analysis: LAP, GT and RAF. GT and RAF prepared the draft of the manuscript. LAP critically revised the manuscript and approved of the final version.

REFERENCES

AKDEMIR, D. et al. Optimization of genomic selection training populations with a genetic algorithm. **Genetics Selection Evolution**, v.47, n.1, p.1–10, 2015. Available from: <https://doi.org/10.1186/s12711-015-0116-6>. Accessed: Nov. 19, 2019. doi: 10.1186/s12711-015-0116-6.

ASSIS, C. et al. Prediction of lignin content in Different Parts of Sugarcane Using Near-Infrared Spectroscopy (NIR), Ordered Predictors Selection (OPS), and Partial Least Squares (PLS). **Applied Spectroscopy**. v.71, n.8, p.2001–2012, 2017. Available from: <https://doi.org/10.1177/0003702817704147>. Accessed: Sept. 02, 2019. doi: 10.1177/0003702817704147.

AZEVEDO, C. et al. Independent component regression applied to genomic selection for carcass traits in pigs. **Pesquisa Agropecuaria Brasileira**, v.48, n.6, p.619–626, 2013. Available from: <https://doi.org/10.1590/S0100-204X2013000600007>. Accessed: Mar. 29, 2020. doi: 10.1590/S0100-204X2013000600007.

BOLFARINE, H.; BUSSAB, W. O. **Elementos de Amostragem**. São Paulo: ABE - Projeto Fisher, Edgard Blücher, 2005.

BROWN, S. Chemical Systems Under Indirect Observation: Latent Properties and Chemometrics. **Applied Spectroscopy**, v.49, n.12, p.14A-31A, 1995. Available from: <https://doi.org/10.1366/0003702953965876>. Accessed: Apr. 29, 2020. doi: 10.1366/0003702953965876.

CROSSA, J. et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v.186, n.2, p.713–724, 2010. Available from: <https://doi.org/10.1534/genetics.110.118521>. Accessed: Jan. 5, 2020. doi: 10.1534/genetics.110.118521.

DASZYKOWSKI, M. et al. Representative subset selection. **Analytica Chimica Acta**, v. 468, n.1, p.91–103, 2002. Available from: <https://doi.org/10.1016/S0003-2670(02)00651-7>. Accessed: Jul. 04, 2019. doi: 10.1590/S0100-204X2013000600007.

DE JONG, S. SIMPLS: an alternative approach to partial least squares regression. **Chemometrics and intelligent laboratory systems**, v.18, p.251–263, 1993. Available from: <https://doi.org/10.1016/0169-7439(93)85002-X >. Accessed: Jul. 04, 2019. doi: 10.1016/0169-7439(93)85002-X.

DE LOS CAMPOS, G. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v.182, n.1, p.375–385, 2009. Available from: < https://doi.org/10.1534/genetics.109.101501>. Accessed: Nov. 17, 2019. doi: 10.1534/genetics.109.101501.

FERRAGINA, A. et al. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. **Journal of Dairy Science**, 2015. p.1–19. Available from: <http://dx.doi.org/10.3168/jds.2014-9143>. Accessed: Jan. 20, 2019. doi: 10.3168 / jds.2014-9143.

FERREIRA, M.M.C **Quimiometria – Conceitos, Métodos e Aplicações.** Campinas, SP: Editora Unicamp, 2015. 496 f.

GALVÃO, R. K. H. et al. A method for calibration and validation subset partitioning. **Talanta**, v.67, n.4, p.736–740, 2005. Available from: <https://doi.org/10.1016/j.talanta.2005.03.025>. Accessed: Jun. 02, 2020. doi: 10.1016/j.talanta.2005.03.025.

GOGÉ, F. et al. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. **Chemometrics and Intelligent Laboratory Systems**, v.110, n.1, p.168–176, 2012. Available from: <http://dx.doi.org/10.1016/j.chemolab.2011.11.003>. Accessed: Jun. 29, 2020. doi: 10.1016/j.chemolab.2011.11.003.

GOKTAS, A.; YENAY, O. Comparison of partial least squares with other prediction methods via Comparison of partial least squares with other prediction methods via generated data Atila Göktaş & Özge Akkuş. **Journal of Statistical Computation and Simulation**, n. August, 2020. Available from: <https://doi.org/10.1080/00949655.2020.1793342>. Accessed: Nov. 15, 2020. doi: 10.1080/00949655.2020.1793342.

HE, Z. et al. Design of a reference value-based sample-selection method and evaluation of its prediction capability. **Chemometrics and Intelligent Laboratory Systems**, v.148, p.72–76, 2015. Available from: <http://dx.doi.org/10.1016/j.chemolab.2015.09.001>. Accessed: Jan. 10, 2020. doi: 10.1016/j.chemolab.2015.09.001.

HONORATO, F. A. et al. Calibration transfer in multivariate methods. **Quimica Nova**, v.30, n.5, p.1301–1312, 2007. Available from: <https://doi.org/10.1590/S0103-50532010000100019>. Accessed: Jan. 10, 2020. doi: 10.1590/S0103-50532010000100019.

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in R**. New York: Springer US, 2014. V. 102. DOI 10.1007/978-1-4614-7138-7.

KENNARD, R.; STONE, L. Computer Aided Design of Experiments. **Technometrics**, v. 11, n. 1, p. 137–148, 1969. Available from: <http://www.jstor.org/stable/1266770>. Accessed Nov. 29, 2020. doi:10.2307/1266770.

LEE, L. C. et al. Iterative random vs. Kennard-Stone sampling for IR spectrum-based classification task using PLS2-DA. **AIP Conference Proceedings**, v.1940, 2018. Available from: <https://doi.org/10.1063/1.5028031>. Accessed: Jan. 22, 2020. doi: 10.1063/1.5028031.

LONG, J. et al. Prevalence and correlates of problematic smartphone use in a large random sample of Chinese undergraduates. **BMC Psychiatry**, v.16, n.1, p.1–12, 2016. Available from: <http://dx.doi.org/10.1186/s12888-016-1083-3>. Accessed Nov. 04, 2019. doi: 10.1186/s12888-016-1083-3.

MEVIK, B.-H.; WEHRENS, R. The pls package: Principal Component and Partial Least Squares Regression in R. **Journal of Statistical Software**, v.18, n.2, 2007. Available from: <https://www.jstatsoft.org/article/view/v018i02>. Accessed: Mar. 19, 2020. doi: 10.18637/jss.v018.i02.

MORAIS, C. L. M. et al. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. **Bioinformatics**. v.35, n.24, p.5257–5263, 2019. Available from:<https://doi.org/10.1093/bioinformatics/btz421>. Accessed: Mar. 19, 2020. doi: 10.1093/bioinformatics/btz421.

MORGANO, M. A. et al. Determinação de umidade em café cru usando espectroscopia NIR e regressão multivariada. **Ciencia e Tecnologia de Alimentos**, Campinas, v.28, n.1, p.12–17, 2008. Available from: <http://dx.doi.org/10.1590/S0101-20612008000100003>. Accessed: Mar. 19, 2020. doi: 10.1590/S0101-20612008000100003.

NASCIMENTO, P. A. M. et al. Robust PLS models for soluble solids content and firmness determination in low chilling peach using near-infrared spectroscopy (NIR). **Postharvest Biology and Technology**, v.111, p.345–351, 2016. Available from: <http://dx.doi.org/10.1016/j.postharvbio.2015.08.006>. Accessed: Mar. 19, 2020. doi: 10.1016/j.postharvbio.2015.08.006.

PARK, T., CASELLA, G. The bayesian lasso. **Journal of the American Statistical Association**, v.103, n.482, p.681-686, 2008. Available from: <https://doi.org/10.1198/016214508000000337>. Accessed: Mar. 19, 2020. doi: 10.1198/016214508000000337.

PASQUINI, C. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. **Analytica Chimica Acta**, v.1026, p.8–36, 2018. Available from: <https://doi.org/10.1016/j.aca.2018.04.004>. Accessed: Mar. 19, 2020. doi: 10.1016/j.aca.2018.04.004.

PÉREZ, P.; LOS CAMPOS, G. DE. Genome-Wide Regression and Prediction with the BGLR Statistical Package. **Genetics**, v.198, p.483–495, 2014. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196607/>. Accessed: Mar. 19, 2020. doi: 10.1534/genetics.114.164442.

R CORE TEAM. **R**: A Language and environment for Statistical Computing. Available from: <https://www.r-project.org/>. Accessed: Nov. 19, 2019.

RAJER-KANDUČ, K. et al. Separation of data on the training and test set for modelling: A case study for modelling of five colour properties of a white pigment. **Chemometrics and Intelligent Laboratory Systems**, v.65, n.2, p.221–229,2003. Available from: <https://doi.org/10.1016/S0169-7439(02)00110-7>. Accessed: Mar. 19, 2020. doi: 10.1016/S0169-7439(02)00110-7.

RESENDE, M. D. V. et al. **Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial**. Minas Gerais: Viçosa, 2012. 291p.

RESENDE, M.D. V. et al. **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**. Minas Gerais: Viçosa, 2014. 881p.

ROQUE, J. V. et al. Comprehensive new approaches for variable selection using ordered predictors selection. **Analytica Chimica Acta**. v.1075, p.57–70, 2019. Available from: <https://doi.org/10.1016/j.aca.2019.05.039>. Accessed: Mar. 19, 2020. doi: 10.1016/j.aca.2019.05.039.

SAPTORO, A. et al. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. **Chemical Product and Process Modeling**, v.7, n.1,

2012. Available from: <https://doi.org/10.1515/1934-2659.1645 >. Accessed: Mar. 19, 2020. doi: 10.1515/1934-2659.1645.

SIANO, G. G.; GOICOECHEA, H. C. Representative subset selection and standardization techniques. A comparative study using NIR and a simulated fermentative process UV data. **Chemometrics and Intelligent Laboratory Systems**, v.88, n.2, p.204-212, 2007. Available from: <https://doi.org/10.1016/j.chemolab.2007.05.002>. Accessed: Mar. 19, 2020. doi: 10.1016/j.chemolab.2007.05.002.

SOUSA, L. C. et al. Development of nirs calibration models for minimization of Eucalyptus spp wood analysis. **Ciencia Florestal**, v.21, n.3, p.91-599, 2011. Available from: <http://dx.doi.org/10.5902/198050983817>. Accessed: Mar. 19, 2020. doi: 10.5902/198050983817.

STEVENS, A.; RAMIREZ-LOPEZ, L. An introduction to the prospectr package. R package version 0.2.0.. **R package Vignette**, 2013. n. May. Available from: <https://cran.r-project.org/web/packages/prospectr/citation.html>. Accessed: Mar. 19.

TEÓFILO, R. F. et al. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. **Journal of Chemometrics**, v.23, n.1, p.32–48, 2009. Available from: <https://doi.org/10.1002/cem.1192>. Accessed: Mar. 19, 2020. doi: 10.1002/cem.1192.

TIBSHIRANI, R. Regression Shrinkage and Selection via the Lasso. **Journal of the Royal Statistical Society**, v.58, n.1, p.267-288, 1996. Available from: <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>. Accessed: Mar. 19, 2020. doi: 10.1111/j.2517-6161.1996.tb02080.x.

TREVISAN, M. G.; POPPI, R. J. Química Analítica de Processos. **Química Nova**, v.29, n.5, p.1065–1071, 2006. Available from: <http://dx.doi.org/10.1590/S0100-40422006000500029>. Accessed: Mar. 19, 2020. doi: 10.1590/S0100-40422006000500029.

VAZQUEZ, A. I. et al. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. **Genetics**, v.192, n.4, p.1493-1502, 2012. Available from: <https://doi.org/10.1534/genetics.112.141705>. Accessed: Mar. 19, 2020. doi: 10.1534/genetics.112.141705.

WOLD, H. Soft modeling: the basic design and some extensions. **Systems under Indirect Observation**, Part 2, North-Holland, Amsterdam, p. 1-54, 1982. Accessed: Mar. 19, 2020.

WOLD, S. et al. Principal component analysis. **Chemometrics and intelligent laboratory systems**, 1987. v.2, p.37–52. Available from: <https://doi.org/10.1016/0169-7439(87)80084-9>. Accessed: Mar. 19, 2020. doi: 10.1016/0169-7439(87)80084-9.

WOLD, S.; et al. Multivariate data Analysis in Chemistry. Dordrecht: **Chemometrics**, p.17–95, 1984a. Available from: <https://doi.org/10.1007/978-94-017-1026-8_2>. Accessed: Mar. 19, 2020. doi: 10.1007/978-94-017-1026-8_2.

WOLD, S.;. et al. The partial least squares (PLS) approach to generalized inverses. **SIAM Journal on Scientific and Statistical Computing**, p.735–743, 1984b. Available from: <https://doi.org/10.1137/0905052>. Accessed: Mar. 19, 2020. doi: 10.1137/0905052.

WU, W. et al. Artificial neural networks in classification of NIR spectral data: Design of the training set. **Chemometrics and Intelligent Laboratory Systems**, v.33, n.1, p.35–46, 1996. Available from: <https://doi.org/10.1016/0169-7439(95)00077-1>. Accessed: Mar. 19, 2020. doi: 10.1016/0169-7439(95)00077-1.

ZHOU, N.; WANG, L. Effective selection of informative SNPs and classification on the HapMap genotype data. **BMC Bioinformatics**, v.8, n.1, p.484, 2007. Available from: <https://doi.org/10.1186/1471-2105-8-484>. Accessed: Mar. 19, 2020. doi: 10.1186/1471-2105-8-484.