ROBERTA DE AMORIM FERREIRA

COMPARAÇÃO DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM REGRESSÃO APLICADOS A DADOS GENÔMICOS E DE ESPECTROSCOPIA NIR

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
F383c
2018

Ferreira, Roberta de Amorim, 1990-
Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia NIR / Roberta de Amorim Ferreira. – Viçosa, MG, 2018.
xiv, 53 f. : il. (algumas color.) ; 29 cm.

Orientador: Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: r. 47-53.

1. Análise dimensional. 2. Teoria bayesiana de descisão estatística. 3. Espectroscopia de infravermelho. 4. Marcadores genéticos. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 530.8

ROBERTA DE AMORIM FERREIRA

COMPARAÇÃO DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS EM REGRESSÃO APLICADOS A DADOS GENÔMICOS E DE ESPECTROSCOPIA NIR

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 21 de fevereiro de 2018.

Felipe Lopes da Silva

Camila Ferreira Azevedo

Reinaldo Francisco Teófilo
(Coorientador)

Luiz Alexandre Peternelli
(Orientador)

A minha mãe, Maria Inês.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me dado forças e ter me guiado durante todo o caminho para que fosse possível completar mais uma fase da minha vida.

À minha mãe que esteve sempre ao meu lado, por sua dedicação, amor e paciência sempre me apoiando e me mostrando que nada é impossível, pessoa que sigo como exemplo.

Ao meu irmão Gustavo, e a meu sobrinho Kevin meu agradecimento especial.

A todos os meus familiares, primos, primas, tios e tias, pela torcida e apoio.

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade concedida para realização do curso.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudos.

Ao professor Luiz Alexandre Peternelli pela oportunidade, orientação, compreensão, confiança, paciência e pelos ensinamentos profissionais e de vida. Sem você nada disso seria possível. Obrigada por tudo!

Aos membros da banca examinadora pelas contribuições e críticas ao trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, por contribuírem para minha formação acadêmica.

Aos meus amigos do PPESTBIO pelos momentos de descontração, pelas trocas de experiência, pelas palavras de conforto e constantes incentivos. Em especial ao Denilson, Jaqui, Leísa, Gabi, Carol, Lucas e Ithalo.

As minhas amigas Larissa, Angélica e Mabi pela torcida e amizade.

A todos que de alguma forma contribuíram para a realização deste trabalho.

BIOGRAFIA

Roberta de Amorim Ferreira, filha de Maria Inês de Amorim Ferreira e Roberto Divino Ferreira, nasceu em João Monlevade, Minas Gerais, em 06 de dezembro de 1990.

Em julho de 2015, graduou-se em Licenciatura em Matemática pela Universidade Federal de Viçosa, Viçosa-MG.

Em fevereiro de 2016 ingressou no curso de Mestrado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

FERREIRA, Roberta de Amorim, M.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia NIR.** Orientador: Luiz Alexandre Peternelli. Coorientador: Reinaldo Francisco Teófilo.

Muitas áreas de pesquisa possuem conjuntos de dados com os desafios da alta dimensionalidade e multicolinearidade a serem superados, de modo que métodos específicos para ajuste do modelo devem ser empregados. Embora os métodos existentes sejam eficientes para construção do modelo, frequentemente se faz necessário selecionar as variáveis mais importantes em explicar o modelo, visto que essa prática pode aumentar sua capacidade preditiva, diminuir custos e tempo das análises. Esse trabalho teve como objetivo principal avaliar e construir modelos empregando três métodos de seleção de variáveis aplicados a dados de marcadores SNPs (*Single Nucleotide Polymorphisms*) e a dados de espectroscopia no infravermelho próximo (NIR), além de avaliar a melhoria na qualidade de predição, quando comparado ao uso dos dados completos. Os métodos avaliados foram o de seleção dos preditores ordenados associado a regressão por quadrados mínimos parciais (PLS-OPS), o *Sparce partial least Square* (SPLS) e o Lasso bayesiano (BLASSO) supervisionado, este último é uma adaptação do método BLASSO com a vantagem de selecionar as variáveis. Foram utlizados conjuntos de dados simulados compostos por 100 amostras e 500 marcadores SNPs avaliados em dois cenários que diferem entre si no vetor de coeficientes de regressão utilizado e quatro conjuntos de dados reais, sendo um de SNPs e três de dados NIR. Usou-se o software R para a modelagem dos dados. As amostras foram separadas em conjuntos de treinamento e de teste via algoritmo de Kennard e Stone. A qualidade preditiva do modelo foi avaliada com base no coeficiente médio de correlação (r) entre valores preditos e reais, e a raiz quadrada do erro quadrático médio (RMSE). No conjunto de dados simulados avaliado no primeiro cenário, havia 52 marcadores de maiores efeitos. Os modelos usando o BLASSO supervisionado, o SPLS e o PLS-OPS selecionaram, respectivamente, em média, 100, 310 e 124 variáveis. Em termos de capacidade preditiva os modelos após seleção foram semelhantes quando comparados ao uso dos dados completos. No segundo cenário, 10 marcadores de menor efeito foram escolhidos para serem significativos. Nesse cenário, para escolha do número de variáveis a serem selecionadas pelo BLASSO supervisionado utilizou-se dois critérios: no primeiro 20% das variáveis foram

selecionadas, e no segundo o número de variáveis selecionadas eram iguais ao do SPLS e do PLS-OPS. Em média os modelos apresentaram um desempenho melhor utlizando a seleção de variáveis em relação aos modelos construídos com os dados completos, sendo o SPLS levemente superior, com r = 0,846 e intervalo de RMSE de menor amplitude. Para a predição da produção de grãos em dados de SNPs, o método BLASSO supervisionado foi superior, com menor valor de RMSE (0,56) e maior valor de r (0,569). O PLS-OPS também apresentou bom desempenho nesse conjunto de dados, atestando o uso deste método para dados dessa natureza. No primeiro conjunto de dados NIR em que foi avaliado o teor de fibra da cana-de-açúcar, de maneira geral os valores de RMSE e de r se mantiveram próximos àqueles obtidos para os dados completos. No segundo conjunto de dados reais NIR em que foi avaliado o teor de lignina da cana-de-açúcar, pode-se observar que os melhores resultados foram obtidos com o método BLASSO supervisionado (RMSE = 0,705 e r = 0,956). No terceiro conjunto de dados reais NIR em que foram avaliadas amostras de repolho roxo, os melhores resultados foram obtidos quando utlizou-se o PLS-OPS (RMSE = 13,05 e r = 0,996). No segundo e terceiro conjuntos de dados NIR avaliados as estatísticas obtidas foram próximas às obtidas com os dados completos, porém com a vantagem de possuir menos variáveis. De maneira geral, os métodos funcionam de forma semelhante, mas cada um exibe vantagens sobre o outro em determinadas situações. Ao utilizarmos os métodos de seleção, podemos observar que os modelos se tornaram mais simples, visto que o número de variáveis reduziu significamente em todos os conjuntos de dados estudados.

ABSTRACT

FERREIRA, Roberta de Amorim, M.Sc., Universidade Federal de Viçosa, February, 2018**. Comparison of selection methods of regression variables applied to genomic data and NIR spectroscopy.** Adviser: Luiz Alexandre Peternelli. Co-adviser: Reinaldo Francisco Teófilo.

Researches from many different areas have data sets with the challenges of high dimensionality and multicollinearity still to be overcome, therefore specific methods for model fit must be employed. Although the existing methods are efficient to construct the model, it is often necessary to select the most important variables in explaining the model, once this practice can increase its predictive capacity, reduce costs, and analysis time. The main objective of this work was to evaluate and construct models using three methods of variable selection applied to single nucleotide polymorphisms (SNPs) and near infrared spectroscopy (NIR) data, besides evaluating the improvement in prediction quality, when compared to the use of complete data. The methods evaluated were: the selection of ordered predictors associated with partial least squares regression (PLS-OPS); the Sparce partial least square (SPLS); and the supervised Bayesian Lasso (BLASSO) – the last one is an adaptation of the BLASSO method with advantage of selecting variables. Were used simulated data sets composed of 100 samples and 500 SNP markers evaluated in two scenarios that differ from one another in the regression coefficient vector used, and four real data sets – composed by one set of SNPs and three sets of NIR data. It was used the software R in order to model the data. Samples were separated into training and test sets via Kennard and Stone algorithm. The predictive quality of the model was evaluated based on the mean correlation coefficient (r) between predicted and actual values, and the square root mean square error (RMSE). In the simulated data set evaluated in the first scenario, there were 52 markers of greater effects. The models using supervised BLASSO, SPLS and PLS-OPS selected an average of 100, 310 and 124 variables, respectively. In terms of predictive capacity, the models after selection were similar when compared to the use of the complete data. In the second scenario, 10 lower-effect markers were chosen to be significant. In this scenario, two criteria were used to select the number of variables to be selected by supervised BLASSO: in the first 20% of the variables were selected, and in the second, the number of variables selected were the same as SPLS and PLS-OPS. On average, the models presented a better performance when using the variables selection, than in relation to the models constructed with the complete data, once

the SPLS was slightly higher – with r = 0.846 and a lower amplitude RMSE interval. For the prediction of grain yield in SNP data, the supervised BLASSO method was superior, with a lower RMSE value (0.56) and a higher r value (0.569). PLS-OPS also performed well in this data set, attesting to the use of this method for data of this nature. In the first set of NIR data in which the sugar cane fiber content was evaluated, the RMSE and r values were, in general, close to those obtained for the complete data. In the second set of real NIR data in which the lignin content of sugarcane was evaluated, it can be observed that the best results were obtained with the supervised BLASSO method (RMSE = 0.705 and r = 0.956). In the third set of real NIR data in which samples of purple cabbage were evaluated, the best results were obtained when PLS-OPS (RMSE = 13.05 and r = 0.996) was used. In the second and third NIR data sets, the statistics obtained were close to those obtained with the complete data, but with the advantage of having fewer variables. In general, the methods used work in a similar way; however, each one of them has advantages over another in specific situations. By using the selection methods, it can be observed that the models have become simpler, once the number of variables reduced significantly in all datasets studied.

1. INTRODUÇÃO

Em geral todo experimento a ser estudado pode ser modelado por uma função matemática que contém as variáveis a serem medidas no mesmo. Quando os conjuntos de dados apresentam um grande volume de variáveis com poucas observações (alta dimensionalidade) correlacionadas (multicolinearidade) a serem analisadas os métodos tradicionais para construção de modelos para previsão não podem ser usados. Deve-se então empregar métodos específicos para o ajuste dos modelos. A crescente evolução computacional dos últimos anos proporcionou o desenvolvimento de métodos estatísticos multivariados para melhor análise e modelagem deste tipo de dados.

Embora os métodos multivariados existentes sejam, na maioria das vezes, eficientes para construção do modelo, frequentemente se faz necessário selecionar as variáveis mais importantes em explicar o modelo, visto que essa prática pode aumentar sua capacidade preditiva (FERREIRA, 2015).

Algumas áreas de pesquisa possuem conjunto de dados com os desafios da alta dimensionalidade e da multicolinearidade a serem superados, de forma que a prática de seleção de variáveis pode ser uma ferramenta útil. Podemos citar a associação genômica ampla (GWAS), que consiste em avaliar a associação entre dados de marcadores SNPs (*Single Nucleotide Polymorphisms*) a algum fenótipo de interesse, e a Quimiometria, uma disciplina da química, que consiste em analisar conjuntos de dados multivariados de origem química com métodos estatísticos. Uma maneira de obter informações multivariada em quimiometria é usando a espectroscopia no infravermelho próximo (NIR).

Para realizar a modelagem de dados SNPs, de los Campos et al. (2009) sugere alguns métodos estatísticos sob enfoque bayesiano como, por exemplo, o Lasso Bayesiano (BLASSO). Para dados NIR o método de regressão quadrados mínimos parciais (PLS) mostrou-se efetivo, isso porque é mais eficiente em lidar com ruídos experimentais (TEÓFILO, 2009).

O método BLASSO com seleção de variáveis (BLASSO supervisionado), além de modelar os dados funciona como um método de seleção de variáveis, visto que um limite de significância pode ser estabelecido de forma que os coeficientes de regressão abaixo desse limite foram descartados. Na quimiometria o método de seleção dos preditores ordenados (OPS) proposto por Teófilo et al. (2009), associado ao PLS, possui

como vantagem a capacidade de selecionar variáveis de conjuntos de dados (FERREIRA, 2015).

Além desses métodos podemos destacar o *Sparce partial least Square* (SPLS), proposto por Chun e Keles (2010), que pode ser aplicado a um grande volume de dados altamente correlacionados. O método possui a grande vantagem de, simultaneamente, reduzir a dimensionalidade dos dados e selecionar as variáveis de forma eficiente (CHUN; KELES, 2010).

Adicionalmente ao uso de dados reais na avaliação de diferentes modelos de predição, o uso de técnicas de simulação tem se mostrado uma ferramenta importante pois nos permite obter informações e realizar análises estatísticas de maneira mais econômica, rápida e controlada. Em especial, na área de associação genômica muitos estudos envolvendo dados simulados vem sendo executados com o objetivo de investigar a associação entre SNP e fenótipo (WALDMANN et al., 2013; MARIGORTA; GIBSON, 2014; OLIVEIRA, 2015a; PIRES, 2015).

Esse trabalho teve como objetivos avaliar e construir modelos empregando três métodos de seleção de variáveis (BLASSO supervisionado, SPLS e PLS-OPS) aplicados a dados de marcadores SNPs (*Single Nucleotide Polymorphisms*), reais e simulados, e a dados reais de espectroscopia no infravermelho próximo (NIR), além de avaliar a melhoria na qualidade de predição, quando comparado ao uso dos dados completos.

2. REVISÃO BIBLIOGRÁFICA

2.1 Métodos Estatísticos

Ao usar uma função matemática ou estatística para modelar as variáveis associadas a um experimento o pesquisador deve se preocupar pela correta escolha do modelo que atenda às peculiaridades do referido experimento. Quando o objetivo é determinar um modelo que explica o relacionamento entre variáveis ou fazer previsões, deve-se optar por um método que busca encontrar um modelo que correlacione a matriz **X** (matriz de variáveis explicativas) com o vetor **y** (variável resposta). Para tal, uma análise de regressão pode ser útil.

Na análise de regressão podemos ter modelos simples ou múltiplos que podem ser modelados por uma função. Realizar um experimento pelo processo de regressão múltipla geralmente é mais vantajoso do que pelo de regressão simples aplicado isoladamente a cada variável explicativa, pois além de permitir o estudo de cada variável isoladamente, também pode-se estudar a interação entre elas tendo, assim, um ajuste mais eficiente e consequentemente um modelo estatístico melhor para previsão.

Para obtenção do melhor modelo, devemos buscar um equilíbrio entre o modelo a ser escolhido e sua capacidade preditiva (FERREIRA, 2015). Caso tenhamos um modelo que não descreve bem o comportamento dos dados, ou um com sobre ajuste aos dados, perdemos a capacidade desse modelo em prever valores futuros.

A equação a seguir é um exemplo de um modelo de regressão linear múltipla:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1,2,\cdots,N$$

em que:

x_{ij}, é o valor da variável independente x_j na i-ésima observação, j=1,...,p;

Os parâmetros β_k, $k = 0,1,2...,p$ são os coeficientes da regressão; e,

O y_i é a variável dependente na i-ésima observação.

Para ajustarmos o modelo de uma regressão linear múltipla, devemos inicialmente estimar os valores dos coeficientes de regressão. Para isso, podemos escrever a equação acima em notação matricial, conforme mostrado abaixo, em que os dados foram dispostos em uma matriz **X** com N linhas e p+1 colunas.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{\mathrm{Np}} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \tag{1}$$

em que:

Y é um vetor N x 1 cujo os elementos correspondem as N observações;

X é uma matriz de dimensão N × (p+1), denominada matriz de incidência;

Ɛ é um vetor de dimensão N × 1 cujos elementos correspondem aos erros; e,

β é um vetor (p+1) × 1 cujos componentes são os coeficientes de regressão.

Quando o número de linhas da matriz **X** for maior ou igual ao número de colunas (N > p ou N = p), podemos encontrar os coeficientes de regressão via método quadrados mínimos, é importante ressaltar que para que o método seja válido as colunas da matriz **X** devem ser linearmente independentes (LI), ou seja, não há colinearidade. Portanto, podemos encontrar os coeficientes de regressão da equação (1) da seguinte maneira:

$$\mathrm{L} = \sum_{\mathrm{i}=1}^{\mathrm{p}} \varepsilon_{\mathrm{i}}^2 = \boldsymbol{\varepsilon^{\mathrm{t}}\varepsilon} = (\mathbf{Y} - \mathbf{X\beta})^{\mathbf{t}}(\mathbf{Y} - \mathbf{X\beta}) =$$

$$= \mathbf{Y^tY} - \mathbf{Y^tX\beta} - \mathbf{\beta^tX^tY} + \mathbf{\beta^tX^tX\beta} = \mathbf{Y^tY} - \mathbf{2\beta^tX^tY} + \mathbf{\beta^tX^tX\beta}$$

$$\frac{\partial \mathrm{L}}{\partial \beta} = \mathbf{-2X^tY} + \mathbf{2X^tX\beta}$$

Temos que $\mathbf{Y^tX\beta} = \mathbf{\beta^tX^tY}$, pois resulta em um escalar. Fazendo $\frac{\partial \mathrm{L}}{\partial \beta} = 0$, teremos:

$$(\mathbf{X^tX})\hat{\boldsymbol{\beta}} = \mathbf{X^tY}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X^tX})^{\mathbf{-1}}\mathbf{X^tY} \tag{2}$$

Para garantirmos que $\hat{\beta}$ seja mínimo, devemos ter $\frac{\partial^2 L}{\partial \beta^2} > 0$. No entanto, a soma de quadrados é sempre positiva. Assim, o modelo ajustado e o vetor de resíduos são, respectivamente:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$$

Além disso, podemos ter o caso em que o número de linhas é menor que o número de colunas (N<p), ocasionando sérios problemas de alta dimensionalidade. As estimativas dos coeficientes de regressão não podem ser obtidas via quadrados mínimos, métodos específicos para ajuste dos modelos devem ser aplicados.

A alta dimensionalidade é uma fonte de multicolinearidade. Esta pode ser definida como a presença de um elevado grau de correlação entre as variáveis independentes (FREUND; WILSON; SA, 2006), ou seja, as variáveis explicativas do modelo são linearmente dependentes.

Estes desafios da alta dimensionalidade e da multicolinearidade são encontrados, por exemplo, em dados provenientes da espectroscopia NIR dentro da área de Quimiometria, e em dados de marcadores SNPs dentro da Associação Genômica, que serão brevemente discutidos a seguir.

2.2 Quimiometria

O grande avanço computacional dos últimos anos proporcionou o desenvolvimento da Quimiometria, uma disciplina da química, que consiste em analisar conjuntos de dados multivariados nas variáveis explicativas de origem química com métodos estatísticos. A Quimiometria hoje se encontra muito difundida nas indústrias química, farmacêutica e de alimentos para controle de qualidade e ultimamente tem se destacado também na área médica para tratamento de imagens e na busca de marcadores de doenças (FERREIRA, 2015).

Diante desse desenvolvimento, a espectroscopia NIR somada a métodos estatísticos multivariados, como a regressão PLS, vêm sendo utilizada para fazer previsões em substituição aos métodos laboratoriais de alto custo. Sua utilização é simples, rápida, exata e não gera resíduos no ambiente (VALDERRAMA et al., 2007; MORGANO et al., 2008).

2.2.1 Espectroscopia no Infravermelho Próximo (NIR)

O espectro é a informação que obtemos quando utilizamos instrumentos de espectroscopia.

O espectro do infravermelho é obtido quando comprimentos de onda na região do infravermelho incidem sobre os átomos de uma certa molécula. Parte da radiação é absorvida e parte é transmitida. A radiação transmitida é lida para gerar o espectro.

Nesse trabalho, falaremos especialmente sobre a espectroscopia NIR, que com o avanço da Quimiometria dos últimos anos tem sido bastante utilizada.

O espectro obtido no NIR abrange uma região de radiação que varia de 12800 a 4000 cm^{-1} em número de onda, ou de 700 a 2500 nm em comprimento de onda (BOKOBZA, 1998).

Podemos encontrar na literatura muitas áreas que têm utilizado o NIR para previsões indiretas, como por exemplo na agronomia para determinação de proteína em café cru (MORGANO et al., 2008), na área florestal para avaliação das propriedades energéticas de resíduos de madeiras tropicais (SILVA et al., 2014), na área farmacêutica para controle de qualidade de medicamentos contendo diclofenaco de potássio (SOUZA; FERRÃO, 2006), dentre outras áreas.

Os dados provenientes do NIR são organizados em uma matriz **X** em que cada linha se refere a uma amostra (espectro) e cada coluna aos números de ondas (variáveis) conforme mostrado na Figura 1. Podemos observar que apenas uma única amostra origina muitas variáveis. Em geral, o número de comprimentos de onda lidos no NIR é muito superior ao de amostras coletadas.

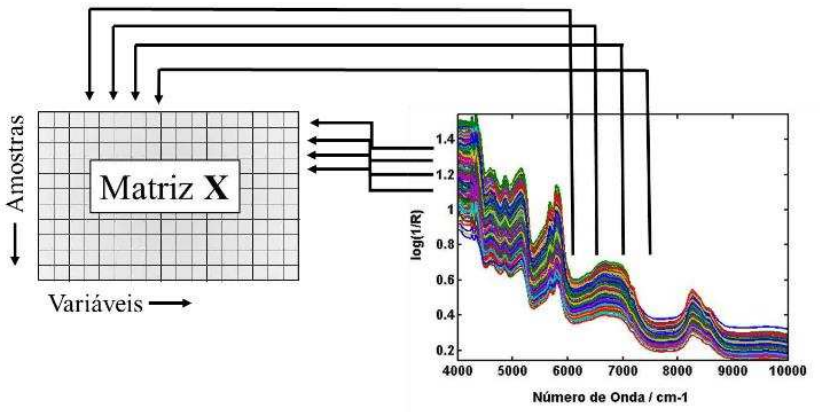

Figura 1- Organização dos dados NIR, FONTE: (ROQUE, 2015)

O vetor **y** contém valores de alguma propriedade específica das amostras, ou seja, seus valores são obtidos por algum método de referência.

Além de o número de colunas em geral ser superior ao número de linhas, podemos observar que os dados provenientes do NIR são altamente correlacionados, ocasionando problemas de multicolinearidade (FREUND; WILSON; SA, 2006). O método dos quadrados mínimos não pode ser aplicado nessa situação. Devemos, então, empregar métodos específicos de compressão de dados para encontrarmos o modelo ajustado (TEÓFILO, 2007).

Para análise e modelagem de dados de espectroscopia NIR, deve-se realizar algumas fases como: pré-tratamentos e calibração multivariada. Essas fases serão abordadas a seguir.

Pré-tratamentos

Geralmente, é necessário realizar pré-tratamentos na matriz de dados antes da construção do modelo estatístico, pois estes minimizam os erros sistemáticos presentes nos dados, tornando a matriz de dados mais adequada para análise (XU et al., 2008, SOUZA; POPPI, 2012). Pré-tratamentos quando aplicado às linhas da matriz **X** são chamados de transformações, e quando aplicado nas colunas são chamados pré-processamento.

Comumente, aplicam-se aos espectros diferentes pré-tratamentos (centragem na média, normalização, alisamento, derivação e correção multiplicativa de sinal (MSC)) (SOUZA; POPPI, 2012; FERREIRA, 2015). Após os tratamentos decide-se quais serão os mais adequados para o conjunto de dados em estudo através de estatísticas que inferem sobre erros de predição.

Mais detalhes e esclarecimentos sobre pré-tratamentos podem ser obtidos em Ferreira (2015).

Regressão multivariada

Regressão multivariada é um método que busca encontrar um modelo matemático que correlacione a matriz **X** com o vetor **y** de referência, objetivando um modelo de previsão satisfatório. No caso de dados NIR, visa buscar uma relação funcional entre os comprimentos de ondas obtidos a partir de cada amostra analisada e a propriedade relevante ao estudo.

Para a construção do modelo, podemos utilizar alguns métodos de regressão multivariada, como: Regressão por componentes principais (PCR) ou Regressão PLS (FERREIRA, 2002). Ambos os métodos visam reduzir a dimensionalidade dos dados. A principal diferença entre eles está na forma como as variáveis latentes (MARÔCO, 2010) são construídas.

Para dados espectrais o método de regressão multivariada PLS mostrou-se mais eficientes do que o PCR, isso porque o PLS é mais eficiente em lidar com ruídos experimentais (TEÓFILO, 2009).

2.3 SNPs - *Single Nucleotide Polymorphisms*

O grande avanço tecnológico dos últimos anos permitiu o desenvolvimento da genética molecular, principalmente no uso de marcadores genéticos moleculares para sequenciamento e genotipagem do DNA. O uso de marcadores moleculares possibilita a identificação dos melhores indivíduos antes mesmo da manifestação do fenótipo, ou seja, antes mesmo que uma determinada característica se expresse.

Atualmente, os marcadores de maior destaque são os do tipo SNPs (VIGNAL, et al., 2002).

Os marcadores SNPs são obtidos de metodologias modernas e de baixo custo (CAETANO, 2009**)**. Essa tecnologia, tem contribuído significativamente para o desenvolvimento da genética e melhoramento animal e vegetal (CAI et al., 2013; CRUZ et al., 2013), detecção de doenças genéticas (FRAZER et al., 2009), investigação de paternidade (JOBIN et al., 2008), dentre outras diversas áreas de estudo.

Os SNPs são os marcadores de maior destaque pois possuem baixa taxa de mutação, ou seja, a probabilidade de que uma cópia de um alelo mude é baixa, são codominantes, além de serem as variações mais abundantes nos genomas (VIGNAL, et al., 2002).

Os marcadores SNPs surgem quando acontecem mutações nas bases nitrogenadas: Adenina (A), Citosina (C), Timina (T) e Guanina (G). Na Figura 2, por exemplo, podemos observar que o indivíduo 1 possui inicialmente a mesma sequência nas fitas de DNA, até que ocorre uma alteração de bases, onde deveria ter TT, aparece TC, essa alteração é o que chamamos de SNP.

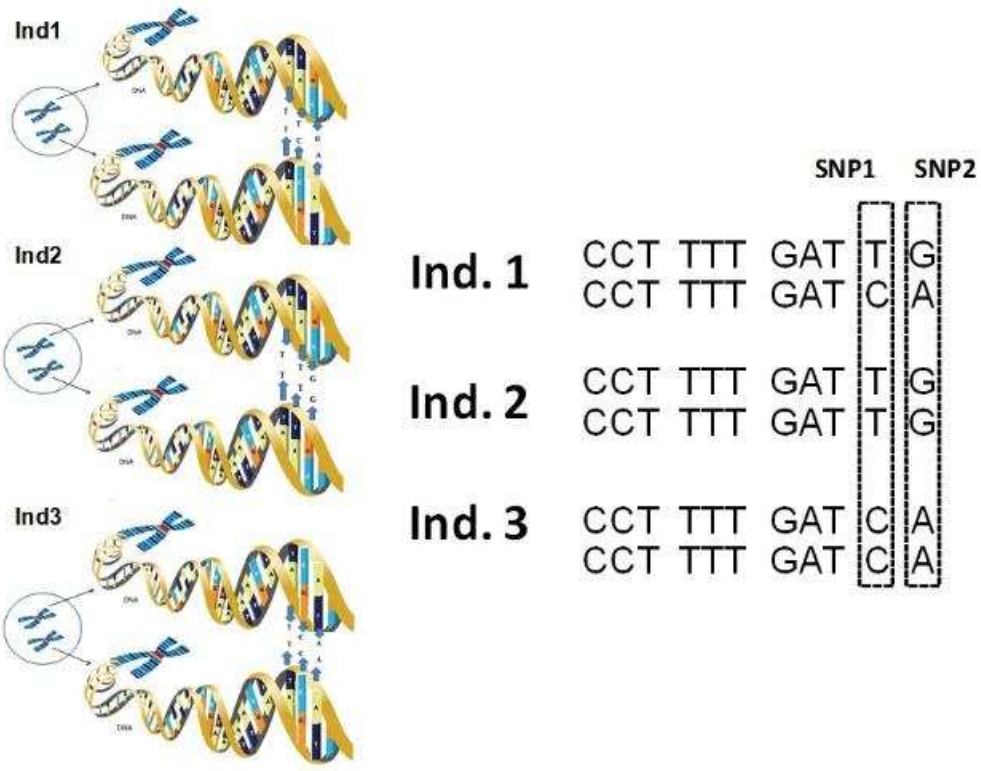

Figura 2-Exemplo de dois SNPs em uma amostra de três indivíduos. FONTE: (OLIVEIRA, 2015a).

Estrutura e organização dos dados

Frequentemente, os SNPs são bialélicos, isto é, o par de bases nitrogenadas (AT) e (GC) que são complementares, podem ser denominadas de alelos **A** e **a** respectivamente (BROOKES, 1999).

Portanto, para analisarmos os dados estatisticamente devemos descodifica-los. Assim, por exemplo: AA = 0, Aa = 1 e aa = 2, ou outra codificação equivalente. Os dados descodificados provenientes de SNPs são organizados em uma matriz **X** em que cada linha se refere ao indivíduo genotipado e cada coluna aos marcadores SNPs (variáveis).

A Figura 3 exemplifica uma amostra com 4 indivíduos (suínos) e 5 variáveis (SNPs). Temos 3 cenários representados. O primeiro nos informa a base de dados inicial original representada por bases nitrogenadas. Podemos observar que no SNP1 por exemplo, os suínos 1 e 2 apresentam os genótipos GG; para o suíno 3, os genótipos GT; e para o suíno 4, os genótipos TT . O segundo cenário é a base de dados inicial, porém com codificação alélica. Podemos observar que o G e o T da codificação apresentada no primeiro cenário foram recodificados, respectivamente, pelos alelos **a** e **A.** Portanto, onde tínhamos GG passou a ser **aa**; onde era GT passou a ser **aA** (ou, equivalentemente, **Aa**) e onde era TT passou a ser **AA**. No último cenário temos a codificação numérica. Na coluna

do SNP1 onde tínhamos **AA**, **Aa** e **aa** temos, agora, 0, 1 e 2, respectivamente. Nota-se que em todos os cenários o valor do fenótipo é sempre o mesmo.

Cenário 1: Base de dados inicial original

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Suíno 1	GG	AA	GG	CT	TT	60,94
Suíno 2	GG	AG	GG	CT	TT	70,45
Suíno 3	GT	AG	AG	TT	TT	100,34
Suíno 4	TT	GG	AA	TT	CC	500,65

Cenário 2: Base de dados inicial com codificação alélica

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Suíno 1	aa	AA	AA	Aa	AA	60,94
Suíno 2	aa	Aa	AA	Aa	AA	70,45
Suíno 3	Aa	Aa	Aa	aa	AA	100,34
Suíno 4	AA	aa	aa	aa	aa	500,65

Cenário 2: Base de dados inicial com codificação alélica

	SNP1	SNP2	SNP3	SNP4	SNP5	Fenótipo
Suíno 1	2	0	0	1	0	60,94
Suíno 2	2	1	0	1	0	70,45
Suíno 3	1	1	1	2	0	100,34
Suíno 4	0	2	2	2	2	500,65

Figura 3-Exemplo hipotético do processo de codificação. Adaptado de Oliveira (2015a).

Podemos observar que um único indivíduo origina muitos SNPs, de modo que o número de marcadores é superior ao número de indivíduos genotipados (alta dimensionalidade). Além do número de colunas ser superior ao número de linhas, podemos notar que os dados provenientes de SNPs são altamente correlacionados, visto que diferentes SNPs apresentam o mesmo perfil genotípico, ocasionando problemas de multicolinearidade. Observa-se que os SNPs mais próximos são muitas vezes mais correlacionados entre si (FENG et al., 2012).

Além da matriz de SNPs, temos o vetor **y** que contém valores das observações fenotípicas de interesse.

Uma vez obtidos os modelos de previsão, podemos selecionar os indivíduos mais produtivos logo após o nascimento, visto que por meio dos marcadores são identificados

os alelos que estão relacionados a uma determinada característica de interesse, acelerando o processo do melhoramento genético (RESENDE et al., 2012).

Para realizar a modelagem de dados de SNPs, Azevedo et al. (2013) sugerem alguns métodos estatísticos como o Ridge Regression (RR-BLUP), Genomic Best Linear Unbiased Predictor (G-BLUP) e o BLASSO versão Bayesiana da regressão LASSO (Least Absolute Shrinkage and Selection Operator – TIBSHIRANI, 1996). Além desses, sugerem o método de regressão via PLS e o via PCR que visam reduzir a dimensionalidade dos dados.

Embora os métodos existentes sejam eficientes para construção do modelo, frequentemente se faz necessário selecionar as variáveis mais importantes em explicar o modelo, visto que essa prática pode aumentar sua a capacidade preditiva (FERREIRA, 2015), diminuir custos e tempo das análises.

2.4 Seleção de variáveis

Realizar seleção de variáveis significa encontrar um subconjunto da matriz **X** que melhor se correlaciona com o vetor resposta **y**, sem que informações relevantes sejam eliminadas. Segundo Zimmer e Anzanello (2014) engenheiros e pesquisadores tem buscado métodos para selecionar o melhor subconjunto de variáveis, visando a diminuição de custos e o aumento da precisão dos resultados.

Existem diversos métodos de seleção para diferentes finalidades. Na GWAS ao realizarmos a seleção de variáveis, podemos identificar quais SNPs são responsáveis pela variação da característica fenotípica de interesse. O método BLASSO (DE LOS CAMPOS et al., 2009) faz alguns coeficientes tenderem a valores próximos a zero (RESENDE et al., 2011). Para que funcione como método de seleção de variáveis um limite de significância pode ser estabelecido de forma que os coeficientes de regressão abaixo desse limite sejam eliminados. Neste trabalho este procedimento será denominado BLASSO supervisionado.

O método SPLS, usado em diferentes conjuntos de dados (FENG et al., 2012; COLOMBANI et al., 2012; ABDEL-RAHMAN et al., 2014), foi desenvolvido com base no PLS com a vantagem de remover algumas variáveis não significativas. Portanto, funciona também como um método de seleção.

Realizar seleção de variável é importante não apenas em dados SNPs, mas de uma maneira geral quando temos dados com alta dimensionalidade. A indústria petrolífera no processo de refino do petróleo, a indústria alimentícia na produção de alimentos são

exemplos de áreas que precisam superar o desafio da alta dimensionalidade, além dessas, podemos citar de maneira geral áreas que envolvam dados químicos (GAUCHI; CHAGNON, 2001).

Na Quimiometria, o grande volume de dados gerados pelo NIR desperta o interesse em usar métodos de seleção de variáveis. Existem alguns métodos de seleção na Quimiometria, porém o mais conceituado e utilizado é o de algoritmos genéticos (FERREIRA, 2015). Alternativamente, o método OPS proposto por Teófilo et al. (2009), mostrou-se eficiente na seleção de variáveis de dados NIR (COSTA; LIMA, 2013; GUIMARÃES et al., 2016; ASSIS et al., 2017; CALIARI et al., 2017). Para construção do modelo com as variáveis selecionadas através do OPS o método PLS poderá ser utilizado.

Para realizar esse trabalho, foram escolhidos os métodos BLASSO supervisionado visto que não possui restrições quanto ao número de coeficientes de regressão, gerando modelos mais eficientes, o SPLS que se mostrou eficiente em lidar com dados altamente correlacionados (CHUN; KELES, 2010) e por fim o método OPS (TÉOFILO et al., 2009) associado ao PLS visto que tem sido eficiente na construção de modelos de NIR e até o momento não foi aplicado a dados de SNPs. Os métodos BLASSO supervisionado, SPLS e PLS-OPS foram escolhidos pelo fato de serem ou funcionarem como métodos de seleção de variáveis.

Esses métodos são meramente conceituais e engenhosos. Dessa maneira, as metodologias referentes ao BLASSO supervisionado, PLS-OPS e o SPLS serão apresentadas de forma mais simples a seguir.

2.5 Método Lasso bayesiano (BLASSO)

As estimativas dos parâmetros de interesse podem ser obtidas por meio de uma abordagem bayesiana ou uma abordagem frequentista. A principal diferença entre as duas é que na frequentista o parâmetro é desconhecido e fixo, enquanto na bayesiana o parâmetro é considerado uma variável aleatória. Dessa forma, a inferência bayesiana além de incluir a informação dos dados por meio da função de verossimilhança, inclui também o conhecimento *a priori* do pesquisador a respeito do parâmetro por meio da distribuição *a priori*.

O método BLASSO, versão bayesiana da regressão via LASSO (TIBSHIRANI, 1996) foi proposto na seleção genômica ampla por de los Campos et al. (2009). Em geral, o BLASSO é mais utilizado que o LASSO, pois não tem restrições quanto ao número de

coeficientes de regressão, além de ser mais estável quando se tem alta dimensionalidade (RESENDE, et al. 2012).

O modelo linear geral para predição dos efeitos é dado por:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

em que:

y é o vetor de observações fenotípicas (N×1), N é o número de indivíduos genotipados e fenotipados;

1 é o vetor coluna de 1's (N×1);

μ é a média da variável **y**;

X é a matriz de incidência (N×p);

β é o vetor que contém os coeficientes de regressão, de dimensão p×1;

e é o vetor de resíduos (N×1).

Segundo de los Campos et al. (2009) ao utilizarmos o BLASSO as seguintes distribuições são assumidas:

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

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

em que:

MNV refere-se à distribuição normal multivariada;

λ é o parâmetro de suavização;

σ^2 é o componente de variância que tem como distribuição a *priori* uma qui-quadrado invertida escalada.

Empregando uma formulação em termos de um modelo hierárquico aumentado, tem-se:

$$b_i|\tau \sim N(0, D\sigma^2)$$

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

em que $D = diag(\tau_1^2, \tau_2^2, \ldots, \tau_p^2)$ Segundo Park e Casella (2008), essa formulação conduz a uma distribuição exponencial dupla para os coeficientes, ou seja:

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

O parâmetro de suavização λ faz com que alguns efeitos sejam aproximadamente zero, mas não efetivamente zero.

2.5.1 Blasso supervisionado

Para que o método BLASSO funcione como um método de seleção de variáveis, dois critérios de seleção foram estabelecidos:

1º Critério: 80% das variáveis menos significativas são descartados e as restantes selecionadas.

2º Critério: o número de variáveis que o BLASSO seleciona será igual, respectivamente, ao número de variáveis que os métodos SPLS e PLS-OPS selecionarem.

2.6 Método de Seleção PLS-OPS

2.6.1 Quadrados Mínimos Parciais (PLS)

Diante da necessidade de lidar com dados fortemente correlacionados, Wold (1970) desenvolveu o método PLS (WOLD; SJÖSTRÖM; ERIKSSON, 2001). O PLS é um método fundamentado na compressão dos dados, ou seja, visa reduzir o espaço das medidas originais mantendo apenas as informações mais importantes, gerando assim alguns subespaços.

Como os dados NIR são altamente correlacionados, podemos observar que existem várias colunas com informações praticamente iguais a respeito da variância. O uso do PLS acarreta na redução dessas colunas similares para apenas uma, de forma a reduzirmos o conjunto de dados original sem perder muita informação do sistema completo de variáveis.

Para execução do PLS, utiliza-se o algoritmo bidiagonal, baseado na decomposição de valores singulares (SVD) que afirma que toda matriz pode ser escrita como:

$$\mathbf{y} \rightarrow \mathbf{X} = \mathbf{URV^t} \tag{5}$$

Denominou-se **UR** matriz de escores e **V** a matriz de loadings.

Conforme apresentado na equação (5) o PLS, além de levar em consideração informações presentes na matriz **X**, também considera as informações do vetor **y** para construção das variáveis latentes (WOLD; SJÖSTRÖM; ERIKSSON, 2001).

As colunas da matriz **U** e as linhas da matriz $\mathbf{V^t}$ criam os novos subespaços conhecidos como variáveis latentes, que possuem informações presentes na matriz de

dados **X** e no vetor **y.** Geralmente, as primeiras variáveis latentes (2 a 10), nos informam quase toda (aproximadamente 100%) informação da matriz de dados **X** original (ROQUE, 2015).

O algoritmo para execução do PLS é apresentado resumidamente a seguir (MARTINS; TEOFILO; FERREIRA, 2010):

1. Inicialize o algoritmo para primeira componente:

$$\mathbf{X} = \mathbf{y}\mathbf{v}_1^{\mathbf{t}}$$

$$\mathbf{X}^{\mathbf{t}} = \mathbf{v_1}\mathbf{y}^{\mathbf{t}}$$

$$\mathbf{X}^{\mathbf{t}}\mathbf{y} = \mathbf{v_1}\mathbf{y}^{\mathbf{t}}\mathbf{y}$$

$$\mathbf{X}^{\mathbf{t}}\mathbf{y}(\mathbf{y}^{\mathbf{t}}\mathbf{y})^{-\mathbf{1}} = \mathbf{v_1}(\mathbf{y}^{\mathbf{t}}\mathbf{y})(\mathbf{y}^{\mathbf{t}}\mathbf{y})^{-\mathbf{1}}$$

$$\mathbf{v_1} = \mathbf{X}^{\mathbf{t}}\mathbf{y}(\mathbf{y}^{\mathbf{t}}\mathbf{y})^{-\mathbf{1}} \quad (6)$$

Para normalizarmos o vetor, basta dividi-lo por sua norma. Usando a norma euclidiana, ou seja, $\|v_1\| = \sqrt{v_1^t v_1}$. O vetor normalizado de (6), será:

$$\mathbf{v_1} = \frac{\mathbf{X^t y}}{\|\mathbf{X^t y}\|} \; ; \; \alpha_1 \mu_{\mathbf{1}} = \mathbf{X v_1}$$

2. Para i = 2, ..., h componentes:

2.1 $y_{i-1}\mathbf{v_1} = \mathbf{X^t}\mu_{\mathbf{i-1}} - \alpha_{i-1}\mathbf{v_{i-1}}$

2.2 $\alpha_i \mu_{\mathbf{1}} = \mathbf{X v_i} - y_{i-1}\mu_{\mathbf{i-1}}$

$$\mathbf{V_h} = (\mathbf{v_1}, \dots, \mathbf{v_h}) \quad \mathbf{U_h} = (\mu_{\mathbf{1}}, \dots, \mu_{\mathbf{i}}) \quad \text{e } \mathbf{R_h} = \begin{pmatrix} \alpha_1 & y_1 & & \\ & \ddots & & \\ & & \alpha_{k-1} & y_{k-1} \\ & & & \alpha_k \end{pmatrix}$$

Sendo h o número de variáveis latentes escolhido, prova-se que $\mathbf{U_h R_h} = \mathbf{X V_h}$ e então $\mathbf{R_{h=} U_h^t X V_h}$.

Portanto, calculadas as matrizes **U**, **V** e **S** podemos estimar a pseudo-inversa de Moore-Penrose de **X**, e estimar o modelo da seguinte maneira:

$$\mathbf{y} = \mathbf{X\beta} \rightarrow \mathbf{X} = \mathbf{U_h R_h V_h^t} \rightarrow \mathbf{y} = \mathbf{U_h R_h V_h^t \beta} \rightarrow \widehat{\boldsymbol{\beta}} = \mathbf{V_h R_h^{-1} U_h^t y}$$

A escolha do número de variáveis latentes (nVL) é uma etapa muito importante para a construção do modelo. Deve-se buscar um equilíbrio na quantidade escolhida pois,

caso contrário, podemos ter um sobre ajuste do modelo aos dados ou informações importantes podem ficar de fora.

O método da validação cruzada geralmente é empregado para a escolha do nVL. Para execução desse método, inicialmente um gráfico é construído. No eixo das abscissas temos os nVL (geralmente de 0 a 20), e no eixo das ordenadas temos a raiz quadrada do erro quadrático médio de validação cruzada (RMSECV). O ponto que possuir menor valor de RMSECV será escolhido (MARTENS; NAES, 1996) e o nVL associado a ele será o melhor para construção do modelo. Podemos calcular o RMSECV, conforme a equação:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

Em que *N* é o número de amostras da validação, y_i é o i-ésimo valor referente ao vetor de referência **y**, o $\hat{y}_i$ é o *i*-ésimo valor previsto pelo modelo.

2.6.2 Método de Seleção dos Preditores ordenados (OPS)

O método OPS é composto por 5 etapas. Na primeira etapa devemos fazer a escolha de um vetor informativo. A segunda etapa consiste em observar as regiões onde há maior intensidade do sinal desse vetor. Quanto maior o sinal, mais importante é a variável. Na terceira etapa é feita uma ordenação das variáveis originais conforme a intensidade do sinal do vetor informativo e, então, na quarta etapa as variáveis são analisadas por métodos de validação cruzada, como por o exemplo o PLS. Um subconjunto inicial de variáveis (janela) é selecionado para construir e avaliar o primeiro modelo. Então, esta matriz é ampliada pela adição de um número fixo de variáveis (incremento) e um novo modelo é construído e avaliado. Novos incrementos são adicionados até que todas variáveis sejam analisadas. A cada subconjunto de variáveis analisadas calcula-se os parâmetros estatísticos de interesse. Finalmente, na quinta etapa faz-se a escolha do melhor subconjunto de variáveis selecionadas, e observa-se aqueles que geraram menores erros com a melhor previsão usando validação cruzada. Na Figura 4, um esquema do método OPS (TEÓFILO et al., 2009).

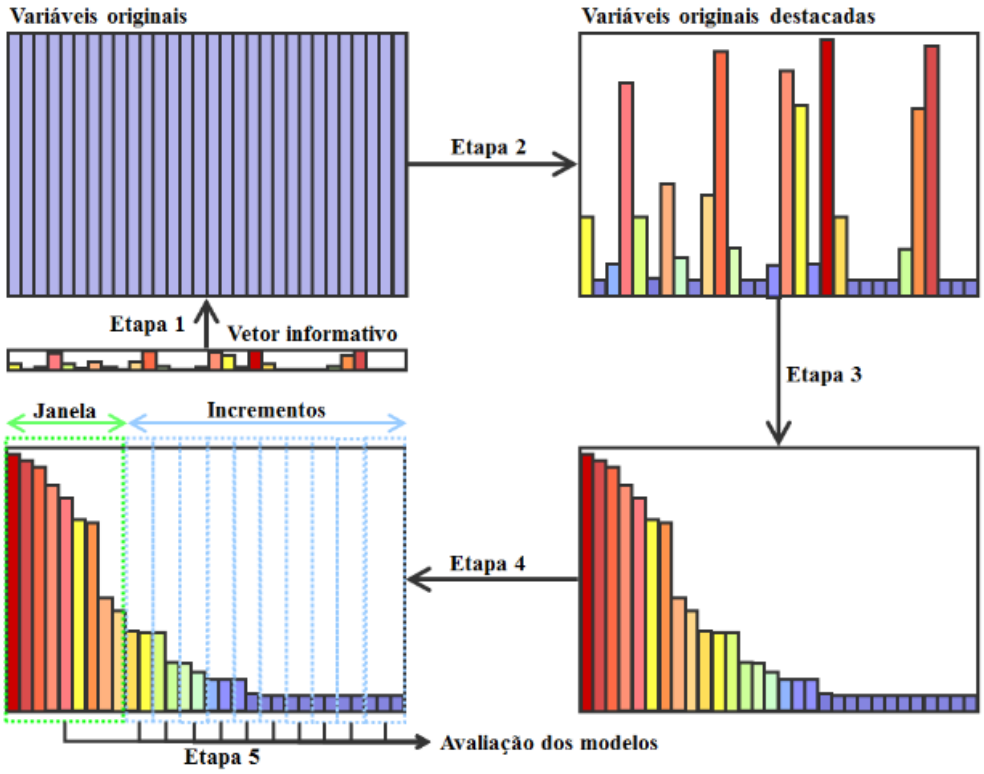

Figura 4- Etapas da seleção de variáveis usando o OPS. Fonte: (TEÓFILO et al., 2009).

Para realizar a modelagem das variáveis selecionados pelo método OPS podemos utilizar o método de regressão via PLS.

2.7 Método *Sparce partial least Square* (SPLS)

O método SPLS proposto por Chun e Keles (2010) é uma adaptação da regressão PLS que possibilita de maneira eficiente reduzir a dimensionalidade e selecionar as variáveis simultaneamente.

O método PLS pode fornecer um modelo com bom desempenho preditivo, porém quando o número de variáveis é maior do que o número de amostras/indivíduos a propriedade de consistência do estimador PLS pode não ser mantida e o seu desempenho decresce. O SPLS surge de maneira a complementar o PLS introduzindo a seleção de variáveis, removendo assim as variáveis insignificantes (FENG et al., 2012). Além disso, o SPLS também é bastante vantajoso quando há alta colinearidade entre as variáveis.

Considere $\mathbf{W}$ a matriz de loadings constituídas por K vetores ($\mathbf{w_1},\ldots,\mathbf{w_K}$), em que K é o número de variáveis latentes. Para cada $\mathbf{w_k}$ (k = 1 a K) temos (FENG et al., 2012):

$$\mathbf{w_k} = \text{argmax}_{\mathbf{w}}\{\mathbf{w^t X^t Y Y^t X w}\} \quad (7)$$

De modo que $\mathbf{w^t w} = 1$ e $|\boldsymbol{w}| \leq \lambda$ para k = 1,....K, onde λ é o parâmetro de ajustamento que determina o grau de redução de variáveis explicativas no modelo, **X** é a matriz de incidência e **Y** é o vetor resposta.

O método SPLS impõe uma restrição em **w** para obter uma solução com menos variáveis explicativas. Portanto, a função (7) é reformulada impondo uma penalidade, na direção de um vetor **c**, em vez do vetor de direção original **w**. A função pode ser escrita como:

$$\min_{\mathbf{w,c}}\{-\mathrm{k}\mathbf{w}^{\mathrm{t}}\mathbf{M}\mathbf{w} + (\mathbf{1} - \mathrm{k})(\mathrm{c} - \mathbf{w})^{\mathrm{t}}\mathbf{M}(\mathbf{c} - \mathbf{w}) + \lambda_1|\mathbf{c}|_1 + \lambda_2|\mathbf{c}|_2\}$$

Em que λ_1 e λ_2 são fatores de penalidade e $\mathbf{w^T w} = 1$ e $\mathbf{M} = \mathbf{X^t Y Y^t X}$. No algoritmo SPLS, uma estimativa de **w** pode ser obtida com a seguinte equação:

$$\tilde{\mathbf{w}} = \left(|\widehat{\mathbf{w}}| - \eta \max_{1\leq \mathrm{i} \leq \mathrm{p}}|\widehat{\mathbf{w}}_{\mathbf{i}}|\right) \mathrm{I}_{(|\widehat{\mathbf{w}}| \geq \mathrm{n}\max_{1\leq \mathrm{i}\leq \mathrm{p}}|\widehat{\mathbf{w}}_{\mathrm{i}}|)} \mathrm{sign}(\widehat{\mathbf{w}}), \text{ com } 0 \leq \eta \leq 1.$$

Para construção dos modelos, bem como sua validação, em dados de NIR e de SNPs, o conjunto das amostras da matriz **X** devem ser separadas em dois subconjuntos: um subconjunto de calibração (ou de treinamento) e outro subconjunto de previsão (ou de teste), sendo um complementar do outro. Esse procedimento pode ser feito usando o algoritmo KS (KENNARD; STONE, 1969), que será resumidamente discutido a seguir.

2.8 Algoritmo Kennard e Stone (KS)

O algoritmo Kennard-Stone foi proposto por R.W. Kennard e L.A. Stone (1969). Esse algoritmo pode ser utilizado para dividir o conjunto de dados original em dois subconjuntos (treinamento e teste), de maneira a selecionar amostras que represente o máximo da variabilidade do conjunto original (SOUZA et al., 2011).

Para a seleção das amostras que formarão o subconjunto de treinamento, o algoritmo utiliza a distância euclidiana para cada par (p, q) de amostras, a partir da seguinte equação:

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

Ao empregar a distância euclidiana as amostras que estão mais distantes entre si são selecionadas e com isso temos uma distribuição mais uniforme do subconjunto de amostras (SOUZA et al., 2011).

A Figura 5 apresenta um exemplo hipotético de um conjunto formado por 6 amostras em que 3 delas serão selecionadas para formar o subconjunto de treinamento. O comprimento da linha entre os números indica as respectivas distâncias euclidianas. O algoritmo inicia selecionando o par de amostras mais distantes (1 e 2); em sequência, seleciona a amostra mais distante (4) a partir das selecionadas anteriormente. Portanto o subconjunto de treinamento será formado pelas amostras 1, 2 e 4.

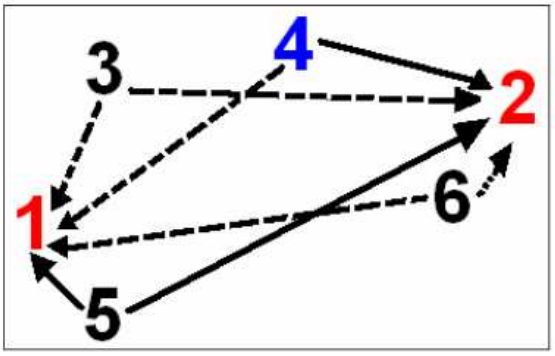

Figura 5-Exemplo de seleção de amostras pelo algoritmo de Kennard-Stone. FONTE: (DANTAS, 2007).

2.9 Construção do modelo e sua validação em dados NIR

Incialmente, deve-se aplicar as transformações e os pré-processamentos no conjunto de calibração, empregar um método de redução da dimensionalidade, escolher o nVL a partir da validação cruzada e por fim construir o modelo de calibração. Após, deve-se validar o modelo, ou seja, verificar sua capacidade preditiva. Para tal, no conjunto de previsão deve-se aplicar novamente os mesmos pré-tratamentos utilizados no conjunto de calibração e aplicar o modelo construído a esses dados, obtendo, então, o valor predito ($\hat{y}_p$).

Por fim, é realizado a validação externa, em que é verificado a eficiência do modelo de calibração construído e sua capacidade preditiva através de estatísticas que inferem sobre o erro de previsão.

Essas etapas descritas, podem ser vistas no esquema apresentado na figura 6:

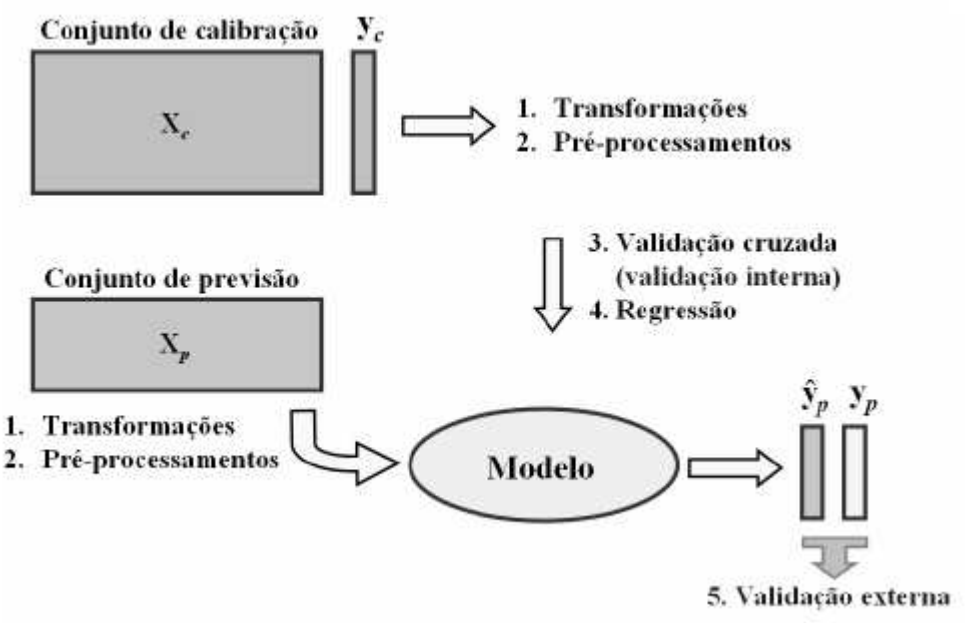

Figura 6 - Esquema usado para construir e validar o modelo (FONTE: TEÓFILO, 2007).

2.10 Construção do modelo e sua validação em dados de SNPs

Após a obtenção da matriz de marcadores, os seus efeitos são estimados e modelos são construídos para predição dos valores genéticos genômicos (VGG).

Os VGG, são obtidos através dos modelos estatísticos construídos. Estes valores nos informam o potencial genético dos indivíduos e podem ser calculados a partir da seguinte expressão (TEIXEIRA, 2015):

$$\mathrm{VGG} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Sendo **X** a matriz de incidência e $\hat{\boldsymbol{\beta}}$ vetor com as estimativas dos efeitos dos SNPs.

Para a construção do modelo bem como sua validação, as linhas da matriz **X** e as correspondentes linhas do vetor **y** de fenótipos devem ser separados em dois subconjuntos: um de treinamento e um outro de teste. Os subconjuntos de treinamento de **X** e **y** são utilizados para estimar os efeitos dos marcadores, ou seja, utilizados na construção dos modelos para predição dos VGG. Podemos observar que nesse conjunto a maior parte dos indivíduos são avaliados (RESENDE et al., 2012). Em geral 80% dos indivíduos são alocados ao subconjunto de treinamento.

Para validar o modelo construído, ou seja, verificar sua capacidade preditiva, aplica-se no modelo construído os dados do subconjunto de teste de **X**, obtendo então os valores preditos ($\hat{y}_p$).

Por fim, é verificado a eficiência do modelo construído e sua capacidade preditiva comparando-se o valor predito ($\hat{y}_p$) com o y real pertencente ao subconjunto de teste.

2.11 Comparação das metodologias

A eficiência dos modelos construídos pode ser verificada utilizando os seguintes parâmetros estatísticos: coeficiente de correlação (r), que mede o grau de associação entre as variáveis, onde $-1 \leq r \leq 1$; e a raiz quadrada do erro quadrático médio (RMSE) que mede as diferenças individuais entre os valores previstos pelo modelo ($\hat{y}_i$) e os observados (y_i) (FERREIRA, 2011).

O ideal seria menores valores de RMSE (FERREIRA, 2015) e maiores valores de coeficiente de correlação.

Esses parâmetros estatísticos podem ser calculados da seguinte maneira:

Coeficiente de correlação:

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

em que:

$\hat{y}$ e $\bar{\hat{y}}$ o valor estimado e médio estimado; e,

y e $\bar{y}$ os valores observados e valores médios observados.

Raiz quadrada do erro quadrático médio:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

em que *N* é o número de amostras pertencente a cada subconjunto (teste ou treinamento) em que está analisando.

3. MATERIAL E MÉTODOS

3.1 Obtenção de dados sintéticos

Foram simulados conjuntos de dados de marcadores SNPs com N indivíduos independentes entre si e p SNPs (SNP_1, SNP_2, ..., SNP_p). Para cada simulação, além da matriz **X** de marcadores, foi obtido o vetor **y** de observações fenotípicas.

As linhas da matriz **X** de marcadores de SNPs são formadas por genótipos de cada indivíduo. Os genótipos são gerados a partir da combinação de dois haplótipos independentes simulados. Sabemos que a maioria dos SNPs são bialélicos, portanto podem ser representados por 0 e 1. Essa situação pode ser exemplificada na Tabela 1, em que foi construído o vetor de genótipos de um indivíduo a partir dos vetores de haplótipos, considerando 10 marcadores SNPs.

Tabela 1- Exemplo hipotético do genótipo de um indivíduo.

	SNP1	**SNP2**	**SNP3**	**SNP4**	**SNP5**	**SNP6**	**SNP7**	**SNP8**	**SNP9**	**SNP10**
Vetores de	0	1	1	1	1	0	0	1	0	1
Haplótipos	1	0	0	1	1	1	0	1	0	0
Genótipo	**0/1**	**1/0**	**1/0**	**1/1**	**1/1**	**0/1**	**0/0**	**1/1**	**0/0**	**1/0**

Para a construção da matriz **X** foi usado a seguinte codificação:

$$\mathbf{X_i} = \begin{cases} -1 \text{ se o i} - \text{ésimo SNP tem genótipo } 0/0, \\ 0 \text{ se o i} - \text{ésimo SNP tem genótipo } 0/1 \text{ ou } 1/0, \\ 1 \text{ se o i} - \text{ésimo SNP tem genótipo } 1/1 \end{cases}$$

Portanto, conforme o exemplo proposto a primeira linha da matriz **X** será formada por:

$$[0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad -1 \quad 1 \quad -1 \quad 0]$$

A simulação foi baseada na proposição apresentado em Feng et al. (2012). Os dados foram gerados utilizando-se o software R (R CORE TEAM, 2017).

3.2 Descrição da simulação

3.2.1 Matriz de marcadores SNPs

Empregou-se a distribuição de Bernoulli que nos fornece como resultado os valores 0 ou 1 para obtenção de cada valor do vetor de haplótipo. Portanto, considerando $\mathbf{H^t} = (H_1,\ldots,H_p)$ um vetor de dimensão p×1 de variáveis de Bernoulli, ou seja, $H_i \sim \text{Bernoulli}(\mu_i)$ com i = 1,...,p, o vetor de médias será dado por $\mu = (\mu_1,\ldots,\mu_p)$. A simulação da matriz **X** de marcadores dos conjuntos de dados, será resumida em 5 etapas que serão discutidas a seguir.

1ª ETAPA: Inicialmente, definiu-se 500 marcadores de SNPs para análise do conjunto de dados.

2ª ETAPA: Obteve-se o vetor de médias marginais de cada SNP. O vetor de médias $\mu_i = (\mu_1,\ldots,\mu_{500})$ com i = 1,...,500 foi obtido a partir da simulação com base na distribuição uniforme (0.1, 0.9).

3ª ETAPA: Calculou-se a matriz de covariâncias **V** de **H**. Para tal, considerando $H_i \sim \text{Bernoulli}(\mu_i)$, inicialmente construiu-se uma matriz em que a diagonal principal continha os valores de $V_{ii} = \text{var}(H_i) = \mu_i(1-\mu_i)$ e os demais zero. Após, obtemos a matriz de correlação ρ_{ij} $i,j = 1,\ldots,500$, em que $\rho_{ij} = 1$**,** quando i = j. Em geral, os SNPs mais próximos, estão mais correlacionados, portando a matriz de correlação foi obtida a partir da simulação de uma distribuição uniforme dependente da distância entre os SNPs. Conforme a Tabela 2:

Tabela 2- Correlação entre SNPs simulada da distribuição uniforme com diferentes intervalos dependentes da diferença da distância |*i-j*| entre dois SNPs.

$\lvert i-j\rvert$	1	2	3	4	5	6	7	8	9	10
Distância	(0.6,0.9)	(0.4,0.6)	(0.3,0.6)	(0.3,0.5)	(0.2,0.5)	(0.2,0.4)	(0.1,0.4)	(0.1,0.3)	(0.1,0.2)	(0,0.1)

Enfim, a matriz **V** de covariâncias de **H**, que é simétrica foi obtida, sendo sua diagonal principal composta pelos valores de $V_{ii} = \text{var}(H_i) = \mu_i(1-\mu_i)$ e os demais por $V_{ij} = V_{ji} = \text{cov}(H_i, H_j) = \rho_{ij}\sqrt{V_{ii}V_{jj}}$ com i ≠ j.

4ª ETAPA: Com a matriz de covariância **V** e o vetor das médias dos SNPs, foram simulados os haplótipos. O algoritmo para encontrarmos os vetores de haplótipos $\mathbf{H_i}$ dos indivíduos pode ser resumido da seguinte maneira:

- Geramos $\mathbf{H_1}$ a partir de uma Bernoulli (μ_i);

- Para i = 2, ..., p consideramos $\mathbf{V_{i-1}}$ a matriz de covariâncias de (H_1, H_2, ..., H_{i-1}) em que o índice i-1 representa as primeiras i-1 linhas e colunas de **V**. Seja um vetor $\mathbf{s_i}$ de dimensão i-1 dado por $\mathbf{s_i} = (\text{cov}(\mathbf{H_1}, \mathbf{H_i}),\ldots,\text{cov}(\mathbf{H_{i-1}}, \mathbf{H_i}))^t$, podemos observar que $\mathbf{s_i}$ é exatamente a primeira i-1 entrada da i-ésima coluna da matriz **V**. Então, $\mathbf{H_i}$ é gerado a partir de uma Bernoulli(μ_i^*), em que μ_i^* é média condicional de $\mathbf{H_i}$ dados os valores de $(\mathbf{H_1},\ldots,\mathbf{H_{i-1}})^t$, dados que $(\mathbf{H_1},\ldots,\mathbf{H_{i-1}})^t = (h_1,\ldots,h_{i-1})^t$, a média condicional de μ_i^* é dada por:

$$\begin{aligned}\mu_i^* &= P(\mathbf{H_i} = 1|h_1,\ldots,h_{i-1}) \\ &= \mu_i + \mathbf{V_{i-1}^{-1}}\mathbf{s_i}[(h_1,\ldots,h_{i-1})^t - (\mu_1,\ldots,\mu_{i-1})^t]\end{aligned}$$

5ª ETAPA: Definiu-se 100 indivíduos, ou seja, a matriz **X** terá 100 linhas. Para obtenção dos genótipos de cada indivíduo, combinou-se dois haplótipos independentes. Obtendo assim, a matriz **X** de marcadores de SNPs.

3.2.2 Vetor de observações fenotípicas

O vetor **y** de fenótipos é obtido de um modelo de regressão linear múltipla, dado por:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + \varepsilon_i$$

Sendo, X_{ij} os elementos da matriz de marcadores de SNPs com i = 1,...,N e j = 1,...,p, β_k os coeficientes de regressão com k= 0,...,p e **Ɛ** o vetor de erros aleatórios que possui uma distribuição normal com média zero e variância σ^2, ou seja, $\varepsilon_i \sim N(0, \sigma^2)$. Matricialmente esse modelo é descrito por:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{8}$$

Para encontrarmos o vetor **y**, devemos utilizar a matriz **X** de marcadores e obtermos os vetores **β** e **Ɛ**. Considerando $\sigma^2 = 1$, o vetor **Ɛ** foi gerado de uma distribuição normal (0,1).

Os dados foram avaliados em dois cenários diferentes, que diferem entre si de acordo com o vetor de **β** utilizado. No primeiro cenário o **β** foi gerado de uma distribuição

normal (0,1) e os valores maiores do que módulo de 1,6 foram escolhidos para serem significativos, ou seja, responsáveis pela variação da característica fenotípica de interesse, já no segundo cenário foram escolhidos alguns valores aleatoriamente, variando de 0,4 a 2,1. Obtidos os vetores $\boldsymbol{\beta}$, $\boldsymbol{\mathcal{E}}$ e a matriz $\mathbf{X}$ de marcadores SNPs, foi encontrado o vetor $\mathbf{y}$ de observações fenotípicas, conforme a equação (8).

O processo descrito acima foi repetido 1000 vezes, produzindo assim 1000 conjunto de dados. A cada simulação as matrizes $\mathbf{X}$ e $\boldsymbol{\mathcal{E}}$ variavam, produzindo diferentes valores para o vetor $\mathbf{y}$**.** O vetor $\boldsymbol{\beta}$ foi gerado uma única vez e então os seus valores foram fixados, objetivando verificar a frequência com que os métodos selecionam esses valores fixos.

Em cada conjunto simulado, as amostras/indivíduos de $\mathbf{X}$ e $\mathbf{y}$ foram separadas em dois subconjuntos (treinamento e teste) 20% das amostras formaram o conjunto de teste e as restantes o conjunto de treinamento, segundo o algoritmo KS (KENNARD; STONE, 1969). Ao fim das análises, em cada cenário, para cada simulação tínhamos 20 amostras no conjunto de teste e 80 no de treinamento.

Com o conjunto de treinamento foram ajustados os modelos usando os métodos propostos (BLASSO supervisionado, SPLS e PLS-OPS). O conjunto de teste foi utilizado para verificar a eficiência do modelo construído e sua capacidade preditiva.

De maneira geral, a simulação pode ser descrita resumidamente conforme a Figura 7. Incialmente, entra-se com o número de SNPs e indivíduos que deseja avaliar, obtendo a matriz $\mathbf{X}$ de marcadores SNPs. Posteriormente, obtém-se o vetor $\boldsymbol{\beta}$ e o fixa. Com a $\mathbf{X}$ obtém-se o $\mathbf{y_r}$ que é o vetor de fenótipos reais. Para obtenção do vetor de fenótipos observados ($\mathbf{y_{obs}}$), basta somar o $\mathbf{y_r}$ com o vetor $\boldsymbol{\mathcal{E}}$, obtido através de simulação de uma distribuição normal (0,1). Em seguida, particiona-se os conjuntos utilizando o algoritmo KS (KENNARD; STONE, 1969), obtendo as matrizes $\mathbf{X_{trn}}$ e $\mathbf{X_{tst}}$ e os vetores $\mathbf{y_{obs\_trn}}$, $\mathbf{y_{obs\_tst}}$, então, ajusta-se os modelos utilizando a matriz $\mathbf{X_{trn}}$ e o vetor $\mathbf{y_{obs\_trn}}$ obtendo-se o vetor de coeficientes da regressão estimado ($\widehat{\boldsymbol{\beta}}$) através dos métodos propostos. Na sequência obtém-se a estimativa do vetor de fenótipos **(**$\hat{\boldsymbol{y}}$), utilizando o $\widehat{\boldsymbol{\beta}}$ e a matriz $\mathbf{X_{tst}}$, por fim avalia a capacidade preditiva comparando o $\hat{\boldsymbol{y}}$ com $\mathbf{y_{obs\_tst}}$, através de estatísticas que inferem sobre erros de predição, como coeficiente de correlação (r) e o RMSE.

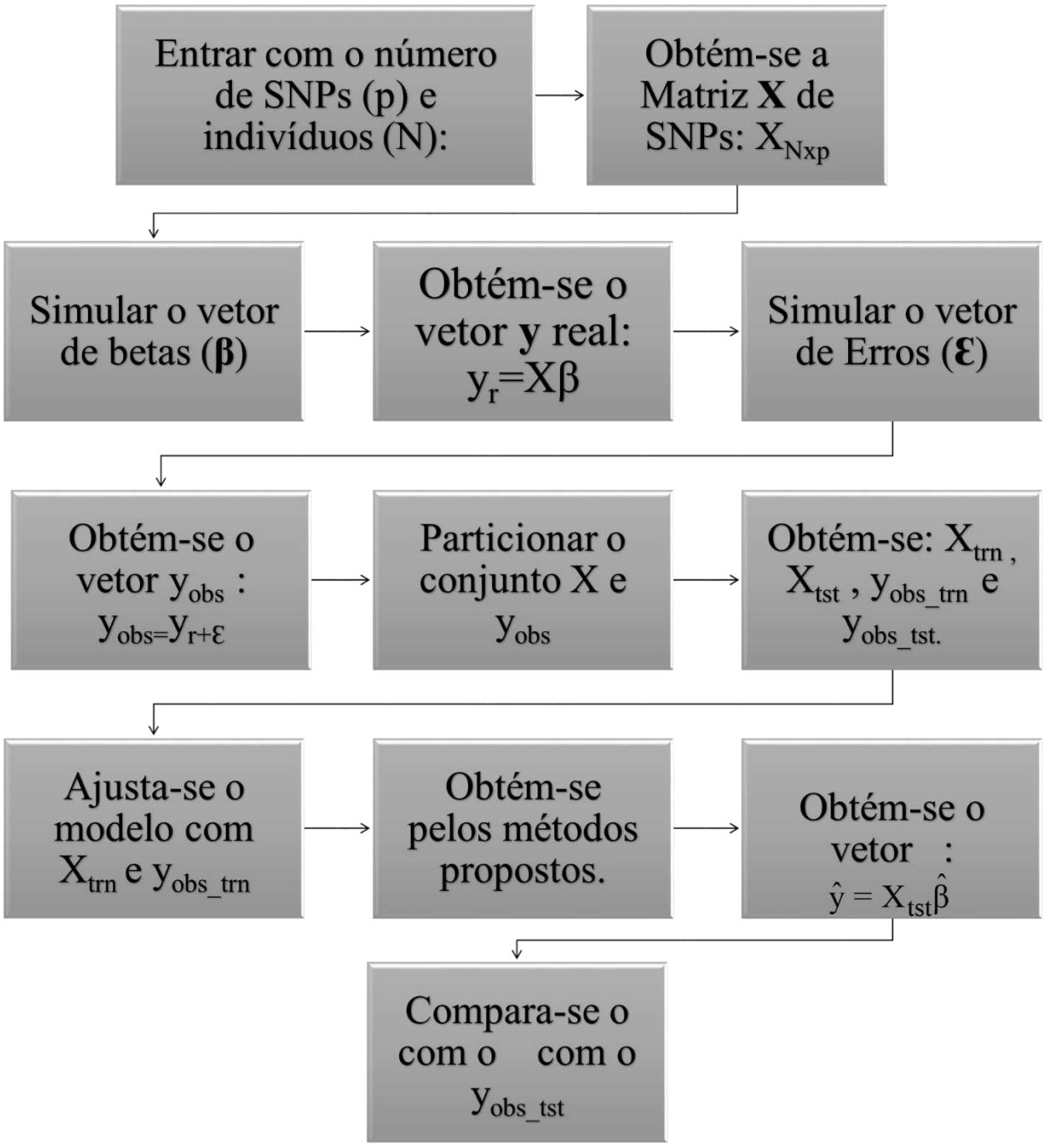

Figura 7- Esquema dos passos para a simulação e análise de marcadores SNPs.

3.3 Conjunto de dados reais

3.3.1 Conjunto de dados SNPs

Foi utilizado um conjunto de dados de produção de milho em condição irrigada apresentados por Crossa et al. (2011). Para tal, avaliou-se a produção de grãos como característica quantitativa.

O número de linhas (indivíduos) no conjunto de dados produção de grãos é de 264. Os marcadores disponíveis para a análise foram 1135. Ao utilizarmos o algoritmo KS o conjunto de treinamento foi constituído de 211 indivíduos e o de teste 53 indivíduos.

3.3.2 Conjunto de dados NIR - Teor de fibra da Cana-de-açúcar

O conjunto de dados utilizado foi o do teor de fibra (FIBRA) da Cana-de-açúcar obtidos de um experimento realizado no programa de melhoramento genético da cana-de-açúcar da Universidade Federal de Viçosa (PMGCA-UFV).

Neste estudo utilizou-se 168 amostras contrastantes em relação à FIBRA. Os espectros foram obtidos no terço médio do colmo. Os dados referentes aos espectros foram dispostos em uma matriz **X**, com 168 linhas e 3113 colunas.

Nesse estudo os melhores pré-tratamentos foram: alisamento, centragem na média e correção multiplicativa de sinal MSC.

As amostras foram separadas em dois conjuntos: um de calibração e em outro de previsão, usando o algoritmo KS (KENNARD; STONE, 1969). O conjunto de calibração continha 138 amostras (cerca de 82 %) e o de previsão as 30 restantes (cerca de 18%) com 8 VL.

Todas as rotinas dos métodos empregados foram implementadas no software R, que serão brevemente discutidas a seguir.

3.3.3 Conjunto de dados NIR- Teor de lignina da Cana-de-açúcar

O conjunto de dados em estudo foram fornecidos pelo banco de germoplasma do Programa Genético de Melhoramento de Cana-de-Açúcar (PMGCA) da Universidade Federal de Vicosa, Vicosa, Minas Gerais (MG), Brasil (ASSIS et al., 2017).

Um total de 256 análises foram realizadas na folha com o objetivo de predizer o teor de Lignina na Cana-de-Açúcar. Os espectros NIR foram obtidos diretamente da folha verde, sem procedimento de preparação da amostra. Os dados referentes aos espectros foram dispostos em uma matriz **X**, com 256 linhas e 1038 colunas.

Nesse estudo os melhores pré-tratamentos foram: correção de linha de base, centragem na média, segunda derivada e correção multiplicativa de sinal MSC.

O conjunto de previsão foi constituído de 40 amostras, e as restantes formaram o conjunto de calibração, segundo o algoritmo KS (KENNARD; STONE, 1969).

3.3.4 Conjunto de dados NIR – Repolho Roxo

As amostras dos conjuntos de dados foram obtidas no comércio em Viçosa, Minas Gerais, entre outubro de 2013 e outubro de 2014 (OLIVEIRA et al., 2018c). Este experimento teve por objetivo determinar as propriedades antioxidantes do extrato de 82 amostras de repolho roxo utilizando espectroscopia NIR.

Os melhores pré-tratamentos foram: correção de linha de base, centragem na média e alisamento. Os dados referentes aos espectros foram dispostos em uma matriz **X**, com 82 linhas e 3113 colunas.

O conjunto de previsão foi constituído de 20 amostras, e as restantes formaram o conjunto de calibração, segundo o algoritmo KS (KENNARD; STONE, 1969).

Os espectros foram obtidos através do software Matlab 7.9 (Math Works, Natick, USA) no Laboratório de Instrumentação e Quimiometria (LINQ) .

3.4 Recursos Computacionais

3.4.1 *Sparce partial least Square* (SPLS)

Para ajuste dos modelos, utilizando o método *Sparce partial least Square* (SPLS) foi utilizado a função spls do pacote SPLS (CHUN; KELES, 2010) do software R.

3.4.2 Seleção dos Preditores Ordenados associado a regressão PLS (PLS-OPS)

As rotinas referentes ao método PLS-OPS (TEÓFILO et al., 2009), foram todas implementadas no software R e para ajuste do modelo usou-se o pacote pls. As rotinas computacionais implementadas no programa serão apresentadas em breve numa publicação científica.

3.4.3 BLASSO supervisionado

Nesse estudo, para ajuste do modelo com o método BLASSO supervisionado, foi utilizado a função bglr do pacote BGLR do software R (DE LOS CAMPOS;

RODRIGUEZ, 2016). Após uma etapa de testes decidiu-se por utilizar 25000 iterações, das quais 10000 foram descartadas (burn-in) para assegurar o aquecimento da cadeia e com seleção de uma a cada 3 iterações (thin).

Visto que o BLASSO supervisionado faz muitos coeficientes de regressão tenderem a valores próximos de zero, foi criado inicialmente um critério de seleção em que 80% das variáveis menos significativas eram descartados e as restantes selecionadas.

Para melhor comparação entre os métodos propostos no segundo cenário de dados simulados, além do critério de seleção descrito anteriormente, um segundo critério foi adotado para o BLASSO supervisionado. Para tal, igualamos o número de variáveis que o BLASSO supervisionado selecionaria ao número de variáveis selecionados pelos métodos SPLS e PLS-OPS respectivamente.

Após a seleção desses coeficientes, um novo vetor que contém as estimativas dos fenótipos ($\hat{y}$) foi encontrado e então estatísticas que inferem sobre erros de predição foram realizadas.

3.5 Avaliação do desempenho dos métodos

Para avaliarmos o desempenho dos modelos construídos, bem como sua capacidade preditiva, utilizamos dois critérios:

1. Frequência com que os métodos propostos selecionaram os coeficientes verdadeiro nos conjuntos de dados simulados;
2. Os valores médios de correlação e os intervalos de variação de RMSE no conjunto de teste.

4. RESULTADOS E DISCUSSÃO

4.1 Conjuntos de dados simulados (Primeiro Cenário)

No primeiro cenário, o vetor de coeficientes de regressão que representam os efeitos dos SNPs reais foi constituído de 52 elementos, cada um correspondente a um marcador.

4.1.1 Modelo Completo

Podemos observar na Tabela 3 os valores do coeficiente de correlação entre os valores preditos e os valores reais e o intervalo de variação da raiz quadrada do erro quadrático médio nos conjuntos de teste usando o método PLS e o BLASSO sobre todas as variáveis da matriz de dados nas 1000 simulações.

Tabela 3 - Coeficiente de correlação médio (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste nas 1000 simulações, avaliadas no primeiro cenário, pelos métodos BLASSO e PLS, sobre o conjunto de dados completos.

	BLASSO	PLS
r	0,688	0,702
RMSE	4,22 a 12,02	3,71 a 18,51

4.1.2 Modelos com seleção de variáveis

A frequência com que os modelos utilizando os métodos BLASSO supervisionado, SPLS e PLS-OPS selecionaram os SNPs reais podem ser vistos nas Figuras 8, 9 e 10.

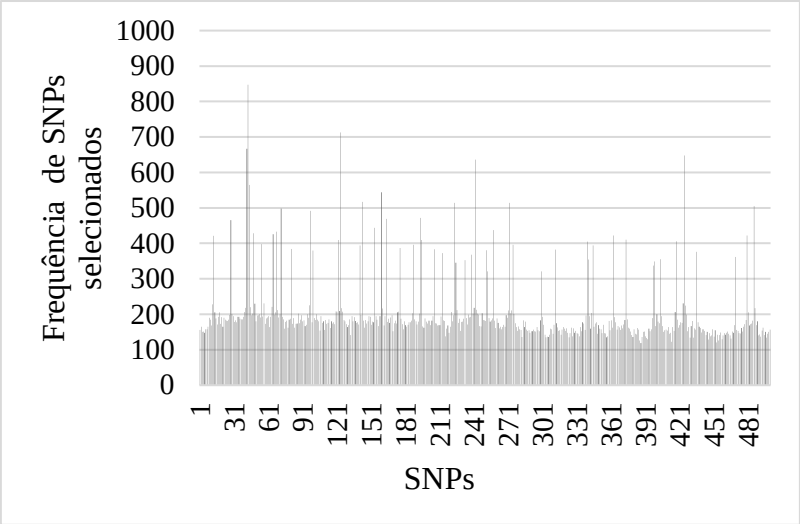

Figura 8- Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do BLASSO supervisionado como método de seleção avaliados no primeiro cenário.

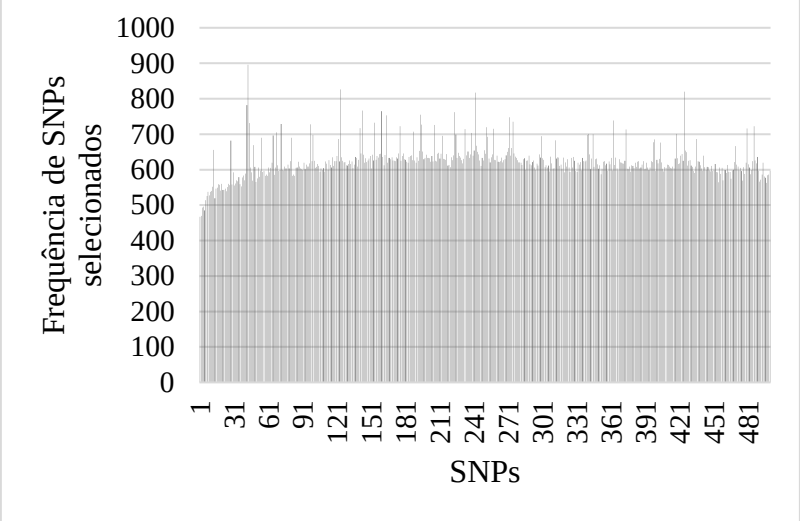

Figura 9- Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do SPLS como método de seleção avaliados no primeiro cenário.

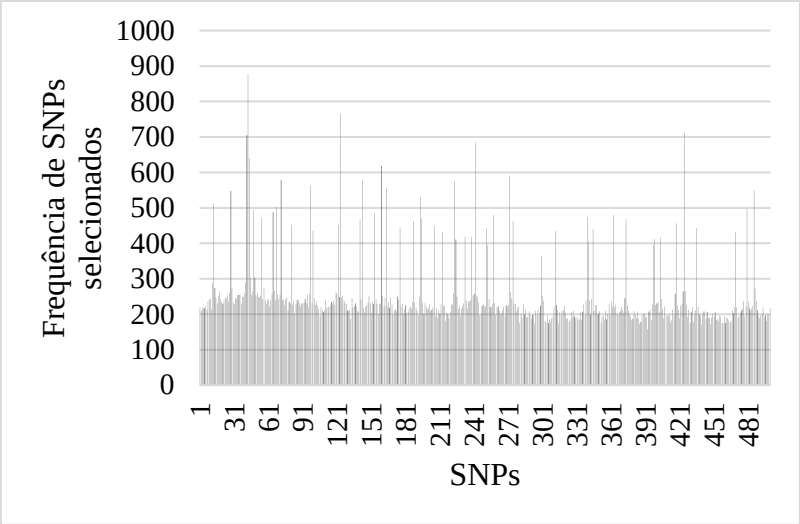

Figura 10 - Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do PLS-OPS como método de seleção avaliados no primeiro cenário.

Nesse trabalho, simulamos a matriz de correlação a partir de uma distribuição uniforme que variava conforme a distância entre os SNPs. Segundo Feng et al. (2012), os SNPs mais próximos são mais correlacionados, porém alguns SNPs mais distantes podem ter associações mais fortes do que os mais próximos, justificando o fato que além das variáveis originalmente significativas os métodos também selecionaram outras variáveis, principalmente as mais próximas das significativas. Comparando-se as Figuras 8, 9 e 10 podemos observar que o SPLS seleciona os SNPs não significativos mais frequentemente do que o BLASSO supervisionado e PLS-OPS. Já o PLS-OPS e o BLASSO supervisionado possuem comportamento semelhantes.

De maneira geral os SNPs tomados como os responsáveis pela variação fenotípica foram os mais selecionados, destacando-se os SNP43, SNP124, SNP242 e SNP425 que foram os mais frequentes em todos os métodos utilizados, conforme a Tabela 4.

Tabela 4- Frequência com que os modelos selecionaram alguns SNPs de maiores efeitos avaliados no primeiro cenário. Total de 1000 simulações.

	Frequência		
	BLASSO supervisionado	SPLS	PLS-OPS
SNP 43	847	896	876
SNP 124	713	826	766
SNP 242	636	817	685
SNP 425	648	820	713

Podemos notar que esses SNPs mais selecionados foram os que possuem maiores efeitos.

O modelo que contêm todos os SNPs significativos avaliado nesse cenário pelos métodos BLASSO supervisionado e PLS-OPS não foi selecionado nenhuma vez, porém em 5 vezes 31 dos 52 SNPs reais foram selecionados pelo BLASSO supervisionado e em 7 vezes 38 SNPs reais foram selecionados pelo PLS-OPS. Já o SPLS, selecionou o modelo exato 168 vezes.

Em média os modelos construídos pelo método SPLS selecionaram cerca de 310 variáveis a cada simulação, enquanto que pelo PLS-OPS selecionaram em média 124 variáveis. Já o BLASSO supervisionado conforme o critério de seleção adotado, a cada simulação os modelos selecionaram 100 variáveis das 500 presentes.

A Tabela 5, apresenta os valores referente ao coeficiente de correlação médio (r) e o intervalo de variação do RMSE referente a cada método no conjunto de teste.

Tabela 5- Coeficiente de correlação médio (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste nas 1000 simulações, avaliadas no primeiro cenário.

	BLASSO supervisionado	SPLS	PLS-OPS
r	0,703	0,690	0,696
RMSE	4,05 a 11,27	4,27 a 18,64	3,44 a 17,44

Podemos notar que os métodos funcionam de maneira semelhante com valores de r muito próximos, ou seja, em termos de capacidade preditiva estes métodos são similares.

Comparando-se o modelo completo com o de seleção de variáveis, podemos perceber que a capacidade preditiva dos modelos foram similares (Tabelas 3 e 5).

Para melhor entendimento das estatísticas apresentadas a respeito do RMSE e do r, tomou-se aleatoriamente como exemplo o conjunto simulado número 3.

Nessa simulação, os valores de RMSE e r referentes ao BLASSO supervisionado foram de RMSE = 6,91 e r = 0,75, referentes ao SPLS foram de RMSE = 6,20 e r = 0,75, e referentes ao PLS-OPS foram de RMSE = 6,56 e r = 0,75. Segundo Ferreira (2015) o modelo é considerado adequado quando o RMSE é bem menor que o desvio padrão dos dados no conjunto de teste, ou quando a razão entre o desvio padrão dos dados no conjunto de teste e o valor do RMSE é um número em torno de 10. O desvio padrão dos dados originais no conjunto de teste foi de 9,03 e a razão entre o desvio padrão e o RMSE foi de aproximadamente 1,45 para o BLASSO supervisionado, aproximadamente 1,42 para o SPLS e aproximadamente 1,4 para o PLS-OPS. Podemos, então, inferir que para o conjunto de dados escolhido o modelo de predição ainda não está adequado.

De maneira geral, após o uso dos métodos para selecionar as "melhores" variáveis, observamos que os valores de RMSE e de r se mantiveram bastante próximos comparado ao modelo completo. Aparentemente a grande vantagem desses modelos de seleção de variáveis seria identificar regiões mais influentes na variável em estudo, além de trabalhar com um número reduzido de variáveis.

4.2 Conjuntos de dados simulados (Segundo Cenário)

Para análise do segundo cenário, dez SNPs (SNP2, SNP5, SNP10, SNP34, SNP49, SNP 73, SNP76, SNP139, SNP153, SNP199) foram escolhidos aleatoriamente para contribuírem para a variação fenotípica, com coeficientes de regressão dados respectivamente por $\beta = (0,4; 0,6; 0,7; 0,9; 1,2; 1,5; 1,6; 1,9; 2; 2,1)$. Esse segundo cenário difere do primeiro cenário em que os coeficientes reais correspondiam a valores acima do módulo de 1,6. Foram escolhidos valores menores para verificar a capacidade dos modelos construídos em detectar esses SNPs de menores efeitos.

4.2.1 Modelo Completo

Na Tabela 6 são apresentados os valores do coeficiente de correlação entre os valores preditos e os valores reais e o intervalo de variação do RMSE nos conjuntos de teste usando o método PLS e o BLASSO sobre todas as variáveis da matriz de dados nas 1000 simulações.

Tabela 6- Coeficiente de correlação médio (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste nas 1000 simulações, avaliadas no segundo cenário, pelos métodos BLASSO e PLS sobre o conjunto de dados completo.

	BLASSO	PLS
r	0,665	0,632
RMSE	1,24 a 3,88	1,30 a 5,90

4.2.2 Modelos com a seleção de variáveis

Nesse cenário, para escolha do número de variáveis a serem selecionadas pelo BLASSO supervisionado utilizou-se dois critérios, no primeiro selecionou 20% das variáveis presentes e no segundo o número de variáveis selecionadas foi igual ao número de variáveis selecionadas pelos métodos SPLS e PLS-OPS respectivamente. As Figuras 11, 12 e 13 apresentam os gráficos de frequências referente a cada método, sendo o BLASSO supervisionado avaliado no primeiro critério.

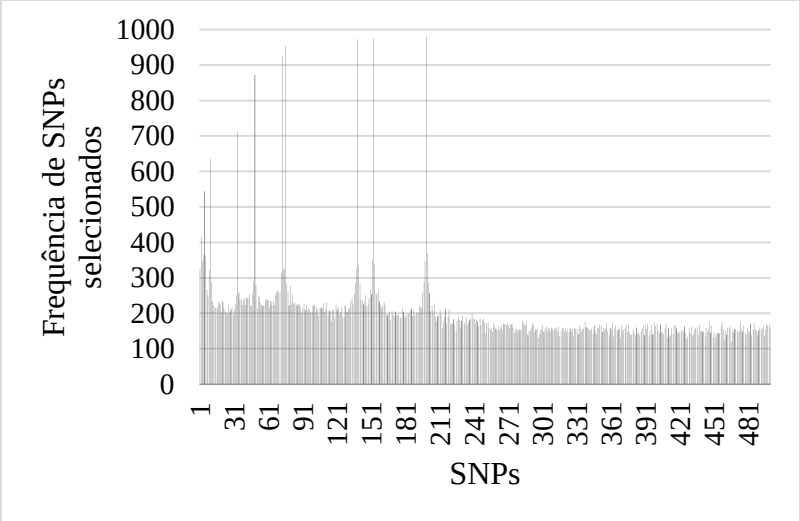

Figura 11- Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do BLASSO supervisionado (20% das variáveis) como método de seleção avaliados no segundo cenário.

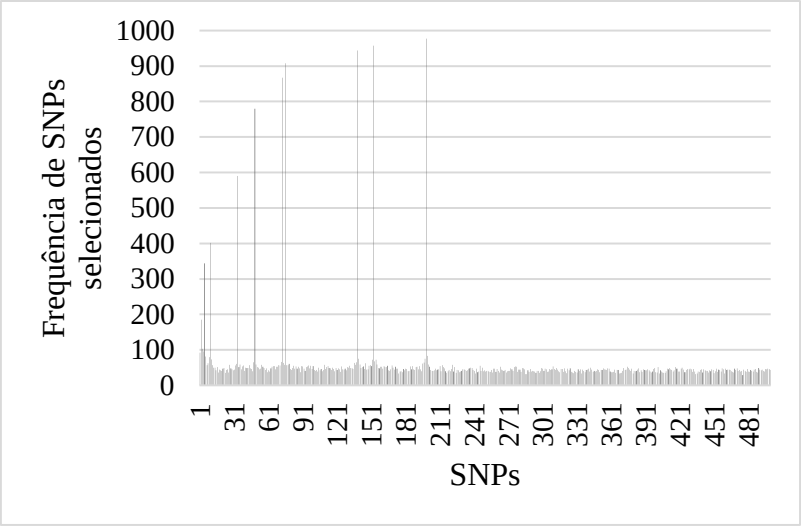

Figura 12 - Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do SPLS como método de seleção avaliados no segundo cenário.

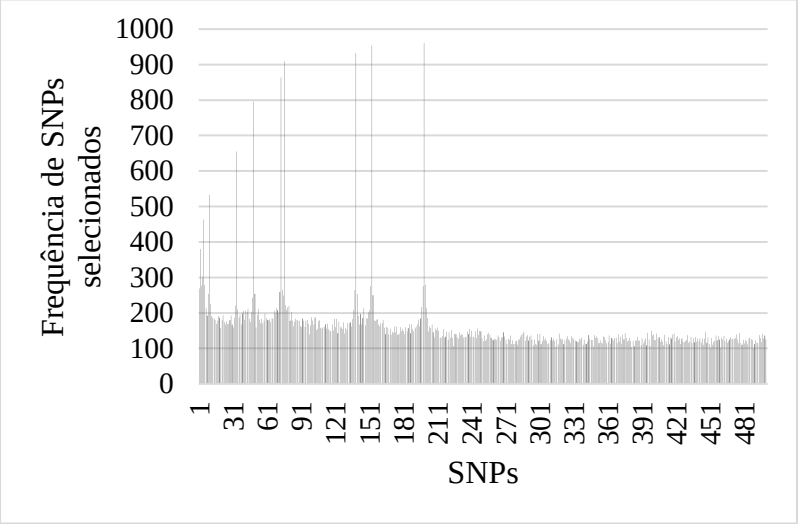

Figura 13- Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do PLS-OPS como método de seleção avaliados no segundo cenário.

Em média os modelos construídos pelos métodos BLASSO supervisionado, SPLS e PLS-OPS selecionaram respectivamente 100, 30 e 80 em cada simulação. Analisando as Figuras 11, 12 e 13 podemos notar que quando os valores dos coeficientes são menores o SPLS seleciona poucas variáveis irrelevantes em relação aos outros métodos. O mesmo resultado foi obtido por Feng et al. (2012), no qual objetivou-se através de estudos de simulação comparar os métodos SPLS e LASSO para a seleção de SNPs. Neste trabalho também foram avaliados SNPs de menores efeitos. Os resultados da simulação indicaram que o SPLS selecionou menos variáveis irrelevantes do que o LASSO.

De maneira geral, os SNPs tomados como os responsáveis pela variação fenotípica foram os mais selecionados pelos métodos (Tabela 7).

Tabela 7 - Frequência com que os modelos selecionaram os SNPs significativos definidos no segundo cenário.

	Frequência		
	BLASSO supervisionado	SPLS	PLS-OPS
SNP 2	419	185	380
SNP 5	544	344	463
SNP 10	635	402	533
SNP 34	712	590	655
SNP 49	872	780	795
SNP 73	925	867	864
SNP 76	953	908	910
SNP 139	973	944	932
SNP 153	975	957	954
SNP 199	981	977	960
Todos os SNPs reais	83	31	75

Observa-se (Tabela 7) que nos três métodos o SNP2, que é o de menor efeito, foi poucas vezes selecionado, apontando uma falha nos métodos em detectar os SNPs de baixo efeito. Por outro lado, os de maiores efeitos foram frequentemente identificados e selecionados nas simulações, como era de se esperar.

O modelo verdadeiro exato avaliado nesse cenário foi mais vezes selecionado quando se utilizou o método BLASSO supervisionado. Apesar do modelo exato não ter sido muito selecionado nas simulações, podemos observar (Tabela 8) o número de vezes que 7, 8 e 9 dos 10 SNPs reais foram selecionados.

Tabela 8- Números de vezes que os SNPs reais foram selecionados respectivamente pelos métodos BLASSO supervisionado, SPLS e PLS-OPS.

Número de SNPs reais	Número de vezes que os SNPs reais foram selecionados		
	BLASSO supervisionado	SPLS	PLS-OPS
7	218	257	211
8	338	224	286
9	264	144	202

A Tabela 9, apresenta os valores referente ao coeficiente de correlação médio (r) e o intervalo de variação do RMSE referente a cada método no conjunto de teste.

Tabela 9- Coeficiente de correlação médio (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste nas 1000 simulações. Avaliados no segundo cenário.

	BLASSO supervisionado	SPLS	PLS-OPS
r	0,753	0,846	0,705
RMSE	1,13 a 3,35	0,92 a 4,66	0,86 a 5,32

Em termos de capacidade preditiva podemos verificar que o SPLS superou os outros métodos. Os desvios padrão na população de teste dos dados variaram entre 1,78 a 5,71. As razões entre os desvios padrão e os valores de RMSE referente a cada método ainda não estão adequadas para um modelo satisfatório.

Os gráficos de frequências nas Figuras 14, 15, refere-se ao BLASSO supervisionado avaliado no segundo critério proposto, no qual igualamos o número de variáveis que o BLASSO supervisionado selecionaria ao número de variáveis selecionadas pelos métodos SPLS e PLS-OPS.

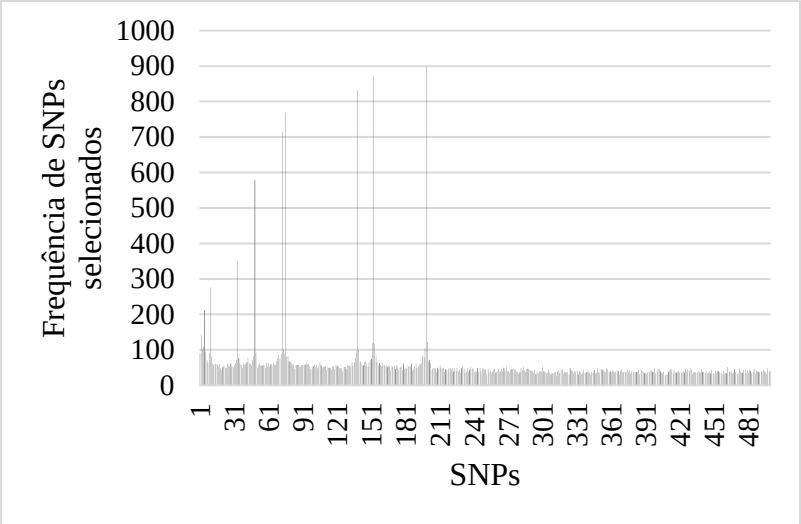

Figura 14- Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do BLASSO supervisionado como método de seleção, adotando-se o mesmo número de variáveis selecionadas em cada simulação pelo método SPLS.

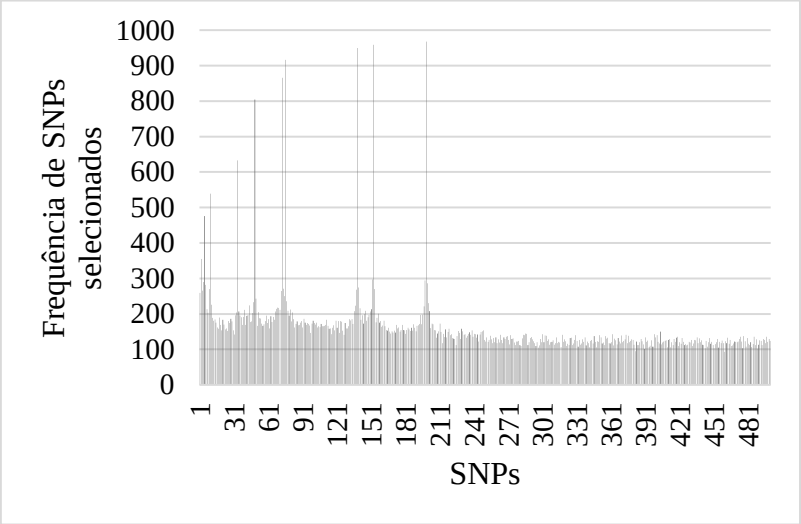

Figura 15- Frequência de seleção dos 500 SNPs nos 1000 conjuntos de dados simulados pelo uso do BLASSO supervisionado como método de seleção, adotando-se o mesmo número de variáveis selecionadas em cada simulação pelo método PLS-OPS.

Podemos observar que quando o BLASSO supervisionado utiliza os critérios do SPLS e PLS-OPS para escolha do número de variáveis selecionadas, este seleciona menos variáveis irrelevantes quando avaliado sob o primeiro critério (20% das variáveis), principalmente quando se utiliza o SPLS, obtendo melhores valores de coeficiente de correlação. O valor do coeficiente de correlação do BLASSO supervisionado passou a ser 0,80 com o RMSE variando de 0,74 a 3,25 ao adotarmos o número de variáveis do SPLS, selecionando o modelo correto 31 vezes.

Analisando o BLASSO supervisionado com o critério do PLS-OPS obteve-se um valor de coeficiente de correlação de 0,77 e RMSE variando de 0,79 a 3,76, selecionando o modelo correto 68 vezes.

Comparando-se os modelos completos com os de seleção podemos perceber que em todos os métodos o coeficiente de correlação médio aumentou significativamente e o intervalo de variação do RMSE diminuiu, indicando que realizar a seleção quando os efeitos dos SNPs são de menores efeitos é uma boa alternativa.

4.3 Conjuntos de dados reais

4.3.1 Conjunto de dados de SNPs

Na Tabela 10 são apresentados os valores do coeficiente de correlação entre os valores preditos e os valores reais e a raiz quadrada do erro quadrático médio nos conjuntos de teste usando os métodos BLASSO e PLS sobre todas as variáveis (1135 colunas) da matriz de dados. Na Tabela 11, temos as mesmas estatísticas após o uso dos métodos de seleção propostos.

Tabela 10 - Coeficiente de correlação médio (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste no conjunto de dados reais de marcadores SNPs, pelos métodos BLASSO e PLS.

	BLASSO	PLS
r	0,525	0,491
RMSE	0,73	2,63

Tabela 11- Coeficiente de correlação médio (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste no conjunto de dados reais de marcadores SNPs.

	BLASSO supervisionado	SPLS	PLS-OPS
r	0,56	0,445	0,552
RMSE	0,56	2,76	2,96

Os resultados obtidos ao utilizarmos o BLASSO supervisionado e o PLS-OPS foram superiores aos do SPLS em termos de coeficiente de correlação. Este mesmo conjunto de dados foi analisado por Crossa et al. (2010) segundo diversos métodos, dentre

os quais o método BLASSO, o qual apresentou melhores resultados. O valor de coeficiente de correlação obtido naquele trabalho foi igual a 0,525 considerando o modelo completo com RMSE=0,73.

Podemos notar que ao aplicarmos o BLASSO supervisionado utilizando o critério seleção de variáveis adotado (20% das variáveis eram selecionadas) e o método PLS-OPS, esse valor de correlação obtido por Crossa et al. (2010) foi superado, indicando que realizar seleção de variáveis é uma boa alternativa e que o método PLS-OPS pode ser utilizado em dados de marcadores de SNPs.

Para este conjunto de dados, o desvio padrão na população de teste foi de 0,89, enquanto que os valores de RMSE para o BLASSO supervisionado, SPLS e PLS-OPS foram respectivamente 0,56, 2,76 e 2,96. Portanto o valor de RMSE = 0,56, com a razão entre o desvio padrão e o RMSE aproximadamente igual a 1,6 referente ao BLASSO supervisionado é o melhor modelo de predição, indicando que o melhor método para modelagem desse conjunto de dados é o BLASSO supervisionado.

Nesse conjunto de dados, o BLASSO supervisionado selecionou 227 SNPs enquanto que o SPLS e PLS-OPS selecionaram, respectivamente, 1011 e 26 SNPs.

4.3.2 Conjunto de dados NIR

Conforme descrito detalhadamente no Material e Métodos foram avaliados três conjuntos de dados NIR:

DADOS 1:Teor de fibra da cana-de-açúcar. A matriz de dados é composta por 168 amostras (linhas) e 3113 variáveis (colunas).

DADOS 2: Teor de lignina da Cana-de-açúcar. Os dados referentes aos espectros foram dispostos em uma matriz **X**, com 256 linhas e 1038 colunas.

DADOS 3: Repolho Roxo. Os dados referentes aos espectros foram dispostos em uma matriz **X**, com 82 linhas e 3113 colunas.

Os espectros NIR das amostras dos DADOS 1,2 e 3, na faixa de 4000 a 10000 cm^{-1}, são apresentados na Figura 16.

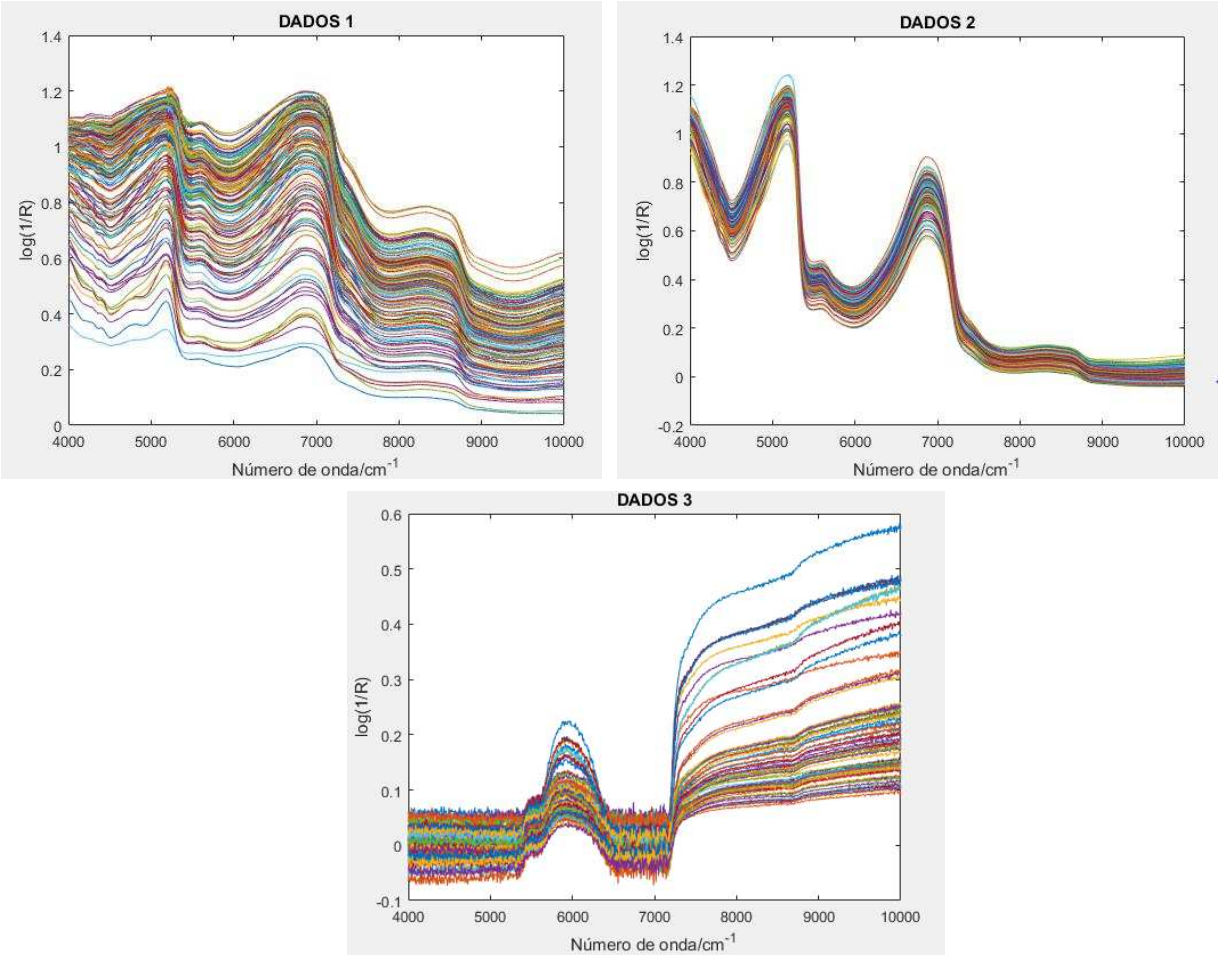

Figura 16- ESPECTROS NIR: (Teor de fibra da cana-de-açúcar (DADOS 1); Teor de lignina da Cana-de-açúcar (DADOS 2); Repolho Roxo (DADOS 3)).

Na Tabela 12 são apresentados os valores do coeficiente de correlação entre os valores preditos e os valores reais e a raiz quadrada do erro quadrático médio nos conjuntos de teste usando os métodos BLASSO e PLS sobre todas as variáveis da matriz dos dados 1, 2 e 3. O número de variáveis latentes (nVL) foi escolhido através do método de validação cruzada leave-one-out (KOHAVI, 1995).

Tabela 12- Coeficiente de correlação (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente ao subconjunto de teste dos 3 conjunto de dados reais de espectroscopia NIR (Teor de fibra da cana-de-açúcar (DADOS 1); Teor de lignina da Cana-de-açúcar (DADOS 2); Repolho Roxo (DADOS 3)). Os métodos BLASSO e PLS foram utlizados sobre todas as variáveis. nVL representa o númeto de variáveis latentes escolhido.

DADOS	**nVL**		**BLASSO**	**PLS**
1	8	r	0,69	0,68
		RMSE	1,74	1,75
2	10	r	0,96	0,93
		RMSE	0,67	0,89
3	6	r	0,99	0,99
		RMSE	14,66	10,49

Na Tabela 13 são apresentados os valores do coeficiente de correlação entre os valores preditos e os valores reais e a raiz quadrada do erro quadrático médio nos conjuntos de previsão (teste) usando os métodos BLASSO supervisionado (20% das variáveis eram selecionadas), SPLS e PLS-OPS respectivamente sobre os conjuntos de dados estudados.

Tabela 13- Coeficiente de correlação (r) e intervalo de variação do RMSE entre o valor predito ($\hat{y}_p$) e o y real pertencente aos subconjuntos de teste dos 3 conjunto de dados reais de espectroscopia NIR avaliados (Teor de fibra da cana-de-açúcar (DADOS 1); Teor de lignina da Cana-de-açúcar (DADOS 2); Repolho Roxo (DADOS 3)). Os métodos BLASSO supervisionado, SPLS e PLS-OPS foram utlizados sobre todas as variáveis. nVL: númeto de variáveis latentes escolhido. nVS: número de variáveis selecionadas.

DADOS		**BLASSO supervisionado**	**SPLS**	**PLS-OPS**
1	r	0,69	0,679	0,676
	RMSE	2,17	2,38	2,83
	nVS	623	2275	42
2	r	0,956	0,83	0,946
	RMSE	0,66	2,35	0,77
	nVS	208	1035	177
3	r	0,995	0,451	0,996
	RMSE	15,45	164,33	13,05
	nVS	623	3111	386

No conjunto de DADOS 1 podemos notar (Tabela 13) que os métodos de seleção funcionam de maneira semelhante, com valores muito próximos de r e RMSE, ou seja, em termos de capacidade preditiva são similares. Porém o número de variáveis selecionadas varia enormemente. O valor do desvio padrão dos dados no conjunto de teste foi de 2,35. O BLASSO supervisionado apresentou o valor de RMSE = 2,02, menor que o obtido nos outros métodos propostos. Porém a razão 2,35/2,02 não atende ao critério definido por Ferreira (2010) para classificar o modelo como adequado. Em termos de capacidade preditiva, o modelo com a seleção de variáveis não melhorou em relação ao modelo completo (Tabelas 12 e 13).

No conjunto de DADOS 2 os melhores métodos de seleção foram o BLASSO supervisionado e PLS-OPS (Tabela 13). Em termos de capacidade preditiva esses modelos foram semelhantes aos obtidos com os dados completos, com a vantagem de ter menos variáveis (Tabelas 12 e 13).

No conjunto de DADOS 3 os melhores métodos de seleção foram novamente os BLASSO supervisionado e PLS-OPS (Tabela 13). Em termos de coeficiente de correlação esses modelos foram semelhantes, porém com valores levemente superiores

de RMSE aos obtidos com os dados completos (Tabelas 12 e 13). Este mesmo conjunto de dados foi analisado por Oliveira et al. (2018c) em que valores encontrados de RMSE e r foram próximos aos obtidos nesse estudo.

5. CONCLUSÕES

Em geral, os métodos BLASSO supervisionado e PLS-OPS proporcionaram predições semelhantes, com estimativas de raiz quadrada do erro quadrático médio e de coeficiente de correlação próximas. O BLASSO supervisionado assegurou modelos mais parcimoniosos, selecionando menos variáveis do que o PLS-OPS. Quando o efeito das variáveis foram de menores magnitudes o SPLS superou os outros métodos, selecionando poucas variáveis irrelevantes.

Ao utilizarmos os métodos de seleção os modelos finais se tornaram mais simples quando comparados aos respectivos modelos com os dados completos, visto que o número de variáveis diminuiu significativamente em todos os conjuntos de dados estudados, sem haver perda significativa de poder de predição.

6. REFERÊNCIAS

ABDEL-RAHMAN, E. M.; MUTANGA, O.; ODINDI, J.; ADAM, E.; ODINDO, A.; ISMAIL, R. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. **Computers and Electronics in Agriculture**, v. 106, p. 11-19, 2014.

ASSIS, C.; RAMOS, R. S.; SILVA, L. A.; KIST, V.; BARBOSA, M. H. P.; TEÓFILO, R. F. 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.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; GUIMARÃES, S. E. F. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, v. 48, n. 6, p. 619-626, 2013.

BOKOBZA, L. Near infrared spectroscopy. **Journal of Near Infrared Spectroscopy**, v. 6, n. 1, p. 3–17, 1998.

BROOKES, A. J. The essence of snps. **Gene**, v. 234, n. 2, p. 177-186, 1999.

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, n. 8, p. 64-71, 2009.

CAI, H.; LAN, X.; LI, A.; ZHOU, Y.; SUN, J.; LEI, C.; ZHANG, C.; CHEN, H. SNPs of bovine HGF gene and their association with growth traits in Nanyang cattle. **Research in veterinary science**, v. 95, n. 2, p. 483-488, 2013.

CALIARI, I. P.; BARBOSA, M. H. P.; FERREIRA, S. O.; TEÓFILO, R. F. Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods. **Carbohydrate polymers**, v. 158, p. 20-28, 2017.

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.

CHUN, H.; KELES, S. **Sparse Partial Least Squares (SPLS) Regression and Classification**,2013. Disponível em: <https://cran.rproject.org/web/packages/spls/spls.pdf>. Acesso em: 21, out. 2017.

COLOMBANI, C. CROISEAU, P.; FRITZ, S.; GUILLAUME, F.; LEGARRA, A.; DUCROCQ, V.; ROBERT-GRANIÉ, C. A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. **Journal of dairy science**, v. 95, n. 4, p. 2120-2131, 2012.

COSTA, R. C.; DE LIMA, K. M. G. Prediction of parameters (soluble solid and pH) in intact plum using NIR spectroscopy and wavelength selection. **Journal of the Brazilian Chemical Society**, v. 24, n. 8, p. 1351-1356, 2013.

CROSSA, J.; DE LOS CAMPOS, G.; PÉREZ, P.; GIANOLA, D.; BURGUEÑO, J.; ARAUS, J. L.; MAKUMBI, D.; SINGH, R. P.; DREISIGACKER, S.; YAN, J.; ARIEF, V.; BANZIGER, M.; BRAUN, H. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v. 186, n. 2, p. 713-724, 2010.

CRUZ, M. F. A.; BUENO, R. D.; SOUZA, F. B.; MOREIRA, M. A.; BARROS, E. G. Identificação de SNPs para conteúdo de ácidos graxos em soja pela técnica HRM. **Pesquisa Agropecuária Brasileira**, v. 48, n. 12, p. 1596-1600, 2013.

DANTAS FILHO, H. A. **Desenvolvimento de técnicas quimiométricas de compressão de dados e de redução de ruído instrumental aplicadas a óleo diesel e madeira de eucalipto usando espectroscopia NIR.** 2007. 158 f. Tese (Doutorado em Química) Universidade Estadual de Campinas, Campinas, 2007.

DE LOS CAMPOS, G.; RODRIGUEZ, P. P. **Bayesian Generalized Linear Regression**. 2016. Disponível em: <https://cran.r-project.org/web/packages/BGLR/BGLR.pdf>. Acesso em: 21, out. 2017.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, n. 1, p. 375-385, 2009.

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.

FERREIRA, M. M. C. Multivariate QSAR. **J. Braz. Chem. Soc**., São Paulo, SP, v.13, n.6, p.742-753, 2002.

FERREIRA, M. M. C. **Quimiometria – Conceitos, Métodos e Aplicações.** Campinas, SP: Editora Unicamp, 2015. 493 f.

FERREIRA, S.P. **Estudo comparativo do pós-processamento estatístico aplicado ao modelo brams.** 2011. 151f. Dissertação (Mestrado em Sensoriamento Remoto) - Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2011.

FRAZER, K. A.; MURRAY, S. S.; SCHORK, N.J.; TOPOL, E.J. Human genetic variation and its contribution to complex traits. **Nature Reviews Genetics**, v. 10, n. 4, p. 241-251, 2009.

FREUND, R. J.; WILSON, W. J.; SA, P. Regression analysis – Statistical Modeling of a response variable. **Elsevier**, Inc., San Diego, 459p, 2006.

GAUCHI, J.P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 171-193, 2001.

GIANOLA, D.; ENCISO, M. P.; TORO, M. A. On marker-assisted prediction of genetic value: beyond the ridge. **Genetics**, v. 163, n. 1, p. 347-365, 2003.

GUIMARÃES, C. C.; ASSIS, C.; SIMEONE, M.L.F.; SENA, M.M. Use of near-infrared spectroscopy, partial least-squares, and ordered predictors selection to predict four quality parameters of sweet sorghum juice used to produce bioethanol. **Energy & Fuels**, v. 30, n. 5, p. 4137-4144, 2016.

JOBIM, M.R.; EWALD, G.; WILSON, M.J.; CHAMUM, B.; JOBIN, L.F. Novos testes de DNA na investigação de paternidade com o suposto pai falecido. **RT/Fasc**, v. 104, n. 874, p. 55-69, 2008.

KENNARD, R. W.; STONE, L. A. Computer Aided Design of Experiments. **Technometrics**, v. 11, p. 137–148, 1969.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. **Ijcai**, v. 14, n. 10, p. 1137-1145, 1995.

LEGARRA, A.; MISZTAL, I. Computing strategies in genome-wide selection. **Journal of Dairy Science**, v. 91, n. 1, p. 360-366, 2008.

LIMA, L. P. **Seleção genômica não paramétrica via distância genética entre subpopulações.** 2017. 101 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, MG, 2017.

MARIGORTA, U. M.; GIBSON, G. A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. **Frontiers in genetics**, v. 5, p. 225, 2014.

MARÔCO J. **Análise de equações estruturais: fundamentos teóricos, software e aplicações.** Report Number, 2010.

MARTENS, H.; NAES, T. **Multivariate calibration**. New York: Wiley, 1996. 438f.

MARTINS, J. P. A.; TEOFILO, R. F.; FERREIRA, M. M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. **Journal of Chemometrics**, v. 24, n. 6, p. 320–332, 2010.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics** , v.157, p.1819 – 29, 2001.

MONTGOMERY, D.C.; PECK, E.A. **Introduction to linear regression analysis**. New York : J. Wiley, 1981. 504f.

MORGANO, M. A; FARIA, C.G.; FERRÃO, M.F.; BRAGAGNOLO, N.; FERREIRA, M.M.C. Determinação de umidade em café cru usando espectroscopia NIR e regressão multivariada. **Ciência Tecnologia Alimentos**, v. 28, n. 1, p. 12-17, 2008.

OLIVEIRA, F. C. **Um método para seleção de atributos em dados genômicos**. 2015. 273 f. Tese (Doutorado em Modelagem Computacional) - Universidade Federal de Juiz de Fora, Juiz de Fora, MG, 2015a.

OLIVEIRA, I. R. N. **Aplicação de espectroscopia de infravermelho e métodos multivariados para avaliação de características funcionais em extratos de**

antocianinas. 2015. 132 f. Tese (Doutorado em Ciência e Tecnologia de Alimentos) - Universidade Federal de Viçosa, Viçosa, MG, 2015b.

OLIVEIRA, I. R.; ROQUE, J. V.; MAIA, M. P.; STRINGHETA, P. C.; TEÓFILO, R. F. New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, 2018c.

PASQUINI, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, p. 198 219, 2003.

PIEPHO, H. P. Ridge regression and extensions for genomewide selection in maize. **Crop Science**, v. 49, n. 4, p. 1165-1176. 2009.

PIRES, M. P. **Comparação entre modelos de análise genômica utilizando dados simulados e dados reais em ovinos**. 2015. 71 f. Tese (Doutorado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinárias – Unesp, Jaboticabal, SP, 2015.

R Core Team (2017). **R: A language and environment for statistical computing. R Foundation for Statistical Computing**, Vienna, Austria. Disponível em: http://www.R-project.org/. Acesso em: Nov. 2017.

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, n. 56, p. 63, 2008.

RESENDE, M. D.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística, 291f., 2012. Disponível em: < http://www.cnpso.embrapa.br/gws/selecao_genomica.pdf >. Acesso em Nov. 2017.

RESENDE, M. D. V.; SILVA, F.F.; VIANA, J.M.S.; PETERNELLI, L.A.; JÚNIOR, M.F.R.R.; VALLE, P.M.D. **Métodos estatísticos na seleção genômica ampla**. 1 ed. Colombo, PR, Embrapa Florestas, 2011. 107f.

ROQUE, J. V. **Desenvolvimento de modelos de regressão multivariada para determinação de ésteres de forbol em sementes de jatropha curcas L. usando espectroscopia e quimiometria.** 2015. 84 f. Dissertação (Mestrado em Agroquímica) - Universidade Federal de Viçosa, Viçosa, 2015.

SILVA, D. A.; ALMEIDA, V. C.; VIANA, L. C.; KLOCK, U.; MUÑIZ, G. I. B. Avaliação das propriedades energéticas de resíduos de madeiras tropicais com uso da espectroscopia NIR. **Floresta e Ambiente**, Seropédica, v. 21, n. 4, p. 561-568, 2014.

SOUSA, L. C.; GOMIDE, J. L.; MILAGRES, F. R.; ALMEIDA, D. P. Desenvolvimento de modelos de calibração NIRS para minimização das análises de madeiras de Eucalyptus spp. **Ciência Florestal**, v. 21, n. 3, p. 591-599, 2011.

SOUZA, J.S; FERRÃO, M.F. Aplicações da espectroscopia no infravermelho no controle de qualidade de medicamentos contendo diclofenaco de potássio. Parte I: Dosagem por regressão multivariada **Revista Brasileira de Ciências Farmacêuticas,** V.42, n.3, p. 437-445, 2006.

SOUZA, A. M.; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: Um tutorial, parte I. **Quimica Nova**, v. 35, p. 223–229, 2012.

TEIXEIRA, F. R. F. **Análise de fatores aplicada na seleção genômica em suínos**. 2015. 76f. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, MG, 2015.

TEÓFILO, R. F. **Métodos quimiométricos em estudos eletroquímicos de fenóis sobre filmes de diamante dopado com boro**. 2007. 329 f: Tese (Doutorado em Química) - Universidade Estadual de Campinas, 2007.

TEÓFILO, R. F. **Métodos Quimiométricos: Uma Visão Geral- Conceitos básicos de quimiometria.** Viçosa: Universidade Federal de Viçosa/Departamento de Química, 118p., 2013. Disponível em: < http://www.deq.ufv.br/area/publicacao/26>. Acesso em: Nov. 2017.

TEÓFILO, R. F.; MARTINS, J. P. A; FERREIRA, M. M. C. 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.

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

VALDERRAMA, P.; BRAGA, J. W. B.; POPPI, R. J. P. Validation of Multivariate Calibration Models in the Determination of Sugar Cane Quality Parameters by Near Infrared Spectroscopy. **J. Braz. Chem. Soc**. v. 18, n. 2, p. 259-266, 2007.

VIGNAL, A.; Milan, D.; SanCristobal, M.; Eggen, A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genetics Selection Evolution**, v. 34, n. 3, p. 275, 2002.

XU, L.; ZHOU, Y.; TANG, L.; WU, H.; JIANG, J.; SHEN, G.; YU, R.; Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration. **Analytica Chimica Acta**, v. 616, p. 138–143, 2008.

WALDMANN, P.; MÉSZÁROS, G.; GREDLER, B.; FUERST, C.; SÖLKNER, J. Evaluation of the lasso and the elastic net in genome-wide association studies. **Frontiers in genetics**, v. 4, p. 270, 2013.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: A basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**. v.58, p.109-130, 2001.

ZIMMER, J.; ANZANELLO, M. J. Um novo método para seleção de variáveis preditivas com base em índices de importância. **Production.** Porto Alegre, RS. v. 24, n. 1, p. 84-93., 2014.