PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS

PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA

PROGRAMA DE MESTRADO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS

PARTICIONAMENTO DE CONJUNTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMAS DE CALIBRAÇÃO MULTIVARIADA

André Luiz Alves

2017

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS

PROGRAMA DE MESTRADO EM ENGENHARIA DE PRODUÇÃO E SISTEMAS

PARTICIONAMENTO DE CONJUNTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMAS DE CALIBRAÇÃO MULTIVARIADA

ANDRÉ LUIZ ALVES

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para qualificação da pesquisa de Mestre em Engenharia de Produção e Sistemas.

Orientador: Clarimar José Coelho, Dr.

Coorientador: Gustavo Teodoro Laureano, Dr.

Goiânia

Setembro, 2017

A474p Alves, André Luiz
Particionamento de conjunto de dados e seleção de variáveis em problemas de calibração multivariadas[manuscrito]/ André Luiz Alves.-- 2017.
49 f.; 30 cm

Texto em português com resumo em inglês
Dissertação (mestrado) - Pontifícia Universidade Católica de Goiás, Programa de Pós-Graduação Stricto Sensu em Engenharia de Produção e Sistemas, Goiânia, 2017

Inclui referências f. 47-49

1. Computação - Matemática. 2. Álgoritmos. 3. Computação. 4. Calibração Multivariada. I.Coelho, Clarimar José. II.Pontifícia Universidade Católica de Goiás. III. Título.

CDU: 004.421(043)

PARTICIONAMENTO DE CONJUNTO DE DADOS E SELEÇÃO DE VARIÁVEIS EM PROBLEMAS DE CALIBRAÇÃO MULTIVARIADA

ANDRÉ LUIZ ALVES

Esta Dissertação julgada adequada para obtenção do título de Mestre em Engenharia de Produção e Sistemas, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás em setembro de 2017.

Prof. Marcos Lajovic Carneiro, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia de Produção e Sistemas

Banca Examinadora:

Prof. Clarimar José Coelho, Dr.
Orientador

Prof. Anderson da Silva Soares, Dr.

Profa. Carmen Cecilia Centeno, Dra.

Goiânia – Goiás
Setembro de 2017

À memória de meus pais, Agenor Cosme Alves e Magnólia Gomes Barbosa Alves, exemplos de amor, humildade e superação.

AGRADECIMENTOS

À Deus.

Aos meus pais pelas bases de um lar pleno de amor, respeito, carinho e trabalho.

À minha esposa pelo apoio, companheirismo e compreensão pelas ausências durante esta etapa.

Aos meus filhos por a cada dia renovarem em mim o desejo de ir adiante.

Ao professor Dr. Clarimar José Coelho pela oportunidade, paciência, orientação e empenho imprescindíveis para a realização deste trabalho.

Ao professor Dr. Gustavo Teodoro Laureano pela coorientação.

Ao professor Dr. Igor Savioli Flores pelas valiosas contribuições.

Ao professor Me. Arlindo Galvão Filho pela ajuda com o Algoritmo das Projeções Sucessivas.

Ao professor Me. Daniel Vitor de Lucena pela ajuda, amizade e incentivo.

Ao colega Agamenon Lima do Vale pela amizade e ajuda com o algoritmo Ransac.

Ao colega Marcilon Fonseca de Lima pela amizade e apoio.

Aos demais amigos do Programa de Mestrado em Engenharia de Produção e Sistemas pelo companheirismo.

Aos professores do Programa, em especial à professora Maria José Pereira Dantas pelo incentivo, apoio e participação nos artigos elaborados.

À Fundação de Amparo à Pesquisa do Estado de Goiás pela bolsa concedida.

“Se soubesse o que estou fazendo, não seria pesquisa.”

Albert Einstein

RESUMO

O objetivo do trabalho é comparar um algoritmo proposto baseado no método consenso de amostra aleatória (RANdom SAmple Consensus, RANSAC) para seleção de amostras, seleção de variáveis e seleção simultânea de amostras e variáveis com o algoritmo de projeções sucessivas (Sucessive Projections Algorithm, SPA) a partir de conjuntos de dados químicos no contexto da calibração multivariada. O método proposto é baseado no método RANSAC e regressão linear múltipla (Multiple Linear Regression, MLR). A capacidade preditiva dos modelos é medida empregando o erro de previsão da raiz quadrada do erro quadrático médio (Root Mean Square Error Of Prediction, RMSEP). Os resultados permitem concluir que o Algoritmo das Projeções Sucessivas melhora a capacidade preditiva do Ransac. Conclui-se que o SPA influi positivamente no algoritmo Ransac para seleção de amostras, para seleção de variáveis e também para seleção simultânea de amostras e variáveis.

Palavras-chave: Particionamento de dados, Seleção de amostras e de variáveis, Calibração multivariada.

ABSTRACT

The objective of this work is to compare a proposed algorithm based on the RANdom SAmple Consensus (RANSAC) method for selection of samples, selection of variables and simultaneous selection of samples and variables with the Sucessive Projections Algorithm (SPA) from a chemical data set in the context of multivariate calibration. The proposed method is based on the RANSAC method and Multiple Linear Regression (MLR). The predictive capacity of the models is measured using the Root Mean Square Error of Prediction (RMSEP). The results allow to conclude that the Successive Projection Algorithm improves the predictive capacity of Ransac. It is concluded that the SPA positively influences the Ransac algorithm for selection of samples, for selection of variables and also for simultaneous selection of samples and variables.

Keywords: Data partitioning, Sample selection and variable selection, Multivariate calibration.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

APS	Algoritmo das Projeções Sucessivas
BLR	*Bayesian Linear Regression*
GA	*Genetic Algorithm*
GB	*Gigabyte*
GUI	*Graphical User Interface*
HD	*Hard Disk*
KS	Algoritmo de Kennard-Stone
LDP	*Location Determination Problem*
MLR	*Multiple Linear Regression*
MSEP	*Mean Square Error of Prediction*
NIR	*Near Infrared*
PCA	*Principal Component Analisys*
PRESS	*Predicted Residual Error Sum of Squares*
RAM	*Random Access Memory*
Ransac	*Random Sample Consensus*
RLM	Regressão Linear Múltipla
RMSE	*Root Mean Square Error*
RMSECV	*Root Mean Square Error of Cross Validation*
RMSEP	*Root Mean Square Error of Prediction*
SPA	*Successive Projections Algorithm*
Tb	*Terabyte*
VIS-NIR	*Visible Near Infra Red*

SUMÁRIO

1. INTRODUÇÃO

O particionamento de dados consiste em subdividir um conjunto de dados, de acordo com critérios estabelecidos, para uma finalidade específica. Uma das razões para se subdividir dados em subconjuntos reside na necessidade de se separar parte destes dados para a obtenção de modelos matemáticos que os representem. Na calibração multivariada, uma parte é destinada à calibração, outra é destinada à validação e, outra destinada a fazer predições.

A calibração multivariada tem sido amplamente utilizada na quimiometria, contexto no qual este trabalho se insere. Quimiometria é a área da química analítica que utiliza matemática, estatística e lógica formal para projetar ou selecionar procedimentos experimentais de maneira otimizada, prover o máximo de informação química relevante através da análise de dados químicos e obter conhecimento sobre sistemas químicos. Suas principais áreas de aplicação incluem calibração, validação e teste de significância (HOPKE, 2003; GEMPERLINE, 2006).

A calibração multivariada compreende um conjunto de técnicas com o propósito de medir, explicar e prever o grau de relacionamento entre as variáveis estatísticas resultantes da obtenção de dados em experimentos, possibilitando a obtenção de modelos matemáticos que relacionam variáveis independentes a variáveis dependentes, com o objetivo de se predizer uma determinada informação ou grandeza, por exemplo, a determinação da concentração de uma ou várias substâncias presentes em uma amostra (HAIR, 2009).

Problemas em calibração multivariada envolvem matrizes de dados complexas que refletem a natureza das amostras reais. As variáveis independentes são informações obtidas através de instrumentos utilizadas para construção de modelos matemáticos e são organizadas no formato matricial. Os objetivos de investigação na análise multivariada incluem redução de dados ou simplificação estrutural, classificação e agrupamento de objetos similares, investigação das dependências entre variáveis, predição, e formulação e teste de hipóteses (JOHNSON, 2014). Além da predição numérica real da propriedade procurada (y), a calibração fornece vários parâmetros informativos que podem ser utilizados de forma exploratória para investigar a validade do modelo e melhorá-lo (BRO, 2003).

No entanto, os modelos obtidos através da aplicação das técnicas da análise multivariada, estão intrinsicamente relacionados à qualidade dos dados (GEMPERLINE, 2006). Os dados estão sujeitos aos erros humanos, relacionados à aquisição, à coleta, ao preparo, ao armazenamento e ao manejo de amostras; aos erros de método, que aparecem como consequência do comportamento físico ou químico não ideal das substâncias nas quais se baseia a análise; e também aos erros instrumentais, decorrentes de imprecisão dos instrumentos (SKOOG, 2017).

Wold (1995) afirma que independentemente do que se meça, a medida é imprecisa. Qualquer dado, qualquer medida tem alguma variabilidade ou ruído que se deve a, pelo menos, três causas: falta de completo controle das condições experimentais que torna impossível manter exatamente as mesmas condições na segunda vez que se mede alguma coisa ou a segunda vez que se realiza o experimento (fatores de ruído); a instabilidade do instrumento de medida, que pode produzir um valor diferente da segunda vez que se mede a mesma coisa (fatores ambientais); erros de modelo, que são decorrentes das simplificações e aproximações dos modelos científicos e que são usados para especificar o que se espera das medidas.

Estas possíveis causas podem levar à presença de amostras que se configuram como observações discrepantes – *outliers* – no conjunto de dados. São elementos que parecem desviar notoriamente dos demais membros da amostra onde ocorrem. Que apresentam características substancialmente diferentes do conjunto e com pequenas probabilidades de se encaixarem em relação à distribuição de probabilidade verificada (HUBER, 2009). Caso sejam conhecidas as razões físicas para a observação discrepante, recomenda-se rejeitá-la, corrigi-la com base no seu ambiente original ou ainda, rejeitá-la e, se possível, tomar uma observação adicional. Se as razões físicas para a observação discrepante são desconhecidas, usar teste estatístico, e dependendo do resultado, além das opções citadas, empregar algum método de seleção de amostra para descartar as observações discrepantes presentes (GRUBS, 1969).

Selecionar amostras constitui-se em um fator primordial para a obtenção de modelos robustos. O termo “robusto” significa ser não sensível a pequenos desvios decorrentes das pressuposições adotadas (HUBER, 2009). Porém, a adequada seleção de amostras contempla somente parte do problema da qualidade dos dados utilizados na concepção dos modelos.

Tão impactante quanto a presença de dados discrepantes é o fato de algumas variáveis representarem a mesma informação por existir uma forte relação entre elas. Quando tal relação envolve duas características (variáveis preditoras), é denominada colinearidade. Quando envolve mais de duas das respostas instrumentais, é denominada multicolinearidade. Independentemente de sua fonte ou natureza, é imprescindível adotar procedimentos para detectar sua existência, medir sua extensão e identificar sua localização e causas, pois comprometem a robustez dos modelos (FARRAR, 1967).

Multicolinearidade é uma questão de grau, não de presença ou ausência. As razões para se utilizar apenas algumas das variáveis preditoras disponíveis são: estimar ou prever a menor custo, reduzindo o número de variáveis nas quais os dados são coletados; prever com precisão eliminando variáveis que não contém informações além das já consideradas; para descrever um conjunto de dados multivariados de forma parcimoniosa; e para estimar os coeficientes de regressão com pequenos erros, particularmente quando algumas das variáveis preditoras estão altamente correlacionadas (MILLER, 1984). Porém, a eliminação de uma das variáveis que sejam linearmente dependentes requer cuidado, pois dependendo do poder relativo de explicação que a variável independente eliminada tem sobre a variável dependente, o poder preditivo do modelo pode ser comprometido (PAUL, 2006). Se o modelo será usado para fazer predições, economiza-se tempo e recursos não medindo variáveis preditoras redundantes. As razões são óbvias: entre várias formas plausíveis de se explicar um fenômeno, a mais simples é a melhor (SILVEY, 1969).

Assim como a presença de dados discrepantes, a incidência da multicolinearidade tem motivado pesquisadores na busca por soluções que minimizem os efeitos de dados indesejados (GALVÃO, 2008; GALVÃO 2009; GRUBS, 1969; HOCKING, 1976; MILLER, 1984; NUNES, 2008).

Para seleção de amostras, métodos consagrados como Amostragem Aleatória *(Random Samplig*, RS), *Kennard-Sotne* (KS) (KENNARD, 1969), melhorias têm sido avaliadas e evoluções têm sido propostas como o algoritmo SPXY (*Sample set Partitioning based on joint x – y distances*) (GALVÃO, 2005) que consiste em considerar também a distância na variável dependente (y). SAPTORO *et. al* (2012) propuseram o método MDKS (*Mahalanobis Distance Kennard Stone*) que além de incorporar a influência de (y), adota a distância de Mahalanobis ao invés da euclidiana.

O algoritmo Ransac, devido à sua característica de identificar *outliers*, se posiciona com promissora alternativa para a seleção de amostras.

A seleção de variáveis tem sido abordada por meio de várias técnicas, como Análise de Componentes Principais (*Principal Component Analisys, PCA*) (SWINIARSKI, 2003), Algoritmos Genéticos (*Genetic Algorithm*, GA) (LEARDI, 2000), seleção por etapas (*stepwise selection*) e seleção à frente (*forward selection*) (BLANCHET, 2008), Regressão Linear Bayesiana (*Bayesian Linear Regression, BLR*) (CHEN, 2009), Algoritmo das Projeções Sucessivas (GALVÃO *et. al*, 2001), análise de combinação variável de população (YUN, 2014), entre outras.

Tanto para seleção de amostras como para seleção de variáveis, os métodos adotados têm em comum a forma de amostragem, seja no espaço das amostras ou no espaço das variáveis, o método de modelagem e a busca pela solução ótima ou pelo critério de avaliação.

Conhecer os desempenhos dos algoritmos Ransac (*RANdom SAmple Consensus*) e do Algoritmo das Projeções Sucessivas (APS) pode revelar resultados relevantes não somente para a seleção de amostras e seleção de variáveis, como também na seleção simultânea de amostras e variáveis.

Neste sentido, este trabalho aborda o problema da qualidade dos dados utilizados na concepção dos modelos em calibração multivariada mediante a comparação dos resultados da aplicação de implementações destes algoritmos.

Ambos os algoritmos são aplicados para seleção de amostras e para seleção de variáveis. O intuito é identificar e eliminar *outlliers* e também identificar e eliminar a multicolinearidade presentes no conjunto de dados, para que se obtenha o melhor modelo de regressão linear possível.

O objetivo geral consiste em avaliar os resultados obtidos mediante as aplicações dos algoritmos Ransac e APS, tendo como estudo de caso um conjunto de dados químicos obtidos por espectrofotometria na região do infravermelho próximo. Os sinais correspondem aos vários comprimentos de ondas, resultantes do processo de absorção pelo elemento analisado. Busca-se identificar, entre as exploradas, a melhor forma de se selecionar amostras que sejam isentas de dados discrepantes, bem como selecionar as variáveis mais representativas, que possibilite à obtenção de modelos robustos.

Para tanto, são formuladas as seguintes questões como propósito de se realizar a avaliação, segundo as opções experimentadas:

Questão 1: Qual a melhor opção para a seleção de amostras?

Questão 2: Qual a melhor opção para a seleção de variáveis?

Questão 3: Qual efeito o algoritmo APS provoca no Ransac?

O critério adotado para evidenciar o desempenho das várias opções de execução dos algoritmos e para servir de análise comparativa dos resultados é o RMSEP – Raiz do erro quadrático médio de predição (*Root Mean Square Error of Prediction*).

Na sequência, esta dissertação apresenta no Capítulo 2 os Materiais e Métodos, no Capítulo 3 os Resultados e Discussões, e finalmente, no Capítulo 4 as Conclusões.

2. MATERIAIS E MÉTODOS

2.1. Material Experimental

Os dados utilizados neste trabalho foram obtidos através da espectrofotometria, que mede a interação entre o objeto em análise e a energia irradiada. A amostra recebe a radiação e a energia absorvida pode ser medida pelo espectrofotômetro e relacionada com a concentração da propriedade de interesse. A concentração da amostra inteira é obtida mediante a irradiação com diferentes comprimentos de onda simultaneamente segundo a lei de Lambert-Beer, representada pela equação (1), onde $P_0(\lambda)$ representa a radiação emitida pelo equipamento e $P(\lambda)$ é a radiação emitida pela amostra no comprimento de onda λ. (SKOOG, 2017).

$$\mathbf{x(\lambda)} = \mathbf{log}\ \frac{\mathbf{P_0(\lambda)}}{\mathbf{P(\lambda)}} \tag{1}$$

A Figura 1 representa esquematicamente o processo de obtenção dos dados a partir das amostras. Após serem tratados por um modelo matemático permitem que sejam geradas respostas sobre a propriedade de interesse.

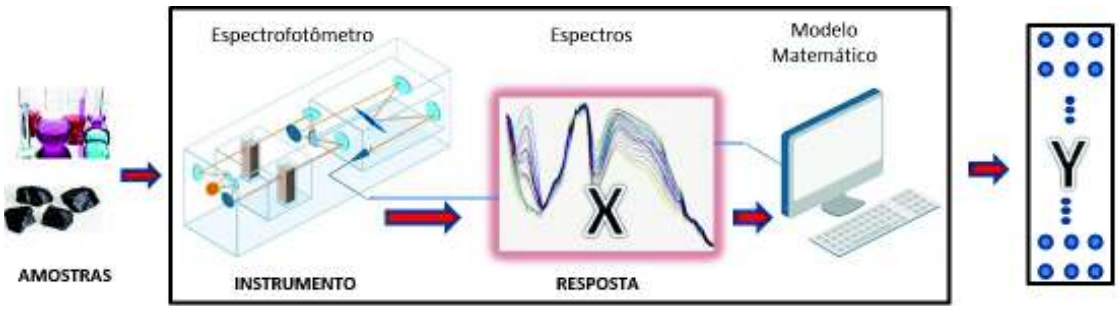


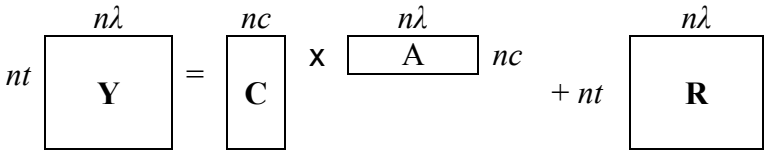
Figura 1 - Representação esquemática do processo de obtenção dos dados de respostas instrumentais por meio de espectrofotometria.

Em procedimentos dessa natureza, é comum a sobreposição de comprimentos de ondas e, consequentemente, dois ou mais sinais podem enviar a mesma informação, significando alta correlação entre as variáveis, ocasionando problemas matemáticos no processo de regressão (SOARES *et. al*, 2013).

A lei de Lambert-Beer, equação (2) estabelece que a absorção total, y_λ, de uma solução em um determinado comprimento de onda, λ, é a soma de todas as contribuições de espécies absorventes dissolvidas A, B, ..., Z, com absorções molares $\varepsilon_{A,\lambda}$, $\varepsilon_{B,\lambda}$,..., $\varepsilon_{Z,\lambda}$.

$$\boldsymbol{Y\lambda = [A]\ \varepsilon_{A,\lambda} + [B]\ \varepsilon_{B,\lambda} + \ldots + [Z]\ \varepsilon_{Z,\lambda}} \tag{2}$$

Considerando-se um espectro, a cada comprimento de onda corresponderá uma equação e o sistema de equações pode ser representado na forma matricial:



O espectro de absorção medido em $n\lambda$ comprimentos de onda, forma vetores de dimensão $n\lambda$ que são as linhas da matriz Y. Se *nt* espectros são medidos *nt* vezes, Y contém *nt* linhas de $n\lambda$ elementos. A estrutura da lei de Beer-Lambert e a lei matemática para multiplicação de matrizes são essencialmente idênticas, a matriz Y pode ser escrita como o produto das duas matrizes C e A, onde C contém as colunas dos perfis de concentração das espécies absorventes. Se existem *nc* espécies absorventes, C tem *nc* colunas, cada uma contendo *nt* elementos. De modo análogo, a matriz A contém, em *nc* linhas, as absorções molares das espécies absorventes, medidas em $n\lambda$ comprimentos de onda, que são os valores $\varepsilon_{X,\lambda}$ da equação N. Devido a imperfeições presentes em quaisquer medidas reais, o produto C x A não resulta exatamente em Y. A matriz R contém a diferença residual (GEMPERLINE, 2006).

Os dados de referência aqui utilizados foram determinados no Laboratório de Pesquisa de Grãos, em Winnipeg, com o uso de espectrofotômetro e se referem a amostras de grãos inteiros de trigo, oriundos da produção vegetal do ocidente canadense. As propriedades amostrais de referência são: a concentração de proteína (em %); teste de peso (em kg/hl); PSI (textura do grão de trigo) (em %); absorção de água por farinografia (em %), tempo de desenvolvimento de massa por farinografia (em minutos), e índice de tolerância à mistura por farinografia. O conjunto de dados para o estudo de calibração multivariada compreende 775 espectros na região visível do infravermelho próximo (*Near Infra Red)* – VIS-NIR, de amostras de todo o grão de trigo, que foram utilizados como dados *shoot-out* em 2008, na Conferência Internacional Refletância Difusa.

A propriedade de interesse escolhida é a concentração de proteína. Os espectros foram adquiridos na faixa de 400-2500nm, com uma resolução de 2nm, empregando-se a região NIR na faixa de 1100-2500nm. Com o propósito de contornar o problema das variações sistemáticas nos espectros derivados, foram calculadas a primeira derivada dos espectros usando um filtro Savitzky-Golay com um polinômio de segunda ordem e uma janela de 11-pontos. Savitzky-Golay tem sido amplamente utilizado para este fim

(DANTAS FILHO *et. al*, 2004; GALVÃO *et. al*, 2005; HONORATO *et. al*, 2005; CASALE *et. al*, 2010). Apenas os dados referentes à concentração de proteína foram usados.

Para a divisão do conjunto de dados destinados à calibração, validação e predição foi utilizado o algoritmo de Kennard-Stone – KS (KENNARD, 1969) em sua forma original, garantindo uma distribuição uniforme ao longo do espaço de dados X. Assim, grupos de dados para este trabalho têm a seguinte constituição:

- Para calibração (matriz X_{Cal}): 389 amostras, 690 variáveis;
- Para validação (matriz X_{Val}): 193 amostras, 690 variáveis;
- Para predição (matriz X_{Pred}) : 193 amostras, 690 variáveis.

2.2. Regressão Linear Múltipla

O primeiro componente do método empregado neste trabalho é a Regressão Linear Múltipla para execução do processo de calibração. É um método de análise apropriado quando o problema de pesquisa envolve uma variável dependente métrica (y) considerada relacionada a duas ou mais variáveis independentes também métricas (x). Trata-se de uma metodologia estatística para a predição de valores de uma variável reposta, a partir de uma coleção de valores de variáveis denominadas preditoras, permitindo assim avaliar os efeitos destas na variável dependente, ou seja, possibilita prever as mudanças na variável dependente como resposta a mudanças nas variáveis independentes. Quando o modelo é linear em se tratando de seus coeficientes, é chamado modelo de Regressão Linear Múltipla (*Multiple Linear Regression,* MLR) (HAIR *et. al*, 1998; GEMPERLINE, 2006; WALPOLE, 2012; JOHNSON, 2014).

A realização de um processo de calibração envolve três partes:

Primeiramente se estabelece o modelo de calibração, mediante o relacionamento da matriz de dados das variáveis obtidas no experimento, denominada matriz X, com a matriz de dados das propriedades de interesse, denominada matriz Y.

Em seguida, realiza-se a validação, na qual se avalia o modelo obtido. A validação pode ser realizada através do método denominado validação cruzada (*Cross Validation*) na qual as próprias amostras usadas na calibração são também usadas na validação, também chamada de validação interna. Outro método de validação, denominado validação externa, a adotada neste trabalho, utiliza um conjunto de dados

diferente dos empregados na calibração, porém com os valores para a propriedade (y) conhecidos (FERREIRA, 1999).

Finalmente realiza-se o processo de previsão mediante a aplicação do modelo obtido em amostras cuja concentração da propriedade de interesse se desconhece.

A regressão linear múltipla é formalmente representada pela equação (3):

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon \tag{3}$$

onde: Y representa a matriz da variável resposta, de ordem n x 1;

X representa a matriz de dados obtidos nas respostas instrumentais, de ordem n x p, ou seja, n observações (amostras) e p características (variáveis);

β representa o vetor de coeficientes de regressão, de ordem n x 1; e

ε representa o vetor de erro residual, de ordem n x 1.

Assim, **Y** se constitui em uma combinação linear das variáveis de **X**, e pode ser representada na seguinte forma matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{12} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Os coeficientes de regressão são calculados pela combinação linear por mínimos quadrados a partir da matriz pseudoinversa de **X**, equação (4):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\,\mathbf{X})^{-1}(\mathbf{X}'\,\mathbf{Y}) \tag{4}$$

O valor estimado para a variável desejada é obtido através da combinação linear entre a matriz de dados e os coeficientes de regressão estimados, equação (5):

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}} \tag{5}$$

E o erro residual é a diferença da variável reposta e seu valor estimado, equação (6):

$$\varepsilon = \mathbf{Y} - \hat{\mathbf{Y}} \tag{6}$$

Medidas de capacidade preditiva.

Dentre as várias estatísticas usadas para medir a capacidade preditiva de um modelo, destacam-se:

- A soma dos quadrados do erro de predição, PRESS, calculada pela equação (7), onde y_i é o valor real de y para o objeto i, $\hat{y}_i$ é o valor predito para o objeto i com o modelo em avaliação, e n é o número de objetos para os quais o valor de $\hat{y}$ é obtido pela predição:

$$\mathbf{PRESS} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_\mathbf{i})^\mathbf{2} \tag{7}$$

- O erro quadrático médio de predição, MSEP, definido como o valor médio de PRESS, como mostrado na equação (8):

$$\mathbf{MSEP} = \frac{\mathbf{PRESS}}{\mathbf{n}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_\mathbf{i})^\mathbf{2} \tag{8}$$

- Sua raiz quadrada é chamada de raiz do erro quadrático médio de predição, RMSEP, calculado pela equação (9):

$$\mathbf{RMSEP} = \sqrt{\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\hat{\mathbf{Y}}_\mathbf{i} - \mathbf{Y}_\mathbf{i})^\mathbf{2}} \tag{9}$$

Todas essas quantidades fornecem a mesma informação. Quanto menor o valor, melhor o modelo. Na literatura relativa à quimiometria, é possível observar que os valores de RMSEP são preferidos, provavelmente porque são expressos nas mesmas unidades que a variável y. Por este motivo, esta é a medida de capacidade preditiva adotada neste trabalho.

2.3. Ransac – *Random Sample Consensus*

O segundo componente do método empregado neste trabalho é o algoritmo Ransac. Técnicas clássicas para estimação de parâmetros, como as baseadas nos mínimos quadrados otimizam o ajuste do modelo a todos os dados presentes, porém, não têm mecanismos internos para detectar e rejeitar dados discrepantes (FISCHLER, 1981). O algoritmo Ransac, baseado no método do consenso de amostra aleatória, não apresenta este problema porque, ao contrário, considera o mínimo necessário de dados na obtenção dos modelos.

Concebido por Fischler e Bolles, se constitui em um método iterativo não determinístico de estimação de parâmetros, com o propósito de tratar dados empíricos

contaminados por *outliers* (FISCHLER, 1981). Foi proposto para a solução do problema de determinação da localização (*LDP – Location Determination Problem*), básico em análise de imagens, que consiste em estabelecer uma correspondência entre elementos de duas representações de uma dada cena. Aplicado tradicionalmente no contexto da Visão Computacional tem sido objeto de estudos e aplicações na pesquisa científica como no reconhecimento de ambientes internos, no deslocamento de robôs, no reconhecimento de íris, no reconhecimento de características da palma da mão, no monitoramento de ambientes por sensoriamento, na identificação de estruturas de construção, na identificação de falhas sísmicas, em diagnósticos por imagens na medicina, entre outras (RUZGIENE, 2005; JINGFU, 2007; CHIAMING, 2009).

A escolha do Ransac para ser aplicado neste trabalho se deve à sua capacidade de identificar e remover *outliers* em conjuntos de dados (RUZGIENE, 2005).

Os autores do Ransac exemplificaram sua proposta evidenciando o impacto que apenas um ponto discrepante causa em um conjunto composto por sete pontos, para o qual se pretende determinar os parâmetros do modelo linear que os representa. Com a utilização do método dos mínimos quadrados, o resultado é fortemente afetado, resultando em um modelo que destoa do conjunto de dados. A aplicação do Ransac resulta em um modelo que representa mais fielmente o conjunto de dados (FISCHLER, 1981).

Para melhor compreender a aderência do Ransac a este trabalho, detalha-se a seguir, as denominações a que tem sido referenciado: paradigma, método e algoritmo.

A designação paradigma foi utilizada pelos autores, pois caracteriza-se como um princípio, uma teoria, um conhecimento, originado da pesquisa em um campo científico. Uma referência inicial que servirá de modelo para novas pesquisas.

Em sua representação formal, o paradigma Ransac estabelece:

- Dado um modelo que requer um mínimo de **n** pontos para instanciar seus parâmetros, e um conjunto de pontos **P** tais que o número de pontos de **P** seja maior ou igual a **n**, deve-se selecionar aleatoriamente um subconjunto **S1** de n pontos de **P** e calcular o modelo **M1** para determinar o subconjunto **S1*** de pontos em **P** que estão dentro de uma tolerância de erro de **M1**.
- O subconjunto **S1*** é chamado consenso de **S1**.

- Se o tamanho de **S1*** é maior que o limite **t,** que é uma função do número estimado de erros grosseiros em **P**, deve-se usar **S1*** para computar (possivelmente usando mínimos quadrados) um novo modelo **M1***.
- Se o tamanho de **S1*** é menor que **t**, então se deve selecionar aleatoriamente um novo subconjunto **S2** e repetir o processo acima. Se, depois de um número predeterminado de tentativas, nenhum subconjunto consenso com **t** ou mais membros for encontrado, resolver o modelo com o maior consenso encontrado ou terminar em falha.

Os autores ressaltam que existem duas melhorias a serem consideradas:

1. Se há problema justificado em relação à seleção aleatória de pontos para formar **S's**, neste caso deve-se usar um processo determinístico de seleção.
2. Uma vez um consenso adequado de **S*** tenha sido encontrado e um modelo **M*** instanciado, adicionar novos pontos de **P** que sejam consistentes com **M*** para obter um novo **S*** e calcular um novo modelo com base neste subconjunto maior.

O paradigma contém três parâmetros não especificados:

1. O erro de tolerância usado para determinar se um ponto é compatível ou não com o modelo,
2. O número de subconjuntos a serem tentados, e
3. O limite **t**, que é o número de pontos compatíveis usados que implique que o modelo correto foi encontrado.

A designação método se justifica por se constituir de um conjunto das normas básicas que devem ser seguidas para a produção de conhecimentos, como requer o rigor da ciência.

A Figura 2 a seguir ilustra o método no caso de um modelo linear:

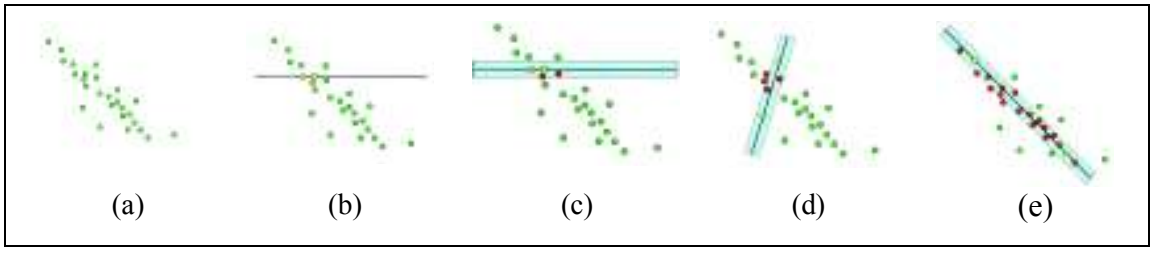


Figura 2 - Exemplo da lógica do método Ransac em um conjunto de pontos.

Nela, se pode observar em:

(a) os pontos que representam os dados;
(b) o modelo resultante de uma escolha aleatória de 2 pontos;
(c) representação dos pontos considerados *inliers* para o modelo obtido em (b);
(d) os pontos considerados uma nova tentativa;
(e) o modelo de consenso com o maior número de *inliers* obtido.

O termo *inlier* referre-se à "boa" observação, em contraposição ao termo *outlier*.

A designação algoritmo tem sido amplamente utilizada nos relatos científicos. Isto por ser constituído por um conjunto de regras e procedimentos lógicos, rigorosamente definidos, que levam à solução de um problema em um número finito de etapas, tornando-o facilmente implementável em linguagens de programação computacional.

O Quadro 1 a seguir apresenta o algoritmo Ransac:

```
Sejam:
    S – menor número requerido de pontos;
    N – número requerido de iterações;
    d – limite usado para identificar inliers;
    T – número requerido de inliers para que o modelo seja considerado bom.
Até que N iterações tenham ocorrido
    Tome uma amostra uniformemente aleatória dos dados de S pontos
    Defina o modelo para os S pontos
    Para cada ponto fora da amostra S
        Compare se a distância entre o ponto e a linha em relação ao valor d,
                Se é menor
                    então o ponto é considerado próximo
        fim
        Se existem T ou mais pontos próximos à linha
                então este é um bom modelo
                    recalcule a linha considerando todos estes pontos
fim
Selecione a melhor linha da coleção, usando o erro assumido como critério.
```

Quadro 1 - Representação do algoritmo Ransac.

O Ransac é conhecido como um método de duas fases:

- Geração de hipótese: fase na qual são escolhidos aleatoriamente o número mínimo de elementos que definem o modelo a ser determinado (dois, no caso de modelo linear, como é o caso deste trabalho);
- Teste de hipótese: fase na qual os coeficientes do modelo obtido na fase anterior são considerados para identificação dos demais elementos do conjunto de dados que são aderentes ao modelo, e, com isso, testar se o modelo obtido é o consenso.

Por ser um algoritmo não determinístico, a quantidade de amostragens necessárias para o Ransac é variável mesmo sendo executado para um mesmo conjunto de dados. Porém, é matematicamente compreensível que esta quantidade, ou número de tentativas é baixo, conforme argumentado pelos seus autores.

Considerando:

e – probabilidade do ponto ser um *outlier;*

s – número de pontos na amostra;

N – número de amostras;

p – probabilidade desejada para que se obtenha uma boa amostra.

Os autores do Ransac mostram que desenvolvendo a equação (10), se obtém a equação (11), que determina o número N.

$$\mathbf{1 - (1 - (1 - e)^{s})^{N} = p} \qquad (10)$$

O Quadro 2 apresenta passo a passo o desenvolvimento da equação (10), para que se obtenha uma amostra isenta de *outliers*.

$\mathbf{(1 - e)}$
É a probabilidade de escolher um ponto que seja um *inlier.*
$\mathbf{(1 - e)^{s}}$
É a probabilidade de escolher S *inliers* na linha. A amostra contém somente *inliers*.
$\mathbf{1 - (1 - e)^{s}}$

É a probabilidade de que um ou mais pontos na amostra sejam *outliers*.
A amostra é contaminada.

$$(1 - (1 - e)^{s})^{N}$$

É a probabilidade de N amostras serem contaminadas.

$$1 - (1 - (1 - e)^{s})^{N}$$

É a probabilidade de que pelo menos uma amostra seja não contaminada.
Pelo menos uma amostra de **S** pontos é composta somente de *inliers*.

Quadro 2 - Quantidade necessária de amostragens na execução do Ransac.

O raciocínio é simples, porém se sustenta em uma lógica que parte do pressuposto estatístico: a de que a probabilidade de se escolher um *inlier* é igual a um menos a probabilidade de se escolher um *outlier*. Seguindo o mesmo raciocínio, a probabilidade de se escolher um subconjunto de amostras que somente contenha *inliers,* é igual a um menos a probabilidade de se escolher um subconjunto de amostra que contenha pelo menos um *outlier* (subconjunto contaminado).

Assim, a equação (11) determina este número:

$$N = \frac{\log(1 - p)}{\log(1 - (1 - e)^{s})} \tag{11}$$

Para ilustrar a eficácia do paradigma Ransac, os autores apresentaram um exemplo para determinar o número de amostras **N**, considerando a probabilidade **p**, de que pelo menos uma amostra aleatória seja livre de *outlier*s.

O exemplo está ilustrado na Figura 3 a seguir, na qual dos 12 pontos do conjunto de dados, 2 são discrepantes:

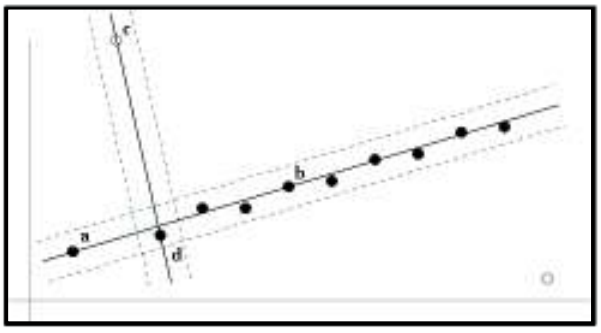


Figura 3 - Exemplo de aplicação do Ransac em um conjunto de dados de 12 pontos (FISCHLER, 1981).

Estabelecendo que a probabilidade desejada para que se obtenha pelo menos uma boa amostra seja de 99%, define-se **p** = 0.99.

Para determinar N, aplica-se a equação (11),

$$\mathbf{N} = \frac{\mathbf{log\,(1-p)}}{\mathbf{log\,(1-(1-e)^{s})}}$$, com os valores de **s** e **e**:

s = 2, que define o tamanho mínimo da amostra, pois se trata de uma reta, e

e = 20%, pois a probabilidade de um ponto ser um *outlier* neste exemplo, é de 2 em 12, ou seja, **e** = 2/12 = 0,166 ~ 0,2 = 20%.

A Figura 4 mostra a tabela utilizada pelos autores do Ransac. Para o exemplo, observando a interseção entre a primeira linha (**S** = 2), terceira coluna (**e** = 20%), evidencia-se que com 5 tentativas se chega ao modelo linear desejado.

	proportion of outliers *e*						
s	5%	10%	20%	25%	30%	40%	50%
2	2	3	5	6	7	11	17
3	3	4	7	9	11	19	35
4	3	5	9	13	17	34	72
5	4	6	12	17	26	57	146
6	4	7	16	24	37	97	293
7	4	8	20	33	54	163	588
8	5	9	26	44	78	272	1177

Figura 4 - Quantidade de amostras em função do percentual de *outliers* (FISCHLER, 1981).

Ou seja, com cinco amostragens (N = 5), há 99% de probabilidade de se obter um modelo (linha) que contenha somente *inliers*.

Implementações do Algoritmo Ransac.

Versões específicas do algoritmo Ransac foram elaboradas para atender às opções, incorporando as adaptações necessárias conforme cada objetivo.

Os dados de entrada para execução para o Ransac são:

- Dados para calibração: matrizes Xcal e Ycal;
- Dados para validação: matrizes Xval e Yval;
- Parâmetros para execução.

Merecem destaques os parâmetros detalhados no Quadro 3 a seguir, para os quais se apresentam os valores assumidos (*default*), caso não sejam fornecidos.

Parâmetro	Descrição	Valor assumido
Sigma	Desvio padrão de ruído	0,05
P_inlier	Probabilidade de que um ponto cujo erro de ajuste seja menor ou igual a Sigma, ou seja, um *inlier*	0,99
Max_iters	Número máximo permitido de iterações	∞
MSS	Conjunto mínimo de amostras (*Minimal Sample Set*)	2

Quadro 3 - Destaque de parâmetros para execução do Ransac.

A Figura 5 apresenta em forma de fluxograma os principais blocos de procedimentos do algoritmo Ransac.

O primeiro bloco trata os valores informados para processamento, fazendo as devidas consistências, e realiza os procedimentos iniciais para execução.

Vale ressaltar que o parâmetro MSS é utilizado pelo algoritmo Ransac para indicar o número de elementos a selecionar em cada iteração, na fase de escolha aleatória para cálculo dos parâmetros β. Para curvas, o valor a ser utilizado é 3, pois três pontos definem uma curva. Em tese, não há limite para este valor. Como neste trabalho o modelo que se busca é linear, este parâmetro tem seu valor *default* igual 2.

O segundo bloco corresponde à fase de "geração de hipótese". O parâmetro MSS determina quantas amostras (linhas da matriz Xcal) são selecionados aleatoriamente a cada iteração. Os parâmetros de calibração β são calculados para a fase seguinte.

O terceiro bloco corresponde à fase de "teste de hipótese". Os parâmetros β, Sigma e P_inlier são utilizados para determinar quais amostras adicionais (linhas da matriz Xcal) constituirão o conjunto de consenso para a iteração.

A cada iteração o conjunto de consenso obtido é comparado com o melhor anteriormente identificado. Caso seja melhor, é executado o quarto e último bloco que atualiza a informação de saída denominada *Consensus_Set*, contendo as amostras (índice das linhas de Xcal) do conjunto de consenso obtido.

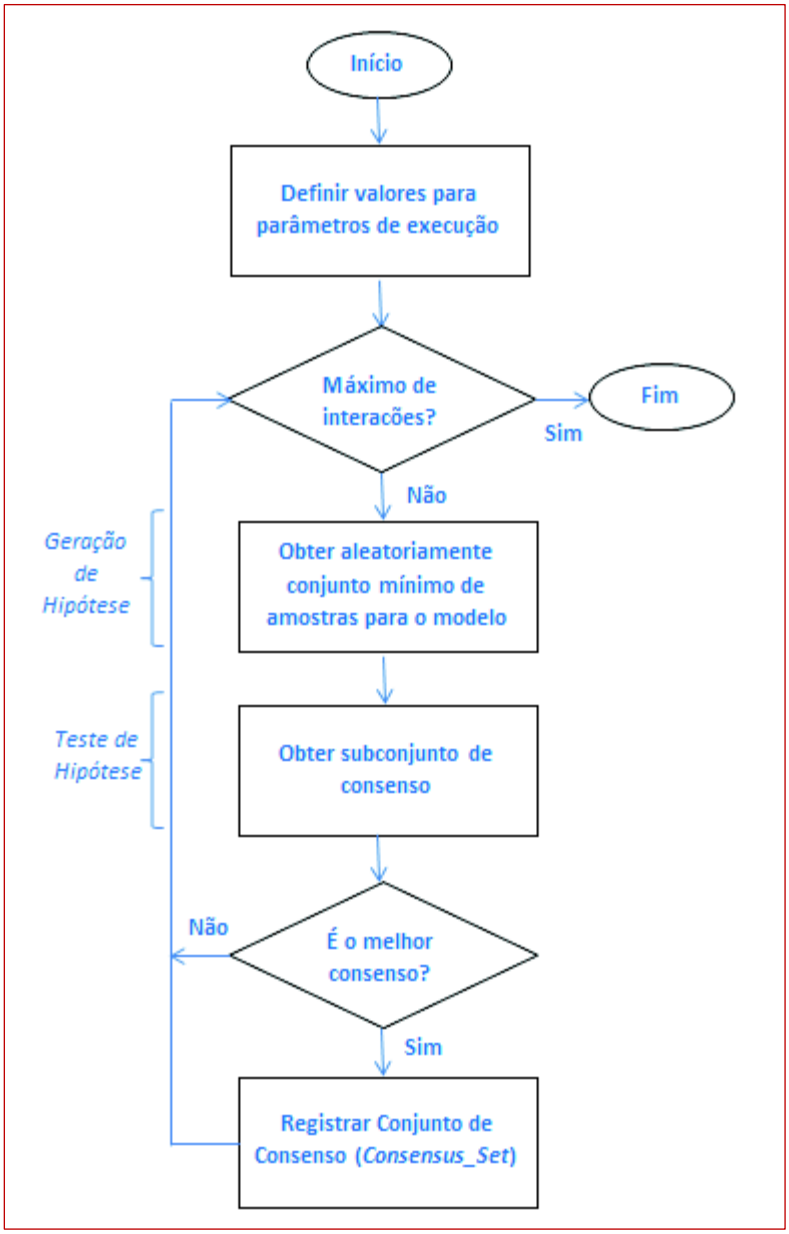


Figura 5 - Fluxograma básico para o algoritmo Ransac.

2.4. Algoritmo das Projeções Sucessivas

O terceiro componente do método empregado neste trabalho é o Algoritmo das Projeções Sucessivas (*Successive Projections Algorithm*, SPA), em razão dos resultados relatados em sua aplicação na seleção de variáveis. Desde sua concepção, tem sido aplicado em análise de multicomponentes espectroscópicos (ARAÚJO *et. al*, 2001), na melhoria da parcimônia em modelos de regressão linear múltipla (GALVÃO, 2008) e principalmente na seleção de variáveis em problemas de calibração multivariada aplicada às espectrometrias no infravermelho próximo (NUNES, 2008; SOARES *et. al*, 2010). Tem sido aplicado em problemas de Regressão Linear Múltipla juntamente com Regressão por Componentes Principais (GALVÃO *et. al*, 2001; DI NEZIO, 2007), com

Algoritmos Genéticos e Redes Neurais Artificiais (GOODARZI, 2009; GOOUDARZI *et. al*, 2009; DANTAS FILHO *et. al*, 2004), com Análise de Discriminante Linear em problemas de classificação (PONTES *et. al*, 2005) e também recebido modificações com propósitos de atuar na presença de interferências desconhecidas (SOARES *et. al*, 2011) e na abrangência de seus resultados (GALVÃO, 2008).

O Algoritmo das Projeções Sucessivas é uma técnica determinística de seleção de variáveis para minimizar problemas de colinearidade em regressão linear múltipla, concebido e proposto por Araújo e outros pesquisadores em 2001. Trata-se de uma técnica do tipo "passo à frente" (*step forward*) na qual dada uma variável inicial, a cada iteração, uma nova variável é inserida até que um número máximo determinado seja atingido. Sua principal finalidade é selecionar as variáveis que contenham o mínimo de redundância possível minimizando o problema de colinearidade, estruturado em três fases (ARAÚJO *et. al*, 2001, DANTAS FILHO *et. al*, 2004):

A primeira consiste em operações de projeção realizadas na matriz Xcal de respostas instrumentais, através da geração de cadeias de variáveis com um mínimo de redundâncias. Essas projeções são usadas para gerar cadeias de variáveis com cada vez mais elementos. Cada elemento de uma cadeia é selecionado de modo que apresente a menor colinearidade com a anterior.

Na fase seguinte, os subconjuntos de variáveis candidatas são avaliados de acordo com o desempenho preditivo RMSEP no modelo MLR, caso seja executado com a opção de validação externa. Caso a opção seja a validação interna, o desempenho preditivo é calculado utilizando-se o RMSECV. São assim avaliados os subconjuntos de variáveis extraídas a partir das cadeias geradas na primeira fase.

A terceira e última fase consiste no procedimento de eliminação de variáveis, em que, por meio de um teste estatístico, verifica-se se a eliminação de uma dada variável não compromete significativamente o erro RMSEP, com o propósito de melhorar a simplicidade do modelo.

O Quadro 4 detalha a sequência dos passos do Algoritmo da Projeções Sucessivas.

Definição inicial de parâmetros e variáveis:
k(n) → comprimento de onda *n* (da *n*-ésima iteração do APS).

X_{cal} → matriz $(M_{cal}$ x $J)$ de respostas instrumentais. M_{cal} são as amostras da calibração e J são os comprimentos de ondas.

Y_{cal} → matriz $(M_{cal}$ x $A)$ dos dados dos analitos. A são os analitos.

Y_{test} → matriz $(M_{test}$ x $A)$ para testes.

N → máximo de comprimentos de ondas para o APS. O primeiro comprimento de onda da primeira iteração é $k(0)$.

N^* → N ótimo encontrado. $k^*(0)$ é o comprimento de onda ótimo encontrado para iniciar as projeções no APS.

Iniciação (antes das projeções):

$n = 1$.

$k_j = j$-ésima coluna de X_{cal} ; $j = 1,..., J$.

Passo 1: S é o conjunto de comprimentos de onda ainda não selecionados.

$S = \{j \vee 1 \leq j \leq J e j \in \{k(0), \dots, k(n-1)\}\}$

Passo 2: Cálculo da Projeção de x_j no subespaço ortogonal para $x_{k(n-1)}$.

$$Px_j = x_j - (x_j^T x_{k(n-1)} x_{k(n-1)})(x_{k(n-1)}^T x_{k(n-1)})^{-1}; \forall j \in S$$

onde P é a projeção do operador.

Passo 3:

$$max(\|Px_j\|; j \in S)$$
$$k(n) = arg$$

Passo 4:

$$x_j = Px_j; j \in S$$

Passo 5:

$$n = n + 1; se\ n < N$$

O número de projeções no processo de seleção pode ser calculado como:

$(N\text{-}1)\ (J\text{-}N/2)$.

Para obter N^* e $k^*(0)$, quando não informados procede-se da seguinte forma:

Especifica-se um conjunto de amostras para validação;

Especifica-se Nmin e Nmax (de um intervalo onde N^* possa estar), sabendo-se que Nmin ≥ A e Nmax ≤ Mcal;

(A) Forma-se um laço de Nmin a Nmax

(B) Forma-se um laço variando de 1 a J

Usam-se os passos de 1 a 5 descritos acima

Constrói-se um modelo MLR de calibração com o

comprimento de onda selecionado

Usa-se o modelo para predizer a concentração do conjunto de validação

Calcula-se a raiz quadrada do erro quadrático médio (RMSE: *Root Mean Square Error*). Se validação interna, com:

$$RMSECV = \sqrt{\frac{1}{AM_{cal}} \sum_{i=1}^{M_{test}} \sum_{j=1}^{A} [\hat{Y}_{cal}(i,j) - Y_{cal(i,j)}]^2}$$

ou, se validação externa, com:

$$RMSEP = \sqrt{\frac{1}{AM_{test}} \sum_{i=1}^{M_{test}} \sum_{j=1}^{A} [\hat{Y}_{test}(i,j) - Y_{test(i,j)}]^2}$$

$$\rho(inicial) = RMSE$$

Volta-se ao laço (B) e incrementa J

$$r(N) = min[\rho(inicial)]; inicial = 1, \dots, J.$$

$$s(N) = arg[min\rho(inicial)]; inicial = 1, \dots, J.$$

Volta-se ao laço (A) e incrementa N

$$minr(N); N = N_{min}, \dots, N_{max}$$

$$N^* = arg$$

$$k^*(0) = s(N^*)$$

Quadro 4 - Passos do Algoritmo das Projeções Sucessivas. Adaptado de (ARAÚJO, 2001).

A Figura 6 representa esquematicamente o procedimento estabelecido pelo algoritmo das projeções sucessivas. Em (a) tem-se a primeira iteração na qual a Observação 1, corresponde a x_3, é o vetor de partida, formando a cadeia $\{x_3, x_1, x_4\}$. Em (b) tem-se a segunda iteração, na qual as projeções calculadas em relação ao subespaço ortogonal a z^1 são realizadas no subespaço ortogonal a z^2, resultando na cadeia $\{x_1, x_2, x_4\}$ (GALVÃO, 2009).

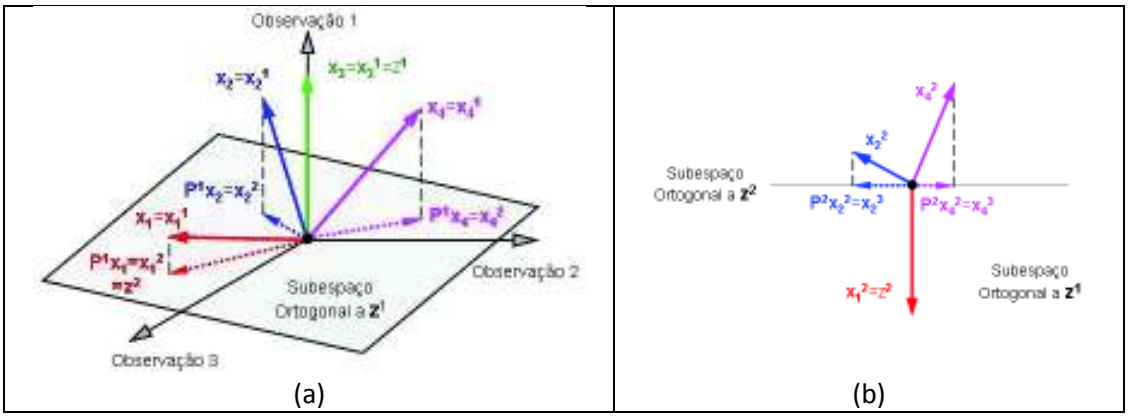


Figura 6 - Representação da sequência de projeções realizadas pelo APS (GALVÃO, 2009).

Implementação do Algoritmo das Projeções Sucessivas:

A Figura 7 apresenta em forma de fluxograma os principais blocos de procedimentos do Algoritmo das Projeções Sucessivas, aqui utilizado conforme sua proposição original.

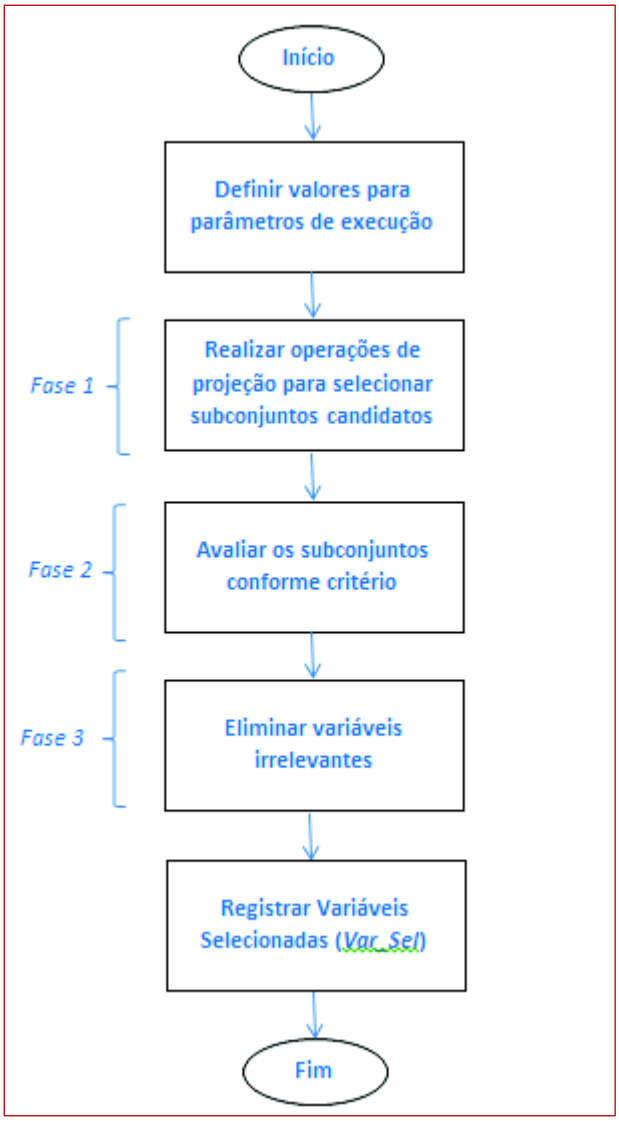


Figura 7 - Fluxograma básico para o Algoritmo das Projeções Sucessivas.

As entradas para execução para o Algoritmo das Projeções Sucessivas são:

- Dados para calibração: matrizes Xcal e Ycal;
- Dados para validação: matrizes Xval e Yval;
- Parâmetros para execução.

O primeiro bloco do fluxograma trata dos parâmetros para execução. Destacam-se os parâmetros detalhados no Quadro 5 a seguir, para os quais se apresentam os valores assumidos (*default*), caso não sejam fornecidos.

Parâmetro	**Descrição**	**Valor assumido**
m_min	Número mínimo de elementos na cadeia	nulo
m_max	Número mínimo de elementos na cadeia	nulo
autoscaling	Indicativo de escalonamento	nulo

Quadro 5 - Destaque de parâmetros para execução do APS.

O parâmetro *autoscaling* indica que antes que as operações de projeção sejam executadas sobre as colunas da matriz Xcal, se proceda à centralização na média, informando o valor "1," ou se faz o escalonamento, valor "0".

O segundo bloco corresponde à "Fase 1". Nesta fase são realizadas as projeções utilizando a função *qr,* suprida pelo Matlab®, para a decomposição ortogonal.

O terceiro bloco corresponde à "Fase 2", que avalia os subconjuntos candidatos gerados na fase anterior. Se há conjunto de dados destinado à validação, então o erro é calculado utilizando a validação externa. Caso contrário, é realizada a validação interna. Neste último caso (*cross validation*), em um laço que é executado envolvendo a cada interação, um dos elementos da matriz de calibração é separado para teste e o erro é calculado.

O quarto bloco corresponde à "Fase 3", que elimina das cadeias geradas os elementos considerados irrelevantes, mediante a aplicação do teste estatístico *F-test*, adequado quando se deseja comparar duas variâncias.

A informação de saída denominada *Var_Sel*, contém as variáveis (índice das colunas de Xcal) selecionadas pela execução.

2.5. Método

O método estabelecido para a realização deste trabalho compreende a aplicação dos algoritmos Ransac e APS em relação a seleção de amostras, seleção de variáveis, e seleção simultânea de amostras e variáveis, utilizando a Regressão Linear Múltipla para a calibração multivariada.

Seleção de Amostras:

Os propósitos das opções I, II e III são identificar qual algoritmo isoladamente resulta no melhor RMSEP e qual a influência do APS no algoritmo Ransac para seleção de amostras.

Opção I: Aplicação do algoritmo Ransac nas matrizes de dados originais para seleção de amostras.

Opção II: Aplicação do algoritmo APS nas matrizes de dados originais para seleção de amostras.

Opção III: Aplicação dos algoritmos Ransac e APS nas matrizes de dados originais para seleção de amostras.

Seleção de Variáveis.

Os propósitos das opções IV, V e VI são identificar qual algoritmo isoladamente resulta no melhor RMSEP e qual a influência do APS no algoritmo Ransac para seleção de variáveis.

Opção IV: Aplicação do algoritmo Ransac nas matrizes de dados originais para seleção de variáveis.

Opção V: Aplicação do algoritmo APS nas matrizes de dados originais para seleção de variáveis.

Opção VI: Aplicação dos algoritmos APS e Ransac nas matrizes de dados originais para seleção de variáveis.

Seleção simultânea de Amostras e Variáveis.

O propósito da opção VII é avaliar o resultado, segundo o RMSEP, de uma solução combinada do algoritmo Ransac com o algoritmo APS.

Opção VII: Aplicação dos algoritmos Ransac e APS de forma combinada nas matrizes de dados originais para seleção simultânea de amostras e variáveis.

Desta forma será possível avaliar o desempenho dos algoritmos Ransac e APS para seleção de amostras e variáveis, mediante a avaliação dos resultados segundo o RMSEP.

2.6. Software e Hardware

O software base utilizado para execução dos algoritmos é o Matlab R2016a®.

Os códigos fonte do Ransac aplicados neste trabalho foram desenvolvidos com base nos disponíveis no site do GitHub (https://github.com/RANSAC/RANSAC-Toolbox) e executados no modo de comandos do Matlab®.

Os códigos fonte do Algoritmo das Projeções Sucessivas utilizados foram desenvolvidos por pesquisadores do Instituto tecnológico da Aeronáutica e Universidade Federal da Paraíba (GALVÃO, 2008). Foram utilizadas as duas versões: uma que provê uma interface gráfica (*Graphical User Interface, GUI*) e outra com execução em modo de comandos, ambos no Matlab®.

O sistema operacional instalado no equipamento é o Microsoft Windows 2007®.

O hardware utilizado para os processamentos, computador DELL® Inspiron 14R, possui processador Intel® Core™ i5, 4 GB de memória RAM e HD de 1 Tb.

3. RESULTADOS E DISCUSSÕES

As execuções dos algoritmos Ransac e APS seguiram rigorosamente o método estabelecido no item 2.5. São relatados a seguir os resultados encontrados.

Seleção de amostras.

Opção I: Aplicação do algoritmo Ransac.

O algoritmo Ransac foi executado como apresentado no fluxograma da Figura 6, com os valores *default* descritos no Quadro 3, exceto para o número máximo de iterações:

- Max_iters = 10.000

Foram realizadas dez execuções desta Opção. Os resultados obtidos para o RMSEP foram:

- RMSEP = 1,299382 – Para o maior número de *inliers* obtido (359 de 390 amostras).
- **RMSEP = 1,256766** – Melhor valor obtido. Selecionadas 242 amostras.

Opção II: Aplicação do algoritmo APS.

A execução desta Opção resultou no seguinte erro apresentado pelo Matlab®:

"Warning: Rank deficiente, rank=3, tol=3,026168.10^{-12}"

Ao investigar a causa deste erro, verificou-se que o problema reside na obtenção de uma configuração com menor número de amostras que de variáveis. Dantas Filho *et. al.* (2004) relatam como realizaram a aplicação do APS para seleção de amostras. Segundo os autores, para evitar problemas de condicionamento da matriz de dados, deve-se realizar primeiramente a seleção de variáveis. Em seguida, transpor a matriz resultante e novamente executar o algoritmo APS, obtendo finalmente a seleção de amostras.

Por esta razão, a alternativa de se selecionar amostras diretamente pelo APS para fins de comparação com o Ransac, não será considerado neste trabalho. A execução exclusiva do APS para seleção de amostras, tendo como requisito sua execução prévia para seleção de variáveis, foge ao escopo deste trabalho.

Opção III: Aplicação dos algoritmos Ransac e APS.

Os algoritmos Ransac e APS foram executados como apresentado no fluxograma da Figura 8 a seguir. Os valores *default* descritos no Quadro 3 foram mantidos, exceto para o número máximo de iterações:

- Max_iters = 10.000

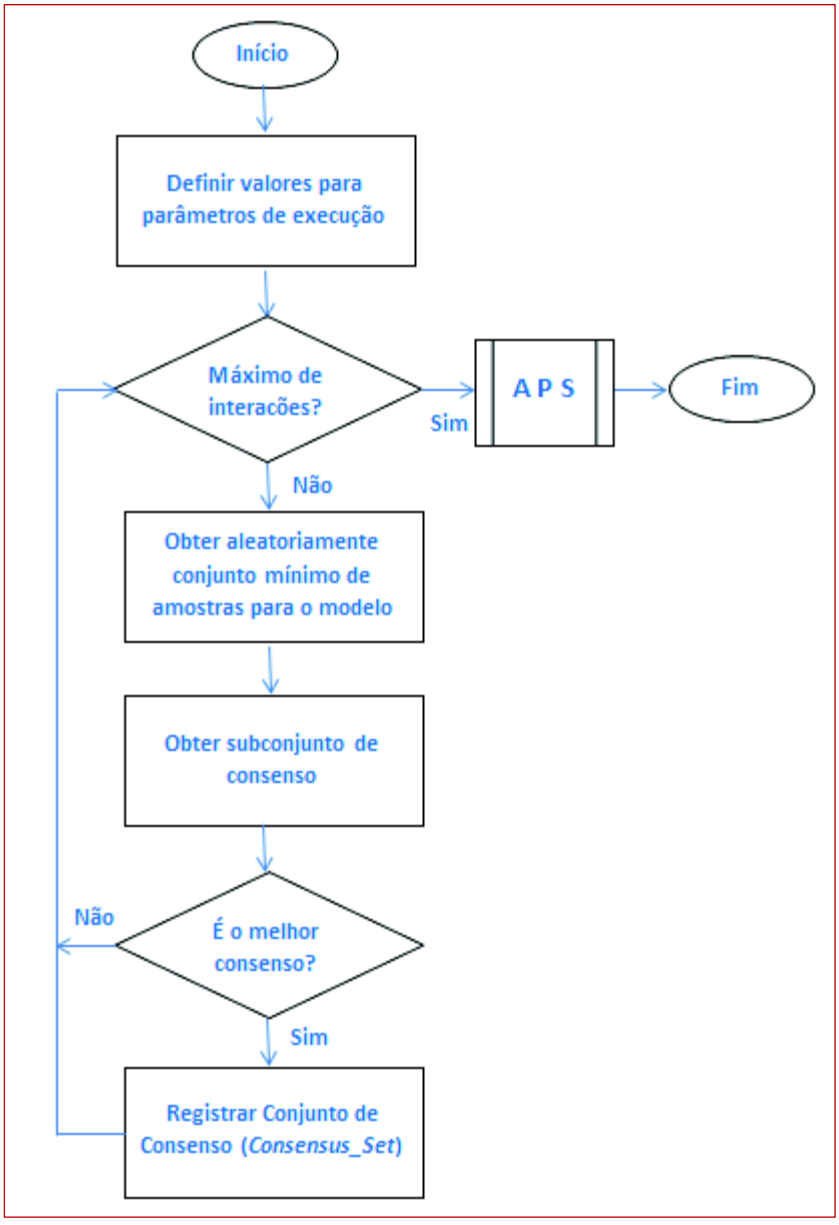


Figura 8 - Fluxograma de execução da Opção II.

A chamada ao Algoritmo das Projeções Sucessivas foi inserida no algoritmo Ransac após se chegar ao limite máximo de iterações.

Os resultados (parciais) obtidos para o RMSEP após a execução do Ransac, mas antes da execução do APS, foram:

- RMSEP = 1,299382 – Para o maior número de *inliers* obtido (359 de 390 amostras).
- RMSEP = 1,256766 – Melhor valor obtido. Selecionadas 242 amostras.

Na sequência, foi executado o algoritmo APS. A matriz de entrada Xcal para o APS foi a resultante do processamento do Ransac, ou seja, o contendo apenas as linhas selecionadas no conjunto de consenso (*Consensus_Set*).

Os resultados obtidos para o RMSEP após a execução do APS foram:

- **RMSEP = 0,23738** – Tendo sido selecionadas 365 amostras.

Foram realizadas dez execuções desta Opção.

Constatação 1:
Constata-se que para a seleção de amostras, a execução do APS após a execução do Ransac resultou no melhor RMSEP.

Quadro 6 - Constatação 1: Seleção de amostras.

Seleção de variáveis.

Opção IV: Aplicação do algoritmo Ransac.

O algoritmo Ransac foi executado como apresentado no fluxograma da Figura 6, com os valores *default* descritos no Quadro 3, exceto para o número máximo de iterações:

- Max_iters = 10.000

Porém, a versão para esta execução foi modificada na fase de geração de hipótese para selecionar colunas (variáveis) e não linhas (amostras), com é o padrão do Ransac.

Foram realizadas dez execuções desta Opção. O melhor resultado obtido para o RMSEP:

- **RMSEP = 1,307060** – Melhor valor obtido.

Opção V: Aplicação do algoritmo APS.

O algoritmo APS foi executado resultando na seleção de 58 variáveis. O valor do RMSEP obtido foi:

- **RMSEP = 0,22234**.

Neste ponto das execuções, uma nova possibilidade se apresentou, baseada nas seguintes considerações:

1. O APS, que é determinístico, selecionou 58 variáveis que não apresentam o problema da multicolinearidade;
2. O resultado de sua execução está armazenado na variável *Var_Sel;*
3. O Ransac pode ser executado com na Opção IV, alterando o tamanho do conjunto mínimo de elementos (MSS) de 2 (linear) para 58;
4. O resultado desta execução permitirá comparar os obtidos com a variável produzida pela execução do APS.

Assim, Foram realizadas dez execuções desta nova versão da Opção IV.

Nenhum dos conjuntos de consenso do Ransac (*Consensus_Set*) coincidiu com o conjunto de variáveis selecionadas do APS desta Opção V (*Var_Sel*).

Constatou-se também que não foi produzido nenhum RMSEP melhor que o obtido na execução original da Opção IV. O melhor resultado obtido para o RMSEP nesta nova tentativa foi:

- RMSEP = 1,319788 – Melhor valor obtido.

Constatação 2:
Constata-se que para a seleção de variáveis, a execução do APS resultou no melhor RMSEP.

Quadro 7 - Constatação 2: Seleção de variáveis: Ransac x APS.

Opção VI: Aplicação dos algoritmos APS e Ransac.

O objetivo desta Opção é identificar como o algoritmo APS influencia no algoritmo Ransac para seleção de variáveis.

Os algoritmos APS e Ransac foram executados como apresentado no fluxograma da Figura 9 a seguir. A chamada ao Algoritmo das Projeções Sucessivas foi inserida no algoritmo Ransac antes da fase de geração de hipótese.

Assim, as matrizes de entrada Xcal e Xval para o Ransac foram geradas a partir do resultado do processamento do APS, ou seja, o contendo apenas as colunas selecionadas (*Var_Sel*).

Os valores *default* descritos no Quadro 3 foram mantidos, exceto para o número máximo de iterações:

- Max_iters = 10.000

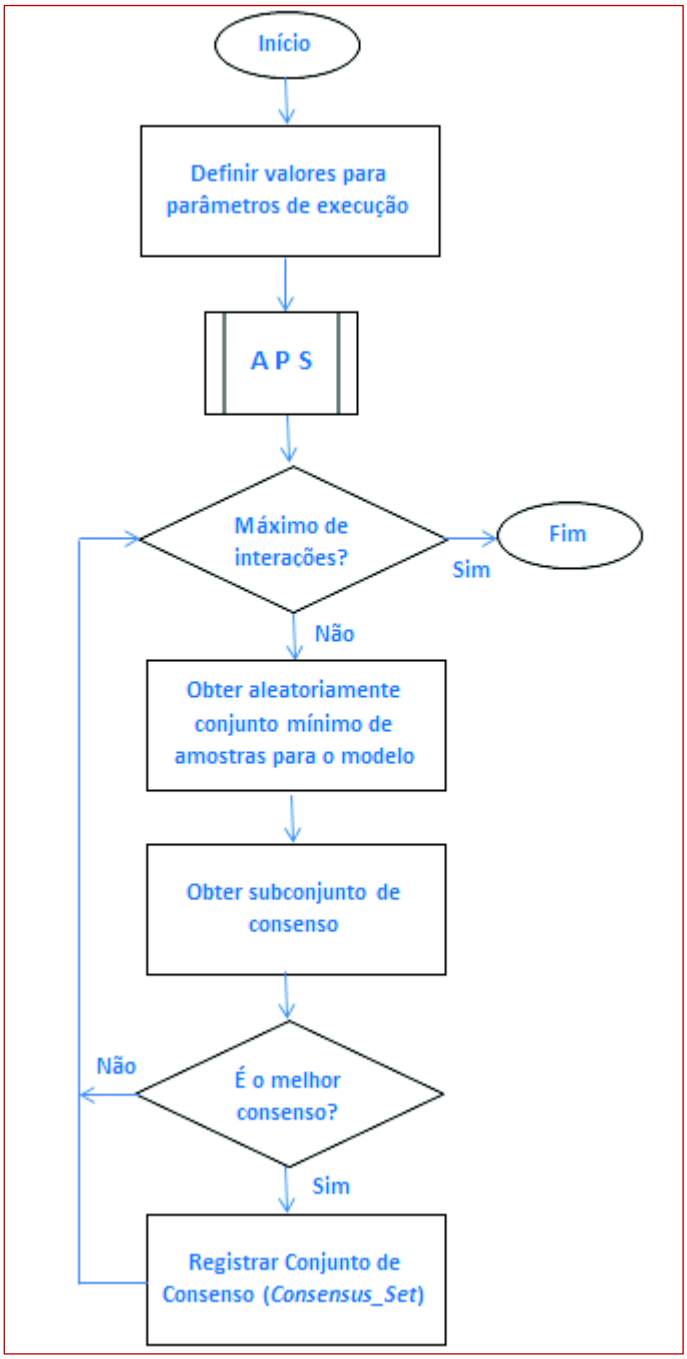


Figura 9 - Fluxograma de execução da Opção VI.

Na sequência, foi executado o algoritmo Ransac tendo como entrada as matrizes resultantes da aplicação do APS e os parâmetros de execução:

- Probabilidade de *inliers* = 0,99
- Máximo de iterações = 10.000
- Variância = 0,05

Foram realizadas dez execuções desta Opção. O resultado obtido para o RMSEP foi:

- **RMSEP = 0,873032** – Melhor valor obtido.

Constatação 3: Constata-se que para a seleção de variáveis, a execução do APS antes da execução do Ransac resultou no melhor RMSEP.

Quadro 8 - Constatação 3: Seleção de variáveis.

Seleção simultânea de amostras e de variáveis.

Opção VII: Aplicação dos algoritmos Ransac e APS de forma combinada.

Os algoritmos Ransac e APS foram executados como apresentado no fluxograma da Figura 10 a seguir.

Os valores *default* descritos no Quadro 3 foram mantidos, exceto para o número máximo de iterações:

- Max_iters = 10.000

O algoritmo APS foi inserido na fase de geração de hipótese, ou seja, durante a seleção aleatória das amostras mínimas.

Nesta configuração, a cada par de amostras aleatoriamente selecionadas pelo Ransac, o algoritmo APS foi executado para eliminar a multicolinearidade presente neste subconjunto de dados, preservando apenas as variáveis representativas.

Assim, a cada iteração foram gerados subconjuntos de dados composto pelas amostras (linhas) aleatoriamente selecionadas na fase de geração de hipótese do Ransac, e pelas variáveis (colunas) selecionadas pelo APS.

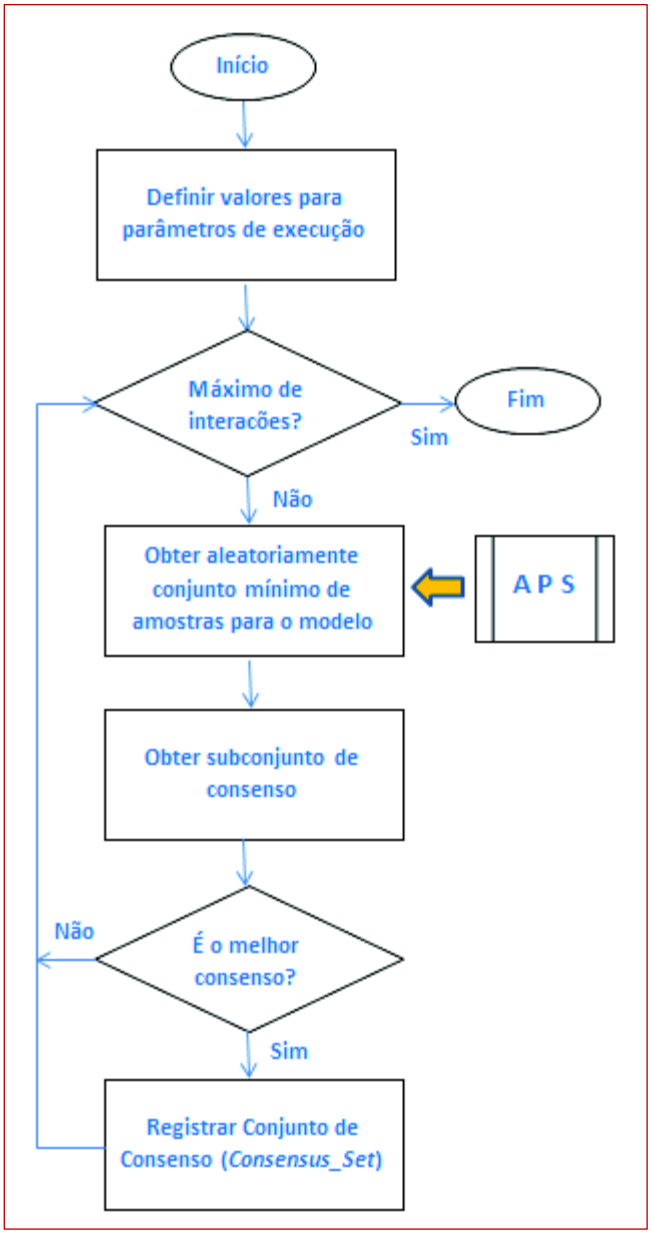


Figura 10 - Fluxograma para execução da Opção VII.

Os parâmetros de estimação do modelo pela regressão linear múltipla foram obtidos a partir desta matriz resultante.

Em seguida, a fase de teste de hipótese para obtenção do subconjunto de dados de consenso (*Consensus_Set*) foi executada, resultando nos seguintes valores do RMSEP:

- RMSEP = 1,236793 – Para o maior número de *inliers* obtido (201 de 390 amostras).
- **RMSEP = 1,230706** – Melhor valor obtido. Selecionadas 73 amostras.

Constatação 4: Constata-se que a seleção simultânea de amostras e de variáveis, mediante a execução do APS incorporada à fase de obtenção da amostra mínima, resulta em um RMSEP melhor que da aplicação do Ransac isolado.

Quadro 9 - Constatação 4: Seleção simultânea de amostras e variáveis.

O Quadro 10 apresenta os resultados observados para o RMSEP de forma sintetizada:

Resultados observados para o RMSEP	
Seleção de Amostras	
Opção I: Algoritmo Ransac	1,256766
Opção II: Algoritmo das Projeções Sucessivas	-
Opção III: Algoritmo Ransac seguido do Algoritmo das Projeções Sucessivas	0,23738
Seleção de Variáveis	
Opção IV: Algoritmo Ransac	1,307060
Opção V: Algoritmo das Projeções Sucessivas	0,22234
Opção VI: Algoritmo das Projeções Sucessivas seguido do Algoritmo Ransac	0,873032
Seleção simultânea de Amostras e de Variáveis	
Opção VII: Algoritmos Ransac e das Projeções Sucessivas combinados	1,230706

Quadro 10 - Síntese dos resultados do RMSEP.

4. CONCLUSÕES

Este trabalho abordou o problema do particioamento de dados para seleção de amostra e de variáveis em um conjunto de dados químicos obtidos pelo processo da escpectrofotometria.

Foram empregados para a realização do método proposto a Regressão Linear Múltipla, o algoritmo Ransac e o Algoritmo das Projeções Sucessivas.

Para avaliação das capacidades preditivas dos modelos obtidos na execução de cada uma das opções foi utilizado o RMSEP.

Versões do algoritmo Ransac foram implementadas especificamente a finalidade deste trabalho, contemplando cada uma as especificidades conforma opção a experimentada.

O Algoritmo das Projeções Sucessivas foi executado sem adaptações, apenas ajustando-se as matrizes de dados de entrada conforme cada opção.

Os resultados observados permitem concluir que:

1. Para a seleção de amostras, a aplicação do Algoritmo das Projeções Sucessivas ao final da execução do Ransac melhora significativamente a capacidade preditiva do algoritmo Ransac (Opção III *vs*. Opção I).

2. Para a seleção de variáveis, a aplicação do Algoritmo das Projeções Sucessivas também melhora a capacidade preditiva do Ransac (Opção VI *vs*. Opção IV). No entanto, o APS isoladamente se apresenta como melhor opção para a seleção exclusiva de variáveis (Opção V *vs*. Opções IV e VI).

3. Para a seleção simultânea de amostras e de variáveis, a aplicação do Algoritmo das Projeções Sucessivas incorporado ao algoritmo Ransac melhora a capacidade preditiva do Ransac aplicado isoladamente para seleção exclusiva de amostras (Opção VII *vs*. Opção I).

4. Para a seleção simultânea de amostras e de variáveis, a aplicação do Algoritmo das Projeções Sucessivas incorporado ao algoritmo Ransac melhora a capacidade preditiva do Ransac aplicado isoladamente para seleção exclusiva de amostras (Opção VII *vs*. Opção I).

5. Para a seleção simultânea de amostras e de variáveis, a aplicação do Algoritmo das Projeções Sucessivas incorporado ao algoritmo Ransac

apresenta uma capacidade preditiva pior em relação à obtida com a aplicação do APS isoladamente para seleção exclusiva de variáveis (Opção VII *vs*. Opção V).

6. Para a seleção simultânea de amostras e de variáveis, a aplicação do Algoritmo das Projeções Sucessivas incorporado ao algoritmo Ransac apresenta uma capacidade preditiva pior em relação à obtida com sua aplicação combinada com o Ransac para seleção exclusiva de variáveis (Opção VII *vs*. Opção VI).

Portanto, do ponto de vista da influência do Algoritmo das Projeções Sucessivas sobre o Ransac, conclui-se que o APS tem um efeito positivo significativo no algoritmo Ransac, tanto na seleção de amostras, como na seleção de variáveis, como também na seleção simultânea de amostras e variáveis.

Como trabalho futuro sugere-se ampliar a aplicação em outros estudos de caso para avaliar a possibilidade de generalização destas conclusões.

REFERÊNCIAS BIBLIOGRÁFICAS

ARAÚJO, M. C. U., *et al.* **The successive projections algorithm for variable selection in Spectroscopic multicomponent analysis**. Chemometrics and Intelligent Laboratory Systems, 57(2):65 – 73, 2001.

BLANCHET, F. G.; Legendre, P.; Borcard, D. **Forward selection of explanatory variables**. Ecology. 89: 2623, 2008.

BRO, R. **Multivariate calibration: What is in chemometrics for the analytical chemist?** Analytica Chimica Acta, v. 500, n. 1, p. 185-194, 2003.

CASALE, M., *et al.* **The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil**. Food Chemistry. 118: 163-170, 2010.

CHEN, T.; Martin, E. **Bayesian linear regression and variable selection for spectrostopic calibration**. Analytica Chimica Acta. 631: 13, 2009.

CHENG, C.; SHANG-HONG, L. **A consensus sampling technique for fast and robust model fitting**. Pattern Recognition. 42: 1318, 2009.

DANTAS FILHO, H. A., *et al.* **A strategy for selecting calibration samples for multivariate modeling.** Chemometrics and Intelligent Laboratory Systems. 72: 83, 2004.

DI NEZIO, M. S., *et al.* **Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water**. Microchemical Journal. 85: 194, 2007.

FARRAR, D. E.; GLAUBER, R. R. **Multicollinearity in regression analysis: the problem revisited.** The Review of Economics and Statistics. 49: 92, 1967.

FERREIRA, M. M. C., Antunes, A. M., Melgo, M. S., Colpe, P. L. O., **Quimiometria I: Calilbração Multivariada, um tutorial**. Química Nova 22, 724-731, 1999.

FISCHLER, M. A.; Bolles, R. C. **Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography**. Graphics and Image Processing, 1981.

GALVÃO, R. K. H., *et al.* **Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry**. Analytica Chimica Acta. 443: 107, 2001.

GALVÃO, R. K. H., *et al.* **A method for calibration and validation subset partitioning**. Talanta. 67: 736-740, 2005.

GALVÃO, R. K. H. **A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm**. Chemometrics and Intelligent Laboratory Systems, Volume 92, Issue 1, (pp 83-91), 2008.

GALVÃO, R. K. H.; Araújo, M. C. U., **Linear regression modelling: variable selection**. In: WALCZAK, B., FERRÉ, R. T., BROWN, S., Compreehensive chemometrics, 2009.

GEMPERLINE, P. **Practical guide to chemometrics**. 2nd Edition. CRC Press, 2006.

GOODARZI, M.; Freitas, M. P.; Jensen, R. **Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase3 Inhibitory Activities**. J. Chem. Inf. Model. 49: 824, 2009.

GOUDARZI N., *et al*. **QSPR Modeling of Soil Sorption Coefficients (Koc) of Pesticides Using SPAANN and SPAMLR**. Journal of Agricultural and Food Chemistry. 57: 7153, 2009.

GRUBBS, F. E. **Procedures for Detecting Outlying Observations in Sample**. TECHNOMETRICS, Vol. 11, N. 1, 1969.

HAIR JR. *, et al*. **Multivariate Data Analysis**. 7th Edition, Prentice Hall. 2009.

HOCKING, R. R. **The analysis and selection of variables in linear regression**. Biometrics, 32, pp. 1-49, 1976.

HONORATO, F. A., *et al*. **Robust modeling for multivariate calibration transfer by the successive projections algorithm**. Chemometrics and Intelligent Laboratory Systems. 76: 65-72, 2005.

HOPE, P. K. **The evolution of chemometrics.** Analytca Chimica Acta. 500, 365-377. 2003.

HUBER, P. J.; Ronchetti, Elvezio M. **Robust Statistics**. Wiley, 2009.

JINGFU, H., *et al*. **Robust fundamental matrix estimation with accurate outlier detection**. Journal of Information Science and Engineering. 23: 1213, 2007.

JOHNSON, R.; WICHERN, D. **Applied Multivariate Statistical Analysis**, 6th ed. Pearson New International Edition, 2014.

KENNARD, R. W.; Stone, L. A. **Computer aided design of experiments**. Technometrics, 11, (pp 137-148), 1969.

LEARDI, R. **Application of genetic algorithm-PLS for feature selection in spectral data sets**. Journal of Chemometrics. 14: 643, 2000.

MILLER, A. J. **Selection of Subsets of Regression Variable**. Journal of the Royal Statistical Society. Vol. 147, No. 3: 389-425, 1984.

NUNES, P. G. A. **Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às espectrometrias UV-VIS e NIR**, Tese de Doutorado. UFPB/CCEN, João Pessoa, 2008.

PAUL, R. K**. Multicollinearity: causes, effects and remedies**. IASRI, New Delli, 2006.

PONTES, M. J. C., *et al.* **The successive projections algorithm for spectral variable selection in classification problems**. Chemometrics and Intelligent Laboratory Systems. 78: 11, 2005.

RUZGIENE, B.; Forstner, W. **Ransac for outlier detection. Geodesy and Cartography**. Vol XXXI: 3, 2005.

SAPTORO, A., Tadé, M. O.; Wuthaluru, H. **Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models**. Chemical Product and Process Modeling. Vol. 7. Iss. 1, Art. 13, 2012.

SILVEY, S. D. **Multicollinearity and Imprecise Estimation**. Journal of the Royal Statistical Society. 31: 539, 1969.

SKOOG, D. A.; HOLLER, F. James; and CROUCH, Stanley R. **Principles of instrumental analysis**. 6th Edition. Cengage learning, 2017.

SOARES, A. S., *et. a*l. **Improving the computational efficiency of the successive projections algorithm by using a sequential regression implementation: a case study involving nir spectrometric analysis of wheat samples**. J. Braz. Chem. Soc., v. 21, n. 4, São Paulo, 2010.

SOARES, A. S., *et a*l. **Spectroscopic Multicomponent Analysis Using Multi-objective Optimization for Variable Selection**. Computer Technology and Application. 4: 466-475, 2013.

SOARES, S. F. C., *et al*, **A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferents**. Analytica Chimica Acta. 689: 22, 2011.

SWINIARSKI, R. W.; Skouron A. **Rough set methods in feature selection and recognition**. Pattern Recognition Letters. 24: 833, 2003.

WALPOLE, R. E*., et al.* **Probabilty.&.Statistics.for.Engineers.&.Scientists**, 9ed. Prentice Hall, 2012.

WOLD, S. **Chemometrics: what do we mean with it, and what do we want from it?** Chemometrics and Intelligent Laboratory Systems. 30: 109, 1995.

YUN, Yong-Huan, *et.al*. **Using variable combination population analysis for variable selection in multivariate calibration**. Analytica Chimica Acta. 826: 14-23, 2014.