Tese apresentada à Pró-Reitoria de Pós-Graduação e Pesquisa do Instituto Tecnológico de Aeronáutica, como parte dos requisitos para obtenção do título de Mestre em Ciências no Curso de Engenharia Eletrônica e Computação, Área de Sistemas e Controle

Arlindo Rodrigues Galvão Filho

CALIBRAÇÃO MULTIVARIADA EMPREGANDO O ALGORITMO DAS PROJEÇÕES SUCESSIVAS COM REAMOSTRAGEM E COMBINAÇÃO DE MODELOS

Tese aprovada em sua versão final pelos abaixo assinados:

Prof. Roberto Kawakami Harrop Galvão
Orientador

Prof. Mario César Ugulino Araújo
Co-orientador

Prof. Celso Massaki Hirata
Pró-Reitor de Pós-Graduação e Pesquisa

Campo Montenegro
São José dos Campos, SP - Brasil
2010

Dados Internacionais de Catalogação-na-Publicação (CIP)

Divisão Biblioteca Central do ITA/CTA

Galvão Filho, Arlindo Rodrigues

Calibração multivariada empregando o algoritmo das projeções sucessivas com reamostragem e combinação de modelos / Arlindo Rodrigues Galvão Filho.

São José dos Campos, 2010.

75f.

Tese de Mestrado – Curso de Engenharia Eletrônica e Computação. Área de Sistemas e Controle – Instituto Tecnológico de Aeronáutica, 2010. Orientador: Prof. Roberto Kawakami Harrop Galvão. Co-orientador: Prof. Mario César Ugulino Araújo.

1. Calibração Multivariada. 2. Seleção de variáveis. 3. Reamostragem. 4. Combinação de modelos. I. Centro Técnico Aeroespacial. Instituto Tecnológico de Aeronáutica. Divisão de Engenharia Eletrônica e Computação. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

GALVÃO FILHO, Arlindo Rodrigues. **Calibração multivariada empregando o algoritmo das projeções sucessivas com reamostragem e combinação de modelos**. 2010. 75f. Tese de Mestrado – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Arlindo Rodrigues Galvão Filho

TITULO DO TRABALHO: Calibração multivariada empregando o algoritmo das projeções sucessivas com reamostragem e combinação de modelos.

TIPO DO TRABALHO/ANO: Tese / 2010

Arlindo Rodrigues Galvão Filho

Av. Cidade Jardim, 679

CEP 12.233-066 – São José dos Campos–SP

CALIBRAÇÃO MULTIVARIADA EMPREGANDO O ALGORITMO DAS PROJEÇÕES SUCESSIVAS COM REAMOSTRAGEM E COMBINAÇÃO DE MODELOS

Arlindo Rodrigues Galvão Filho

Composição da Banca Examinadora:

Prof.	Jackson Paul Matsuura	Presidente	-	ITA
Prof.	Roberto Kawakami Harrop Galvão	Orientador	-	ITA
Prof.	Mario César Ugulino Araújo	Co-orientador	-	UFPB
Dr.	Alexsandro Machado Jacob	Membro Externo	-	ITAÚ
Prof.	Carlos Henrique Quartucci Forster	Membro	-	ITA

ITA

Ao meu pai, que ainda vive em nossos corações.

"Seja a diferença que você que ver no mundo."
— Mahatma Gandhi

Resumo

O termo Calibração Multivariada se refere à construção de um modelo matemático que permita prever o valor de uma grandeza de interesse com base em valores medidos de um conjunto de variáveis explicativas. Neste contexto o Algoritmo das Projeções Sucessivas (APS) é uma técnica de seleção de variáveis que objetiva a minimização de problemas de colinearidade em Regressão Linear Múltipla (RLM). Recentemente, constatou-se que a capacidade preditiva de modelos APS-RLM pode ser aprimorada com o uso de um método de reamostragem e combinação de modelos conhecido como *subagging*. Este trabalho aprofunda o estudo do *subagging* em conjunto com APS, investigando detalhes que não haviam sido anteriormente contemplados. Para isso, apresenta-se um estudo de caso envolvendo a determinação de umidade e proteína em trigo por espectrometria no infravermelho próximo. Em particular, observa-se que a capacidade preditiva e a sensibilidade a ruído dos modelos resultantes são aprimoradas independentemente da fração de reamostragem adotada no *subagging*. Adicionalmente, constata-se que o uso de validação cruzada ou validação por série de teste conduzem a resultados similares. Finalmente, tendo em vista o aumento no tempo de cálculo demandado para implementação do *subagging*, em comparação como APS tradicional, justifica-se o estudo de técnicas para redução da carga computacional envolvida. Neste trabalho propõe-se o uso de uma técnica de regressões sequenciais para facilitar a avaliação de subconjuntos de variáveis na etapa mais demorada do algoritmo.

Abstract

The term Multivariate Calibration refers to the construction of a mathematical model to predict the value of a quantity of interest based on measured values of a set of explanatory variables. In this context, the Successive Projections Algorithm (SPA) is a variable selection technique that aims to minimize collinearity issues in Multiple Linear Regression (MLR). Recently, it was found that the predictive ability of MLR-SPA models can be improved by using a method of resampling and models combination known as subagging. This work furthers the study of subagging in conjunction with SPA, investigating details that were not previously covered. For this purpose, a case study involving the determination of moisture and protein in wheat by near infrared spectrometry is presented. In particular, the results indicate that the predictive capacity and sensitivity to noise of the resulting models are enhanced regardless of the resampling fraction adopted in subagging. Additionally, the use of cross-validation or validation by a separate test set are found to provide similar results. Finally, given the increase in computation time demanded for implementation of subagging compared to traditional SPA, the present work is also concerned with the development of a technique for reducing the computational workload involved. For this purpose, a technique of sequential regressions is employed to facilitate the evaluation of subsets of variables in the most time-demanding stage of the algorithm.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

$\%m/m$	Porcentagem massa/massa, unidade de medida
APS	Algoritmo das Projeções Sucessivas, utilizando série de teste
APS-Sequencial	Algoritmo das Projeções Sucessivas utilizando regressão sequenciais
APS-VC	Algoritmo das Projeções Sucessivas utilizando validação cruzada
APS-*Subagging*	Algoritmo que faz uso do *subagging* em conjunto com APS
APS-VC-*Subagging*	Algoritmo que faz uso do *subagging* em conjunto com APS-VC
Bagging	***B****ootstrap* ***agg****regat****ing***
Bragging	***B****ootstrap* ***r****obust* ***agg****regat****ing***
NIR	*Near Infra-Red*
RAM	*Random Access Memory*
RLM	Regressão Linear Múltipla
RMSE	*Root Mean Square Error of Validation*
RMSEP	*Root Mean Square Error of Predition*
PC	*Principal Component*
PRESS	*Prediction Error Sum of Squares*
Subagging	***Sub****sample* ***agg****regat****ing***
VC	Validação Cruzada

Lista de Símbolos

$(\cdot)_m$	Subscrito que denota modelagem
$(\cdot)_c$	Subscrito que denota calibração
$(\cdot)_v$	Subscrito que denota validação
x	Variável independente
y	Variável dependente
$\hat{y}$	Estimativa de y
β	Coeficiente de regressão
N	Número máximo de observações
K	Número máximo de variáveis
$\mathbf{X}$	Matriz que contém os valores dos espectros medidos
$\mathbf{y}$	Vetor que contém os valores das propriedades em análise
$\mathbf{b}$	Vetor de coeficientes de regressão
$\overline{\mathbf{b}}$	Estimativa do vetor de coeficientes de regressão
$\mathbf{P}$	Matriz de projeções
SEL	Matriz contendo as variáveis selecionadas
M	Tamanho máximo da cadeia de variáveis
α	Coeficiente de significância no teste F
r	Índice de relevância
$\mathbf{v}$	Vetor contendo os índices das variáveis selecionadas pelo APS

P	Número de iterações do *subagging*
I	Índices aleatóriamente gerados no *subagging*
$\|\|\mathbf{b}\|\|$	Norma 2 do vetor de coeficientes de regressão

1 Introdução

O termo Calibração Multivariada se refere à obtenção de um modelo matemático que permita prover o valor de uma grandeza y com base em valores medidos de um conjunto de variáveis explicativas $x_1, x_2, \ldots, x_k$. Em outras palavras, deseja-se encontrar uma função $f = \Re^k \rightarrow \Re$ tal que

$$\hat{y} = f(x_1, x_2, \ldots, x_k) \tag{1.1}$$

sendo $\hat{y}$ uma estimativa de y.

Em diversos casos de interesse prático, um modelo apropriado pode ser obtido com o uso de uma função afim, isto é

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k \tag{1.2}$$

em que $b_0, b_1, \ldots, b_k$ são coeficientes a determinar. Tal relação pode ser amparada por leis físicas, tal como lei de Beer (HOLLER; SKOOG; CROUCH, 1998) que relaciona a absorbância da radiação à concentração de uma espécie absorvente em um dado meio. Em outros casos, modelos da forma 1.2 podem ser empregados para aproximar relações não-lineares em uma pequena faixa de valores de y

O processo de obtenção dos coeficientes do modelo 1.2 com base em um conjunto

de observações de y e $x_1, x_2, \ldots, x_k$ é conhecido como Regressão Linear Múltipla (RLM) (DRAPER; SMITH, 1998), sendo tipicamente realizado pelo método de mínimos quadrados. Porém tal método requer um número de observações maior que o número de variáveis. Em muitas aplicações, o que ocorre é o oposto (mais variáveis que amostras). Para contornar esse problema, pode-se fazer uso de algoritmos de seleção de variáveis.

Neste contexto, algoritmos genéticos têm sido empregados em diversos trabalhos (BROADHURSTA *et al.*, 1997; LEARDI, 2001; RIMBAUD *et al.*, 1995). O algoritmo genético é um método de seleção de variáveis de natureza estocástica inspirado na seleção natural das espécies proposta por Darwin. Os dados são vistos como uma população de indivíduos, e cada indivíduo tem suas características individuais convertidas em um cromossomo, que é uma sequencia de elementos chamados "genes". A aptidão do indivíduo, considerado como a sua capacidade de sobreviver e gerar descendentes, é medida por uma função objetivo, que deve refletir o índice de desempenho a ser maximizado no problema de otimização. Aos indivíduos com melhor aptidão é dada uma maior probabilidade de serem selecionados para o processo de reprodução, ou "*crossover*". Os cromossomos de cada par são então combinados usando operadores genéticos, a fim de gerar descendentes. Esses novos descendentes podem sofrer uma alteração genética, também chamada de "mutação". Finalmente, uma nova população é formada por substituir a geração anterior com a sua prole. Este ciclo evolutivo é repetido até que um critério de paragem dada é satisfeita (GOLDBERG, 1989; WALCZAK; FERRé; BROWN, 2009). A principal desvantagem desse método é que a convergência para um mínimo global não é garantida, podendo-se obter uma solução inadequada como resultado.

Outro algoritmo de seleção de variáveis bastante utilizado é a regressão *stepwise* (regressão "passo a passo") (KOKALY; CLARK, 1999). Nele, a importância apresentada por

cada variável dentro de um dado modelo é investigada. Para isso, as variáveis são adicionadas ou eliminadas do modelo de acordo com algum critério pré-estabelecido. Dessa forma, no decorrer das iterações cada variável á avaliada por um F-teste parcial nas fases de inclusão (também conhecida como *forward selection*) e de eliminação (ou *backward elimination*) (FORINA *et al.*, 2007). A fase de inclusão possibilita a inserção de variáveis no modelo que obtiveram uma avaliação no teste F parcial superior à um valor crítico de entrada. Já na fase de eliminação, cada variável que já está no modelo é novamente avaliada com teste F parcial e, se valor obtido for menor que um valor crítico de saída, a variável é retirada do modelo e novamente armazenada no conjunto das variáveis disponíveis. A seleção termina quando nenhuma variável pode ser adicionada nem removida do modelo de acordo com o critério teste F parcial (DRAPER; SMITH, 1998; WALCZAK; FERRé; BROWN, 2009). Uma desvantagem deste algoritmo é que a escolha adequada dos valores críticos mínimos pode não ser simples, uma vez que suas escolhas dependem da natureza dos dados utilizados. Outro fator crítico do *stepwise* é assumir que os resíduos da regressão siga uma distribuição Gaussiana, o que pode não ocorrer para determinados conjuntos de dados.

O algoritmo das projeções sucessivas (APS), proposto por (ARAúJO *et al.*, 2001), objetiva a redução de colinearidade que é um problema que sabidamente conduz a mal condicionamento da regressão e obtenção de modelos com capacidade de generalização inadequada (NæS; MEVIK, 2001). O APS basicamente é formado por três fases: A primeira pré-seleciona subconjuntos de variáveis que exibam baixa colinearidade. A segunda seleciona o melhor desses subconjuntos com base na habilidade de predição do modelo resultante e a terceira fase elimina as variáveis que não contribuem significativamente para a predição.

O trabalho inicial do APS realizou uma comparação entre a técnica proposta e o algoritmo genético, mínimos quadrados parciais e regressão por componentes principais. O APS alcançou os melhores resultados aplicado à análise espectrofotométrica em questão. Foi utilizado também em problemas de otimização com transformadas *wavelets* (COELHO *et al.*, 2003b; COELHO *et al.*, 2003a; GALVãO *et al.*, 2004), em determinação de solubilidade de sólidos em cervejas (LIU; JIANG; HE, 2009), determinação de glicose em sangue humano (LI; LI; ZHANG, 2009), de parâmetros qualitativos de óleos vegetais (PEREIRA *et al.*, 2008), determinação de cinco fenólicos compostos na água do mar (NEZIO *et al.*, 2007), determinação de enxofre em amostras de diesel (BREITKREITZ *et al.*, 2003), e diversas outras aplicações. Portanto fica claro que bons modelos de calibração multivariada podem ser obtidos com o uso da RLM acoplada ao APS.

Em (GALVãO *et al.*, 2006) foi proposto o uso do APS em conjunto o *subagging* (***sub****sample* ***agg****regat****ing***). O *subagging* é uma técnica de combinação de modelos baseada em sub-amostragem, que foi criada como uma alternativa em relação ao *bagging* (BREIMAN, 1996a; BREIMAN, 1996b; PINO-MEJIAS *et al.*, 2004). Tal técnica é computacionalmente menos custosa mas não compromete substancialmente os ganhos obtidos com o *bagging* (***b****ootstrap* ***agg****regat****ing***) (BÜHLMANN; YU, 2002). Basicamente, para se obter um modelo *subagging*, diversas amostras aleatóriamente geradas são utilizadas para gerar modelos RLM. Posteriormente a combinação dos modelos é feita através da média aritmética dos modelos resultantes, o que possibilita uma redução considerável da variância do preditor.

No estudo de caso desenvolvido em (GALVãO *et al.*, 2006) observaram melhorias em relação ao APS convencional de 33% para determinação espectrofotométrica de quatro parâmetros qualitativos de diesel, contudo houve algumas limitações. Não se investigou

as possíveis alterações nos resultados se os dados estivessem contaminados por ruídos, o que é possível se tratando de dados reais. Também foi investigado somente uma realização do *subagging*, não considerando as variações dos resultados no decorrer de múltiplas realizações. Esta análise é de suma importância devido à estocasticidade do *subagging*. Outro fator não analisado é o comportamento dos resultados com o *subagging* alterando a fração utilizada na subamostragem, o que é um parâmetro crucial para a generalização dos modelos obtidos.

Neste trabalho investiga-se o comportamento do APS em conjunto com o *subagging* utilizando 30 sementes do gerador de números aleatórios diferentes, no intuito de verificar o comportamento médio de múltiplas realizações. Propõe-se também a variação da fração da subamostragem utilizada pelo *subagging*, comparando-a com o APS sem *subagging* utilizando uma divisão do conjunto de dados respectiva à fração analisada. Os modelos resultantes também foram avaliados quanto a sensibilidade a ruído.

Adicionalmente, uma outra configuração do APS foi avaliada utilizando o *subagging*. Esta configuração do APS foi proposta em (GALVãO *et al.*, 2007), e utiliza a validação cruzada como métrica de avaliação na seleção de variáveis. Dessa forma, as mesmas avaliações feitas adicionando o *subagging* para o APS também foram repetidas com o APS com validação cruzada.

No processo de avaliação das técnicas foi observado que, devido as diversas repetições dos algoritmos propostos, o APS gera um esforço computacional considerável. O que é ainda mais penalizado pelas repetições provindas *subagging*. Entretanto este trabalho propõe também uma estratégia para a redução do tempo computacional do APS baseado na regressão sequencial apresentada em (GUSNANTO *et al.*, 2003). Essa nova configuração do APS foi aqui nomeada de APS-Sequencial.

1.1 Organização do texto

O Capítulo 2 descreve o Algoritmo das Projeções Sucessivas (APS) para problemas de regressão. Na Seção 2.1 apresenta o APS-*Subagging*. O Capítulo 3 apresenta a regressão sequencial, juntamente com um exemplo numérico na Seção 3.1 e um exemplo da redução do custo computacional na Seção 3.2. O Capítulo 4 descreve os materiais e métodos utilizados para obtenção dos resultados apresentados no Capítulo 5. O Capítulo 6 apresenta a conclusão do trabalho.

2 Algoritmo das Projeções Sucessivas

Um modelo de regressão linear múltipla (RLM) expressa a relação entre um conjunto de variáveis independentes $x_1, x_2, \ldots, x_K$ (também chamadas explicativas ou preditoras) e uma variável dependente y (resposta) através de uma equação da forma

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{2.1}$$

em que β_k, $k = 1, 2, \ldots, K$ são constantes a serem determinadas e ε é uma parcela aleatória de erro (CHARNET *et al.*, 2008). Neste trabalho o termo "variável" (sem qualificativo) será utilizado para representar uma variável independente.

Os coeficientes de regressão podem ser estimados com base em um conjunto de modelagem contendo N observações $\{(x_{i,1}, x_{i,2}, \ldots, x_{i,K}, y_i), i = 1, 2, \ldots, N\}$, sendo $x_{i,k}$ e y_i os valores de x_k e y na i-ésima observação. Tal estimativa pode ser efetuada através do método de mínimos quadrados. Se os elementos do conjunto de modelagem forem dispostos

na seguinte forma matricial:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ x_{2,1} & x_{2,2} & \dots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,K} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}. \tag{2.2}$$

então a estimativa $\mathbf{b}$ do vetor de coeficientes de regressão $\beta = [\beta_1, \beta_2, \dots, \beta_K]'$ é dada por (DRAPER; SMITH, 1998):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.3}$$

desde que a matriz $\mathbf{X}'\mathbf{X}$ seja não-singular.

Algoritmos de seleção de variáveis tem por finalidade a escolha de um subconjunto de variáveis a serem incluídas no modelo de regressão. O Algoritmo das Projeções Sucessivas (APS), proposto em (ARAúJO *et al.*, 2001), tem por objetivo selecionar um subconjunto de variáveis com baixa colinearidade que portem informação apropriada para construção de um modelo de RLM por mínimos quadrados (GALVãO *et al.*, 2008; LIU; HE, 2009; PONTES *et al.*, 2005; SOUTO *et al.*, 2010). Para aplicação do APS, os dados de modelagem são divididos em dois conjuntos de calibração e validação, contendo N_c e N_v observações, respectativamente, sendo $N_c + N_v = N$. Seguindo uma disposição similar à apresentada na Equação (2.2), os dados de calibração e validação são dispostos em matrizes $\mathbf{X}_c(N_c \times K)$, $\mathbf{X}_v(N_v \times K)$ e vetores $\mathbf{y}_c(N_c \times 1)$, $\mathbf{y}_v(N_v \times 1)$. Em seguida cada coluna das matrizes $\mathbf{X}_c$ e $\mathbf{X}_v$ são centralizadas na média do conjunto de calibração, ou seja, calcula-se a média de cada coluna de $\mathbf{X}_c$ e subtrai-se cada observação das matrizes $\mathbf{X}_c$ e $\mathbf{X}_v$ do respectivo valor médio (FERREIRA *et al.*, 1999).

A execução do APS pode ser dividida em três fases principais. A fase 1 tem por objetivo gerar K cadeias com M variáveis cada uma, sendo

$$M = \min(N_c - 1, K). \tag{2.4}$$

Tais cadeias são geradas através da sequência de passos do Algoritmo 1 apresentado abaixo. Nesse algoritmo, $\mathbf{x}_k$ denota um vetor $(N_c \times 1)$ contendo as N_c observações da variável x_k no conjunto de calibração.

Algoritmo 1: Fase 1 do APS - Geração das cadeias de variáveis

Entradas - *Dados de calibração:* $\mathbf{X}_c$ $(N_c \times K)$

Passo 1: Faça $k = 1$

Passo 2: Faça $i = 1$

Passo 3: Faça $\mathbf{x}_j^1 = \mathbf{x}_j, j = 1, 2, \ldots, K$

Passo 4: Faça $\mathbf{z}^1 = \mathbf{x}_k^1$

Passo 5: Faça $\mathbf{SEL}(1, k) = k$

Passo 6: Calcule a matriz $\mathbf{P}^i$ de projeções no sub-espaço ortogonal a $\mathbf{z}^i$ conforme a Equação (2.5)

$$\mathbf{P}^i = \mathbf{I}\frac{\mathbf{z}^i(\mathbf{z}^i)'}{(\mathbf{z}^i)'\mathbf{z}^i} \tag{2.5}$$

em que $\mathbf{I}$ é uma matriz identidade de dimensões $N_c \times N_c$.

Passo 7: Calcule os vetores projetados $\mathbf{x}_j^{i+1}$ conforme a Equação (2.6)

$$\mathbf{x}_j^{i+1} = \mathbf{P}^i\mathbf{x}_k^j, j = 1, 2, \ldots, K. \tag{2.6}$$

Passo 8: Armazene o índice j^* da variável de maior projeção no elemento $(i+1,k)$ da matriz **SEL**:

$$j^* = \arg \max_{j \in \{1,2,\ldots,K\}} \| \mathbf{x}_j^{i+1} \| \tag{2.7}$$

$$\mathbf{SEL}(i+1,k) = j^* \tag{2.8}$$

Passo 9: Escolha o vetor que define as projeções na próxima iteração:

$$\mathbf{z}^{i+1} = \mathbf{x}_{j^*}^{i+1} \tag{2.9}$$

Passo 10: Faça $i = i+1$. Se $i \leq M$ volte ao Passo 4

Passo 11: Faça $k = k+1$. Se $k \leq K$ volte ao Passo 2

FIM As cadeias de variáveis correspondem aos índices armazenados nas colunas de **SEL**

Saídas - *Índices das variáveis selecionadas:* **SEL** $(M \times K)$

A partir de uma variável inicial $\mathbf{x}_k$, cada elemento da cadeia é escolhido de modo a possuir a menor colinearidade em relação as variáveis anteriores, até que se complete o tamanho máximo M. Para isso, toma-se sempre a variável com maior projeção no subespaço ortogonal ao das variáveis anteriores. O processo é repetido até que cada uma das K variáveis tenha sido usada para inicializar uma cadeia. Ao final, a k-ésima coluna da matriz **SEL** conterá os índices das M variáveis da k-ésima cadeia.

A Figura 2.1 ilustra as operações realizadas na Fase 1 do APS para um caso em que $K = 5$, $N_c = 3$. Neste exemplo, a sequência das projeções é inicializada com $\mathbf{z}^1 = \mathbf{x}_3^1 = \mathbf{x}_3$. Observa-se que a maior projeção obtida foi a referente a $\mathbf{x}_1$, portanto na próxima iteração, $\mathbf{z}^1 = \mathbf{x}_1^2$ será tomado o vetor de referência para a realização das projeções.

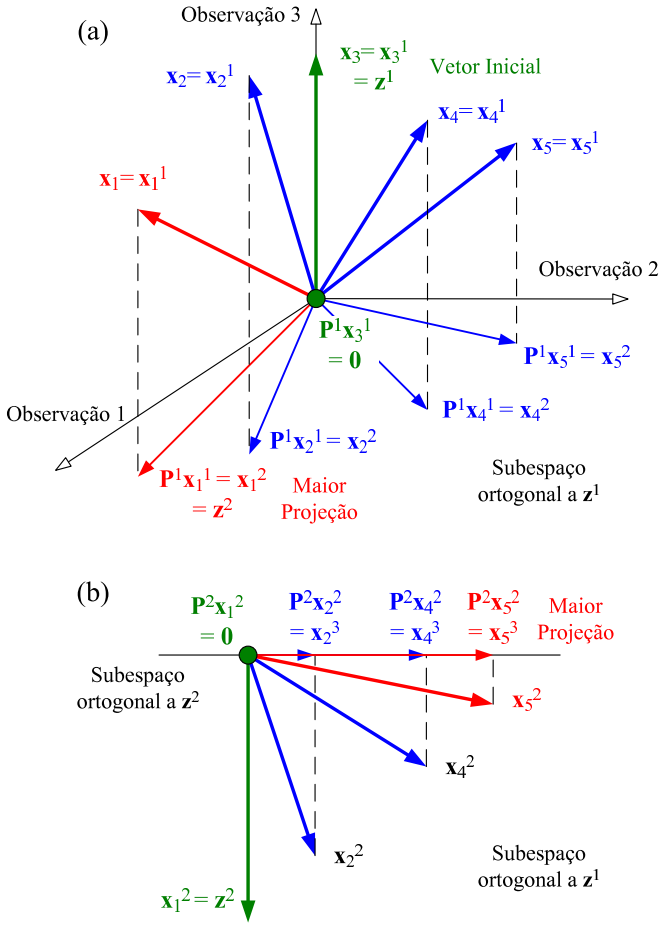


FIGURA 2.1 – Exemplo gráfico das operações realizadas na fase 1.

Na fase 2 o APS utiliza o conjunto de validação para avaliar subconjuntos de variáveis extraídos das cadeias geradas na fase 1, conforme a sequência de passos do Algoritmo 2.

Algoritmo 2: Fase 2 do APS - Avaliação dos subconjuntos de variáveis

Entradas: - *Dados de calibração:* $\mathbf{X}_c$ $(N_c \times K)$, $\mathbf{y}_c$ $(N_c \times 1)$;

- *Dados de validação:* $\mathbf{X}_v$ $(N_v \times K)$, $\mathbf{y}_v$ $(N_v \times 1)$;

- *Índices das variáveis selecionadas:* **SEL** $(M \times K)$.

Passo 1: Faça $k = 1$

Passo 2: Faça $m = 1$

Passo 3: Sejam $\mathbf{X}_{c,mk}$ e $\mathbf{X}_{v,mk}$ matrizes $(N_c \times m)$ e $(N_v \times m)$ formadas pelas colunas de $\mathbf{X}_c$ e $\mathbf{X}_v$ com índices $\mathbf{SEL}(1,k), \mathbf{SEL}(2,k) \ldots, \mathbf{SEL}(m,k)$ (isto é, os m primeiros elementos da k-ésima cadeia gerada na fase 1). Calcule o vetor de coeficiente de regressão $\mathbf{b}_{mk}$ com base em $\mathbf{X}_{c,mk}$ e $\mathbf{y}_c$ de acordo com a Equação (2.10)

$$\mathbf{b}_{mk} = (\mathbf{X}'_{c,mk}\mathbf{X}_{c,mk})^{-1}\mathbf{X}'_{c,mk}\mathbf{y}_c \tag{2.10}$$

e aplique o modelo resultante ao conjunto de validação conforme a Equação (2.11)

$$\hat{\mathbf{y}}_v = \mathbf{X}_{v,mk}\mathbf{b}_{mk} \tag{2.11}$$

Passo 4: Calcule o somatório do quadrado dos erros de validação ou PRESS (*Prediction Error Sum of Squares*) de acordo com a Equação (2.12)

$$\mathbf{PRESS}(m,k) = \sum_{i=1}^{Nv}(y_{v,i} - \hat{y}_{v,i})^2 \tag{2.12}$$

em que $y_{v,i}$, $\hat{y}_{v,i}$ são os i-ésimos componentes dos vetores $\mathbf{y}_v$ e $\hat{\mathbf{y}}_v$, respectivamente.

Passo 5: Faça $m = m + 1$. Se $m \leq M$ volte ao Passo 3

Passo 6: Faça $k = k + 1$. Se $k \leq K$ volte ao Passo 2

Passo 7: Obtenha os índices m e k que correspondem ao menor valor de PRESS conforme a Equação (2.13)

$$(m^*, k^*) = \arg \min_{m\in\{1,2,\ldots,M\}\ k\in\{1,2,\ldots,K\}} \mathbf{PRESS}(m,k) \tag{2.13}$$

FIM Os índices das variáveis escolhidas estão armazenados em

$\mathbf{SEL}(1, k^*), \mathbf{SEL}(2, k^*), \ldots, \mathbf{SEL}(m^*, k^*)$.

Saídas: - *Índices das variáveis selecionadas:* m^*, k^*

Como resultado da fase 2, o melhor sub-conjunto de variáveis é aquele que conduz ao menor valor de PRESS dentre os subconjuntos testados. Vale ressaltar que a quantidade máxima de variáveis nos subconjuntos avaliados na fase 2 pode ser limitada a um número $\overline{M} \leq M$ de modo a reduzir o esforço computacional envolvido.

Na fase 3 o APS utiliza mais uma vez o conjunto de validação, porém agora para reduzir o número de variáveis selecionadas na fase 2, descartando aquelas que não contribuam significantemente para a capacidade preditiva do modelo RLM resultante. Para esse propósito, emprega-se o Algoritmo 3 apresentado abaixo, que foi proposto em (GALVãO *et al.*, 2008). Inicialmente, calcula-se um "índice de relevância" para cada variável selecionada na Fase 2 como sendo o produto entre o desvio padrão amostral e o módulo do coeficiente de regressão dessa variável. Em seguida, as variáveis são progressivamente inseridas em um modelo MLR em ordem decrescente de relevância, obtendo-se o valor de PRESS para o conjunto de validação a cada nova variável inserida. Por fim, determina-se o menor número de variáveis para o qual o valor de PRESS não é significativamente maior do que o valor mínimo obtido nesse processo. Para isso, emprega-se um teste F, como proposto em (HAALAND; THOMAS, 1988).

A escolha de valor muito pequeno para o coeficiente de significância α no teste F pode resultar em modelos "sub-ajustados" (*underfitting*). Em contra partida, se α for muito grande o teste poderá resultar em modelos "sobre-ajustados" (*overfitting*). Para este trabalho foi adotado $\alpha = 0.25$ como sugerido na literatura (HAALAND; THOMAS, 1988; LI *et al.*, 2005).

Vale salientar a possibilidade do uso da Validação Cruzada (VC) para o cálculo do PRESS nas fases 2 e 3, ao invés de se empregar um conjunto de validação separado. Em (GALVãO *et al.*, 2007), realizaram o estudo de caso comparando o uso de um conjunto separado de validação e de VC concluindo-se que não existe diferenças expressivas nos resultados.

Algoritmo 3: Fase 3 do APS - Eliminação final de variáveis

Entradas: - *Dados de calibração:* $\mathbf{X}_c$ $(N_c \times K)$, $\mathbf{y}_c$ $(N_c \times 1)$;

- *Dados de validação:* $\mathbf{X}_v$ $(N_v \times K)$, $\mathbf{y}_v$ $(N_v \times 1)$;

- *Índices das variáveis selecionadas:* **SEL** $(M \times K)$, m^*, k^*;

- *Coeficiente de significância:* α.

Passo 1: Sejam $\mathbf{X}_{c,m^*k^*}$ e $\mathbf{X}_{v,m^*k^*}$ matrizes $(N_c \times m^*)$ e $(N_v \times m^*)$ formadas pelas colunas de $\mathbf{X}_c$ e $\mathbf{X}_v$ com índices $\mathbf{SEL}(1,k^*), \mathbf{SEL}(2,k^*) \ldots, \mathbf{SEL}(m^*,k^*)$ (isto é, as colunas associadas às variáveis selecionadas na fase 2). Calcule o vetor de coeficientes de regressão $\mathbf{b}_{m^*k^*}$ com base em $\mathbf{X}_{c,m^*k^*}$ e $\mathbf{y}_c$ de acordo com a Equação (2.14)

$$\mathbf{b}_{m^*k^*} = (\mathbf{X}'_{c,m^*k^*}\mathbf{X}_{c,m^*k^*})^{-1}\mathbf{X}'_{c,m^*k^*}\mathbf{y}_c \tag{2.14}$$

Passo 2: Calcule o índice de relevância r_j associado à j-ésima variável selecionada, conforme a Equação (2.15)

$$r_j = s_j|b_{m^*k^*}(j)|, j = 1, 2, \ldots, m^* \tag{2.15}$$

em que s_j é o desvio-padrão da j-ésima coluna de $\mathbf{X}_{m^*k^*}$. Seja $j_1, j_2, \ldots, j_{m^*}$ uma

sequência que ordene os valores de relevância de forma decrescente, isto é:

$$r_{j_1} \geq r_{j_2} \geq \ldots \geq r_{j_{m^*}} \tag{2.16}$$

Passo 3: Faça $l = 1$

Passo 4: Seja $\mathbf{X}_{c,l}$ e $\mathbf{X}_{v,l}$ matrizes $(N_c \times l)$ e $(N_v \times l)$ formadas pelas colunas de $\mathbf{X}_{c,m^*k^*}$ e $\mathbf{X}_{v,m^*k^*}$ com índices j_1 até j_l. Calcule o vetor de coeficientes de regressão $\mathbf{b}_l$ com base em $\mathbf{X}_{c,l}$ e $\mathbf{y}_c$ de acordo com a Equação (2.17)

$$\mathbf{b}_l = (\mathbf{X}'_{c,l}\mathbf{X}_{c,l})^{-1}\mathbf{X}'_{c,l}\mathbf{y}_c \tag{2.17}$$

e aplique o modelo resultante ao conjunto de validação conforme a Equação (2.18)

$$\hat{\mathbf{y}}_v = \mathbf{X}_{v,l}\mathbf{b}_l \tag{2.18}$$

Passo 5: Calcule o valor de PRESS de acordo com a Equação (2.19)

$$PRESS(l) = \sum_{i=1}^{N_v}(y_{v,i} - \hat{y}_{v,i})^2 \tag{2.19}$$

Passo 6: Faça $l = l + 1$. Se $l \leq m^*$ volte ao Passo 4

Passo 7: Calcule o mínimo dos valores de PRESS obtidos no Passo 5:

$$PRESS_{min} = \min_{l=1,2,\ldots,m^*} \mathbf{PRESS}(l) \tag{2.20}$$

Passo 8: Seja l^* o menor valor de l para o qual

$$\frac{\mathbf{PRESS}(l)}{PRESS_{min}} \leq F_{crit} \tag{2.21}$$

em que F_{crit} é o valor crítico da distribuição F com (N_v, N_v) graus de liberdade para um coeficiente de significância α escolhido a priori.

FIM: As variáveis selecionadas são aquelas associadas às colunas de $\mathbf{X}_{c,m^*k^*}$ com índices $j_1, j_2, \ldots, j_{l^*}$.

Saídas: - *Índices das variáveis selecionadas:*

$$\begin{aligned} \mathbf{v}(1) &= \mathbf{SEL}(j_1, k^*) \\ \mathbf{v}(2) &= \mathbf{SEL}(j_2, k^*) \\ \vdots &\quad\ \vdots \\ \mathbf{v}(l^*) &= \mathbf{SEL}(j_{l^*}, k^*); \end{aligned}$$

- *Coeficientes de regressão:* $\mathbf{b} = \mathbf{b}_{l^*}$.

2.1 APS-*Subagging*

O APS-*Subagging* foi proposto no artigo (GALVãO *et al.*, 2006) no intuito de melhorar a predição de modelos RLM com base no conceito de *subagging* apresentado em (BÜHLMANN; YU, 2002). No APS-*Subagging* os dados de modelagem são aleatoriamente divididos em um conjunto de calibração e um de validação que são então usados para obter um modelo RLM empregando-se o APS. Tal procedimentos é repetido diversas ve-

zes de forma a obter diferentes modelos que são então promediados ao final. Tal processo encontra-se descrito no Algoritmo 4 abaixo.

Algoritmo 4: APS-*Subagging*

Entradas: - *Dados de modelagem:* $\mathbf{X}$ $(N \times K)$, $\mathbf{y}$ $(N \times 1)$;

- *Nº de amostras a serem usadas na calibração:* N_c;

- *Nº de iterações:* P.

Passo 1: Faça $p = 1$

Passo 2: Gere aleatoriamente uma partição $\{I_c, I_v\}$ do conjunto de índices $I = \{1, 2, \ldots, N\}$ tal que $I_c \cup I_v = I$ e $I_c \cap I_v = \emptyset$, com I_c e I_v contendo N_c e $N - N_c$ índices, respectivamente.

Passo 3: Sejam $\mathbf{X}_c$ e $\mathbf{X}_v$ matrizes $(N_c \times K)$ e $(N_v \times K)$ formadas pelas linhas de $\mathbf{X}_c$ e $\mathbf{X}_v$ com índices pertencentes a I_c e I_v, respectivamente. Sejam também $\mathbf{y}_c$ e $\mathbf{y}_v$ vetores $(N_c \times 1)$ e $(N_v \times 1)$ formados pelas linhas de $\mathbf{y}_c$ e $\mathbf{y}_v$ com índices pertencentes a I_c e I_v, respectivamente.

Passo 4: Execute as fases 1, 2 e 3 do APS empregando as matrizes $\mathbf{X}_c$ e $\mathbf{X}_v$, e os vetores $\mathbf{y}_c$ e $\mathbf{y}_v$ construídos no Passo 3, obtendo os índices das variáveis selecionadas $\mathbf{v}(1), \mathbf{v}(2), \ldots, \mathbf{v}(l^*)$ e o respectivo vetor de coeficientes de regressão $\mathbf{b}$ $(l^* \times 1)$.

Passo 5: Construa um vetor de coeficientes de regressão $\mathbf{b}_p^{aps}$ $(K \times 1)$ com valores iguais a zero nas posições correspondentes às variáveis não-selecionadas, isto é:

$$\mathbf{b}_p^{aps}(k) = \begin{cases} 0, & k \notin \{\mathbf{v}(1), \mathbf{v}(2), \ldots, \mathbf{v}(l^*)\} \\ \mathbf{b}(i), & k = \mathbf{v}(i), i = 1, 2, \ldots, l^* \end{cases} \tag{2.22}$$

Passo 6: Faça $p = p + 1$. Se $p \leq P$ volte para o Passo 2.

Passo 7: Calcule a média $\mathbf{b}^{sub}$ dos p vetores de coeficientes de regressão obtidos no Passo 5, conforme a Equação 2.23

$$\mathbf{b}^{sub} = \frac{1}{P}\sum_{p=1}^{P}\mathbf{b}_p^{aps} \tag{2.23}$$

FIM: O modelo APS-*Subagging* é dado pelo vetor de coeficientes de regressão $\mathbf{b}^{sub}$.

Saídas: $\mathbf{b}^{sub}$

Neste trabalho será investigado também a utilização do *subagging* com o uso do APS empregando VC. Uma descrição da técnica proposta, denominada APS-VC-*Subagging*, é apresentada no Algoritmo 5 abaixo:

Algoritmo 5: APS-VC-*Subagging*

Entradas: - *Dados de modelagem:* $\mathbf{X}$ $(N \times K)$, $\mathbf{y}$ $(N \times 1)$;

- *Nº de amostras a serem usadas na calibração:* N_c;

- *Nº de iterações:* P.

Passo 1: Faça $p = 1$

Passo 2: Extraia aleatoriamente do conjunto de índices $I = \{1, 2, \ldots, N\}$ um subconjunto $\{I_c\}$ com N_c elementos não repetidos.

Passo 3: Seja $\mathbf{X}_c$ uma matriz $(N_c \times K)$ formada pelas linhas de $\mathbf{X}_c$ com índices pertencentes a I_c. Sejam também $\mathbf{y}_c$ um vetor $(N_c \times 1)$ formado pelas linhas de $\mathbf{y}_c$ com índices pertencentes a I_c.

Passo 4: Execute o APS utilizando a validação cruzada nas fases 2 e 3 para o cálculo do PRESS, empregando a matriz $\mathbf{X}_c$ e o vetor $\mathbf{y}_c$ construídos no Passo 3, obtendo os índices das variáveis selecionadas $\mathbf{v}(1), \mathbf{v}(2), \ldots, \mathbf{v}(l^*)$ e o respectivo vetor de coeficientes de regressão $\mathbf{b}$ $(l^* \times 1)$.

Passo 5: Construa um vetor de coeficientes de regressão $\mathbf{b}_p^{aps.vc}$ $(K \times 1)$ com valores iguais a zero nas posições correspondentes às variáveis não-selecionadas, isto é:

$$\mathbf{b}_p^{aps.vc}(k) = \begin{cases} 0, & k \notin \{\mathbf{v}(1), \mathbf{v}(2), \ldots, \mathbf{v}(l^*)\} \\ \mathbf{b}(i), & , k = \mathbf{v}(i) i = 1, 2, \ldots, l^* \end{cases} \tag{2.24}$$

Passo 6: Faça $p = p + 1$. Se $p \leq P$ volte para o Passo 2.

Passo 7: Calcule a média $\mathbf{b}^{vc.sub}$ dos p vetores de coeficientes de regressão obtidos no Passo 5 conforme a Equação (2.25)

$$\mathbf{b}^{vc.sub} = \frac{1}{P} \sum_{p=1}^{P} \mathbf{b}_p^{aps.vc} \tag{2.25}$$

FIM: O modelo APS-VC-*Subagging* é dado pelo vetor de coeficientes de regressão $\mathbf{b}^{vc.sub}$.

Saídas: $\mathbf{b}^{vc.sub}$

Vale salientar que o custo computacional envolvido na obtenção do modelo APS-*Subagging* e APS-VC-*Subagging* pode ser elevado devido à necessidade de executar o APS diversas vezes. Uma possível forma de reduzir este custo é a implementação do APS utilizando regressões sequenciais, como descrito no Capítulo 3.

3 Regressões Sequenciais

A segunda fase do APS tem como objetivo selecionar subconjuntos de variáveis com maior potencial preditivo a partir das cadeias de variáveis resultantes das projeções feitas na fase 1. Em relação ao custo computacional do algoritmo, observa-se que a fase 2 tipicamente tem o maior custo quando comparada às outras fases. Por exemplo, considerando um conjunto de calibração e um conjunto de validação com 200 e 100 objetos aleatoriamente gerados, ambos possuindo 300 variáveis. Ao executar o APS com estes conjuntos as fases 1, 2 e 3 gastaram 3.9, 130.1 e 0.05 segundos, respectivamente. Dessa forma, existe a oportunidade para melhoria computacional do APS eliminando a redundância dos cálculos da fase 2.

Por exemplo, para o modelo com uma variável $\{x_1\}$ faz-se a regressão por mínimos-quadrados para obter b_1. Para a regressão com duas variáveis $\{x_1, x_2\}$ faz-se novamente a regressão para obter b_2. Seria desejável que os coeficientes b_1 pudessem ser reaproveitados para a regressão com duas variáveis $\{x_1, x_2\}$, ao invés de efetuar todo o cálculo novamente. Para isso, foi elaborada uma modificação da fase 2 do APS utilizando o método de regressão sequencial proposto por Gusnanto (GUSNANTO *et al.*, 2003).

A formulação das regressões sequenciais inicia com modelo de uma única variável na

forma

$$y = \beta_1^{(1)} x_1 + \epsilon^{y|x_1} \tag{3.1}$$

sendo $\beta_1^{(1)}$ o coeficiente de regressão e $\epsilon^{y|x_1}$ o resíduo do modelo. Os superescritos (1) e $y|x_1$ denotam que uma variável independente é empregada no modelo e que y é regredida em x_1, respectivamente. A estimativa de mínimos-quadrados de $\beta_1^{(1)}$ é dada por:

$$b_1^{(1)} = \frac{\sum_{i=1}^{n} y_i x_{i,1}}{\sum_{i=1}^{n} (x_{i,1}^2)} \tag{3.2}$$

em que y_i, $x_{i,1}$ são os valores de y e x_1 para o i-ésimo objeto, respectivamente.

Utilizando uma notação similar, o modelo com duas variáveis é escrito como $y = \beta_1^{(2)} x_1 + \beta_2^{(2)} x_2 + \epsilon^{y|x_1,x_2}$. Para obter $b_1^{(2)}$ e $b_2^{(2)}$, x_2 é inicialmente regredida em x_1 segundo um modelo da forma

$$x_2 = \delta_1^{x_2|x_1} x_1 + \epsilon^{x_2|x_1} \tag{3.3}$$

em que o coeficiente $\delta_1^{x_2|x_1}$ pode ser estimado por regressão univariada como

$$\widehat{\delta}_1^{x_2|x_1} = \frac{\sum_{i=1}^{n} x_{i,2} x_{i,1}}{\sum_{i=1}^{n} (x_{i,1}^2)} \tag{3.4}$$

Então, como mostrado em (GUSNANTO *et al.*, 2003), $b_1^{(2)}$ e $b_2^{(2)}$ podem ser obtidos como

$$b_2^{(2)} = \frac{\sum_{i=1}^{n} e_i^{y|x_1} x_{i,2}}{\sum_{i=1}^{n} e_i^{x_2|x_1} x_{i,2}}, \qquad b_1^{(2)} = b_1^{(1)} - \widehat{\delta}_1^{x_2|x_1} b_2^{(2)} \tag{3.5}$$

em que $e_i^{y|x_1} = y_i - b_1^1 x_{i,1}$ e $e_i^{x_2|x_1} = x_{i,2} - \widehat{\delta}_1^{x_2|x_1} x_{i,1}$

O procedimento pode ser generalizado para obter um modelo com m variáveis a partir de um modelo com $(m-1)$ variáveis, com m variando de dois até M. A nova variável x_m

a ser inserida no modelo é inicialmente regredida em $x_1, x_2, \ldots, x_{m-1}$ de acordo com um modelo da forma

$$x_m = \delta_1^{x_m|x_1,\ldots,x_{m-1}} x_1 + \delta_2^{x_m|x_1,\ldots,x_{m-1}} x_2 + \ldots + \delta_{m-1}^{x_m|x_1,\ldots,x_{m-1}} x_{m-1} + \epsilon^{x_m|x_1,\ldots,x_{m-1}} \quad (3.6)$$

Os coeficientes $b_1^{(m)}, b_2^{(m)}, \ldots, b_m^{(m)}$ do modelo com m variáveis podem então ser estimados como

$$b_m^{(m)} = \frac{\sum_{i=1}^n e^{y|x_1,\ldots,x_{k-1}} x_{i,m}}{\sum_{i=1}^n e_i^{x_m|x_1,\ldots,x_{m-1}} x_{i,m}} \quad (3.7)$$

$$b_{m-j}^{(m)} = b_{m-j}^{(m-1)} - \widehat{\delta}_{m-j}^{x_m|x_1,\ldots,x_{m-1}} b_m^{(m)}, \qquad j = 1, \ldots, m-1 \quad (3.8)$$

sendo

$$e_i^{x_m|x_1,\ldots,x_{m-1}} = x_{i,m} - (\widehat{\delta}_1^{x_m|x_1,\ldots,x_{m-1}} x_{i,1} + \widehat{\delta}_2^{x_m|x_1,\ldots,x_{m-1}} x_{i,2} + \ldots + \widehat{\delta}_{m-1}^{x_m|x_m,\ldots,x_{m-1}} x_{i,m-1}) \quad (3.9)$$

$$e_i^{y|x_1,\ldots,x_{m-1}} = y_i - (b_1^{(m-1)} x_{i,1} + b_2^{(m-1)} x_{i,2} + \ldots + b_{m-1}^{(m-1)} x_{i,m-1}) \quad (3.10)$$

3.1 Exemplo numérico da regressão sequencial

Sejam $\mathbf{X}$ (3×3) e $\mathbf{y}$ (3×1) as matrizes geradas aleatoriamente abaixo:

$$\mathbf{X} = \begin{bmatrix} 0.9528 & 0.5982 & 0.8368 \\ 0.7041 & 0.8407 & 0.5187 \\ 0.9539 & 0.4428 & 0.0222 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 0.3759 \\ 0.8986 \\ 0.4290 \end{bmatrix}. \quad (3.11)$$

Inicialmente supõe-se um cenário em que seja utilizada apenas a primeira coluna de $\mathbf{X}$, a saber $\mathbf{x}_1 = [0.9528 \quad 0.7041 \quad 0.9539]'$. O coeficiente de regressão $b_1^{(1)}$ é então obtido

substituindo os respectivos valores na Equação (3.2), o que leva a

$$\begin{aligned} b_1^{(1)} &= \frac{(0.3759 \times 0.9528) + (0.8986 \times 0.7041) + (0.4290 \times 0.9539)}{(0.9528)^2 + (0.7041)^2 + (0.9539)^2} \\ &= 0.6052. \end{aligned} \tag{3.12}$$

Agora adiciona-se a segunda coluna de $\mathbf{X}$, isto é $\mathbf{x}_2 = [0.5982 \quad 0.8407 \quad 0.4428]'$. Para obter os coeficientes de regressão $\beta_1^{(2)}$, $\beta_2^{(2)}$ deve-se primeiramente regredir $\mathbf{x}_2$ em $\mathbf{x}_1$ de acordo com a Equação (3.4) e encontrar $\mathbf{e}^{y|x_1}$ e $\mathbf{e}^{x_2|x_1}$, como mostrado abaixo:

$$\begin{aligned} \widehat{\delta}^{x_2|x_1} &= \frac{(0.5982 \times 0.9528) + (0.8407 \times 0.7041) + (0.4428 \times 0.9539)}{(0.9528)^2 + (0.7041)^2 + (0.9539)^2} \\ &= 0.6848, \end{aligned} \tag{3.13}$$

$$\mathbf{e}^{y|x_1} = \begin{bmatrix} 0.3759 \\ 0.8986 \\ 0.4290 \end{bmatrix} - 0.6052 \begin{bmatrix} 0.9528 \\ 0.7041 \\ 0.9539 \end{bmatrix} = \begin{bmatrix} -0.2007 \\ 0.4725 \\ -0.1483 \end{bmatrix} \tag{3.14}$$

e

$$\mathbf{e}^{x_2|x_1} = \begin{bmatrix} 0.5982 \\ 0.8407 \\ 0.4428 \end{bmatrix} - 0.6848 \begin{bmatrix} 0.9528 \\ 0.7041 \\ 0.9539 \end{bmatrix} = \begin{bmatrix} -0.0534 \\ 0.3586 \\ -0.2104 \end{bmatrix}. \tag{3.15}$$

Substituindo os valores encontrados na Equação (3.5), tem-se:

$$\begin{aligned} b_2^{(2)} &= \frac{(-0.2007 \times 0.5982) + (0.4725 \times 0.8407) + (-0.1483 \times 0.4428)}{(-0.0534 \times 0.5982) + (0.3586 \times 0.8407) + (-0.2104 \times 0.4428)} \\ &= 1.2032, \end{aligned} \tag{3.16}$$

$$b_1^{(2)} = 0.6052 - (0.6848 \times 1.2032)$$

$$= \quad -0.2188. \tag{3.17}$$

Finalmente, será adicionada a terceira coluna de $\mathbf{X}$, isto é $\mathbf{x}_3 = [0.8368 \quad 0.5187 \quad 0.0222]'$. O cálculo do coeficiente de regressão será feito através da estimativa de mínimos quadrados (LAWSON; HANSON, 1974) como mostra a Equação (2.3). De modo semelhante ao modelo univariado, como mostra a Equação (3.6), deve-se primeiramente regredir $\mathbf{x}_3$ em $\mathbf{x}_2$ e $\mathbf{x}_1$:

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})^{-1} &= \left(\begin{bmatrix} 0.9528 & 0.7041 & 0.9539 \\ 0.5982 & 0.8407 & 0.4428 \end{bmatrix} \times \begin{bmatrix} 0.9528 & 0.5982 \\ 0.7041 & 0.8407 \\ 0.9539 & 0.4428 \end{bmatrix} \right)^{-1} \\
&= \left(\begin{bmatrix} 2.2134 & 1.5842 \\ 1.5842 & 1.2607 \end{bmatrix} \right)^{-1} \\
&= \begin{bmatrix} 3.0998 & -3.8952 \\ -3.8952 & 5.6879 \end{bmatrix}
\end{aligned} \tag{3.18}$$

$$\begin{aligned}
\widehat{\delta}^{x_3|x_1,x_2} &= \begin{bmatrix} 3.0998 & -3.8952 \\ -3.8952 & 5.6879 \end{bmatrix} \times \begin{bmatrix} 0.9528 & 0.7041 & 0.9539 \\ 0.5982 & 0.8407 & 0.4428 \end{bmatrix} \times \begin{bmatrix} 0.8368 \\ 0.5187 \\ 0.0222 \end{bmatrix} \\
&= \begin{bmatrix} -0.0176 \\ 0.7728 \end{bmatrix}
\end{aligned} \tag{3.19}$$

Para obter $\mathbf{e}^{x_3|x_2,x_1}$ e $\mathbf{e}^{y|x_2,x_1}$ tem-se:

$$\begin{aligned}\mathbf{e}^{x_3|x_1,x_2} &= \begin{bmatrix}0.3759\\0.8986\\0.4290\end{bmatrix} - \left(-0.0176\begin{bmatrix}0.9528\\0.7041\\0.9539\end{bmatrix} + 0.7728\begin{bmatrix}0.5982\\0.8407\\0.4428\end{bmatrix}\right)\\ &= \begin{bmatrix}0.3913\\-0.1187\\-0.3033\end{bmatrix}\end{aligned} \tag{3.20}$$

$$\begin{aligned}\mathbf{e}^{y|x_1,x_2} &= \begin{bmatrix}0.8368\\0.5187\\0.0222\end{bmatrix} - \left(-0.2188\begin{bmatrix}0.9528\\0.7041\\0.9539\end{bmatrix} + 1.2032\begin{bmatrix}0.5982\\0.8407\\0.4428\end{bmatrix}\right)\\ &= \begin{bmatrix}-0.1354\\0.0410\\0.1049\end{bmatrix}\end{aligned} \tag{3.21}$$

Os coeficientes de regressão $b_1^{(3)}$, $b_2^{(3)}$ e $b_3^{(3)}$ são obtidos substituindo os respectivos valores na Equação (3.7):

$$\begin{aligned}b_3^{(3)} &= \frac{(-0.1354\times 0.8368)+(0.0410\times 0.5187)+(-0.1049\times 0.0222)}{0.3913\times 0.8368)+(-0.1187\times 0.5187)+(-0.3033\times 0.0222)}\\ &= -0,3459,\end{aligned} \tag{3.22}$$

$$\begin{aligned}b_2^{(3)} &= 1.2032 - 0.7728\times(-0.3459)\\ &= 1.4706\end{aligned} \tag{3.23}$$

$$\begin{aligned} b_1^{(3)} &= -0.2188 - (-0.0176) \times (-0.3459) \\ &= -0.2249. \end{aligned} \tag{3.24}$$

Finalmente, para fins de comparação, a regressão clássica por mínimos quadrados foi calculada para o exemplo em questão conforme mostrado abaixo.

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \left(\begin{bmatrix} 0.9528 & 0.7041 & 0.9539 \\ 0.5982 & 0.8407 & 0.4428 \\ 0.8368 & 0.5187 & 0.0222 \end{bmatrix} \times \begin{bmatrix} 0.9528 & 0.5982 & 0.8368 \\ 0.7041 & 0.8407 & 0.5187 \\ 0.9539 & 0.4428 & 0.0222 \end{bmatrix} \right)^{-1} \\ &= \left(\begin{bmatrix} 2.3134 & 1.5842 & 1.1837 \\ 1.5842 & 1.2607 & 0.9465 \\ 1.1837 & 0.9465 & 0.9698 \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} 3.1010 & -3.9476 & 0.0678 \\ -3.9476 & 7.9925 & -2.9821 \\ 0.0678 & -2.9821 & 3.8587 \end{bmatrix} \end{aligned} \tag{3.25}$$

$$\begin{aligned} \mathbf{b} &= \begin{bmatrix} 3.1010 & -3.9476 & 0.0678 \\ -3.9476 & 7.9925 & -2.9821 \\ 0.0678 & -2.9821 & 3.8587 \end{bmatrix} \times \begin{bmatrix} 0.9528 & 0.7041 & 0.9539 \\ 0.5982 & 0.8407 & 0.4428 \\ 0.8368 & 0.5187 & 0.0222 \end{bmatrix} \times \begin{bmatrix} 0.3759 \\ 0.8986 \\ 0.4290 \end{bmatrix} \\ &= \begin{bmatrix} -0.2249 \\ 1.4706 \\ -0.3459 \end{bmatrix} \end{aligned} \tag{3.26}$$

Como se pode observar, o resultado expresso em (3.26) é o mesmo ao qual se havia

chegado nas Equações (3.22), (3.23) e (3.24).

3.2 Exemplo da redução do custo computacional

Para constatar a redução do custo computacional, foi gerada aleatoriamente uma matriz $\mathbf{X}$ de (500×500) e um vetor $\mathbf{y}$ de (500×1).

Utilizando a regressão sequencial, estes dados foram testados variando a quantidade de variáveis de $2, 3, \ldots, 300, 350, \ldots, 500$. O tempo gasto na execução é apresentado na Figura 3.1.

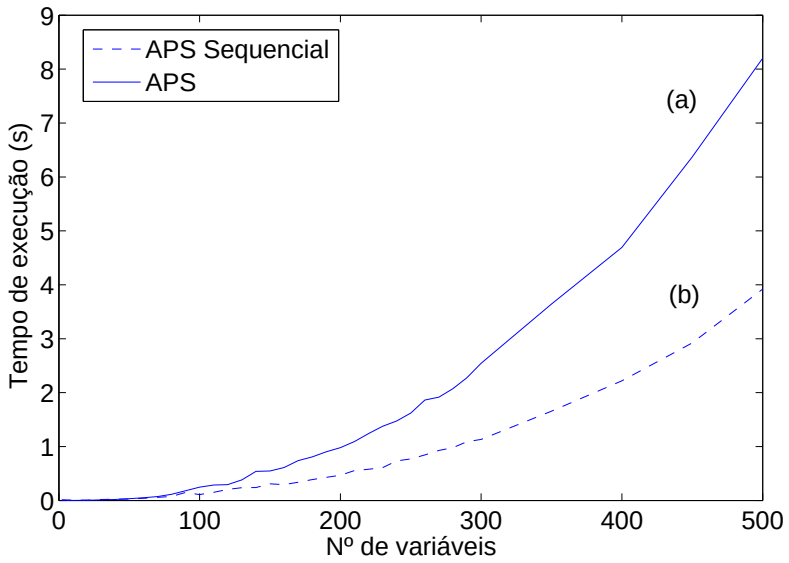


FIGURA 3.1 – Tempo computacional da regressão com (a) configuração original e (b) com a formulação proposta.

É possível observar uma redução considerável do tempo gasto ao utilizar a formulação de regressão seqüencial (b) ao invés da técnica tradicional que requer o cálculo da inversa (a); constatando a eficácia da técnica.

O uso da técnica de regressões sequenciais no APS foi relatada em (SOARES *et al.*, 2010). Nesse artigo, apresentou-se um estudo de caso envolvendo dados espectrofoto-

métricos de amostras de trigo. Como resultado, obteve-se ganhos de até 10 vezes com respeito ao APS tradicional.

Os testes realizados com os algoritmos utilizados neste trabalho foram implementados em ambiente *MATLAB*® versão 7.9.0 (R2009b) (*The MathWorks, Inc.*). A configuração do computador utilizado consiste em um processador *Intel*® *Core i5*^{TM} (*Intel Corp.*), com *clock* de (2×) 2.27 GHz e 4GB de memória RAM (*Random Access Memory*), com sistema operacional *Windows7*® (*Microsoft Corp.*) *Home Premium*.

4 Material e Métodos

4.1 Conjunto de dados

Neste trabalho foi empregado um conjunto de dados de espectros de reflectância no infravermelho próximo (*Near Infrared*, NIR) publicamente disponíveis no site: `ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas/` (Kalivas, citado por (FORINA *et al.*, 2007)).

Este conjunto de dados contém 100 amostras de trigo para as quais foram medidas espectros NIR de reflectância difusa em três espectrofotômetros diferentes, porém serão utilizados somente os dados referentes ao primeiro deles, com faixa espectral de 1100 – 2498nm e resolução de 2nm resultando num total de 701 variáveis espectrais (FORINA *et al.*, 2007). O conjunto de dados também contém os teores de umidade e proteína para cada uma das 100 amostras. A faixa de variação dessas propriedades é mostrada na Tabela 4.1

TABELA 4.1 – Faixa de variação das propriedades analisadas do trigo (%m/m).

Propriedade	Umidade	Proteína
Faixa de variação	12,45-17,36	7,75-14,28

A Figura 4.1 apresenta os 100 espectros do conjunto de dados. Como pode ser observado na Figura 4.1*a*, os espectros apresentam deslocamento de linha de base. Sabe-se

que tais deslocamentos tipicamente estão relacionados a variações ambientais e/ou de posicionamento da amostra e não a alterações de sua composição química (BEEBE; PELL; SEASHOLTZ, 1998). Para corrigir tal problema, optou-se por trabalhar com os espectros derivativos obtidos com o uso de um filtro de Savitzky-Golay empregando polinômio de 2ª grau e janela de 21 pontos (FORINA *et al.*, 2007). Como resultado o número de variáveis foi reduzido para 681. Os espectros derivativos estão apresentados na Figura 4.1*b*. Em seguida foram eliminadas as variáveis cuja máxima intensidade de sinal para todos os espectros derivativos não excedessem 2% da intensidade máxima do sinal no conjunto total de dados (HONORATO *et al.*, 2005), resultando 652 variáveis. Os espectros resultantes que foram utilizados ao longo do trabalho são apresentados na Figura 4.2.

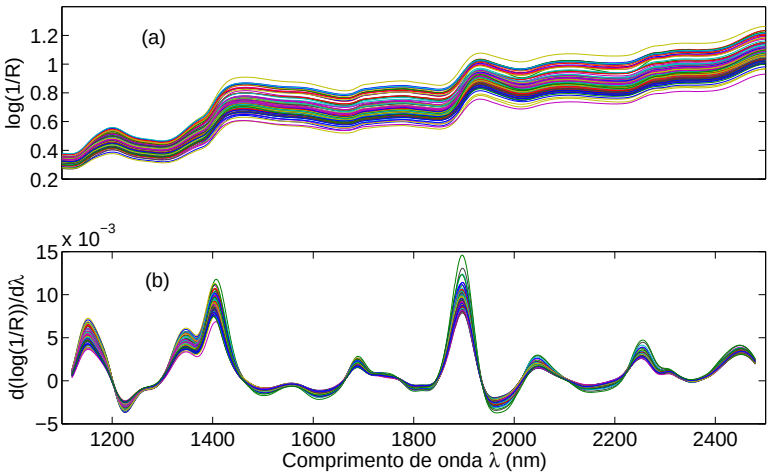


FIGURA 4.1 – Espectros das 100 amostras de trigo: (a) Brutos e (b) Derivativos.

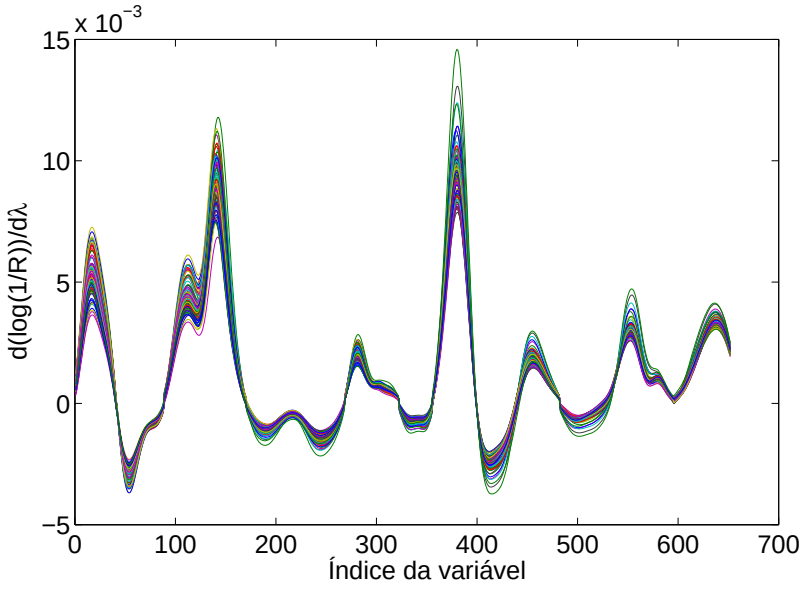


FIGURA 4.2 – Espectros derivativos das 100 amostras de trigo após o descarte das variáveis cuja máxima intensidade não excedessem 2% da intensidade máxima do sinal no conjunto total.

No intuito de avaliar a capacidade preditiva dos modelos a serem obtidos dividiu-se os dados em um conjunto de modelagem e outro de predição. Para esse propósito aplicou-se o algoritmo Kennard-Stone (KENNARD; STONE, 1969) nos espectros derivativos (Figura 4.2) das 100 amostras disponíveis. Tal algoritmo realiza uma amostragem uniforme porque procura maximizar a menor distância entre os objetos selecionados em um certo conjunto.

Aplicando-se o algoritmo de Kennard-Stone aos espectros da Figura 4.2 os dados foram divididos em 70 objetos de modelagem e 30 de predição. Um esquema desta divisão é apresentado na Figura 4.3. O conjunto de predição em nenhum momento é utilizado na obtenção dos modelos, ficando reservado exclusivamente para o processo de avaliação.

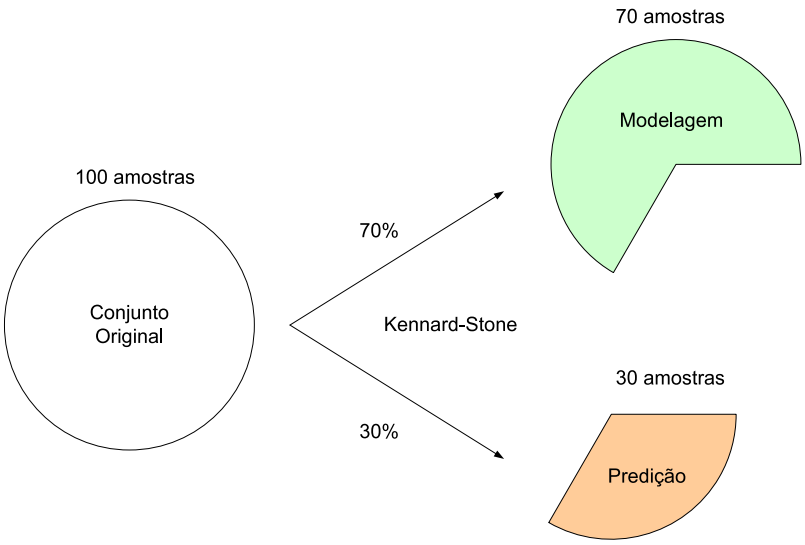


FIGURA 4.3 – Divisão dos dados de trigo em modelagem e previsão.

O resultado da aplicação do algoritmo Kennard-Stone pode ser visualizado na Figura 4.4. Observa-se em (a) e (d) os histogramas referentes ao teor de umidade e de proteína. Os objetos selecionados para modelagem e predição encontram-se indicados em (b) e (c).

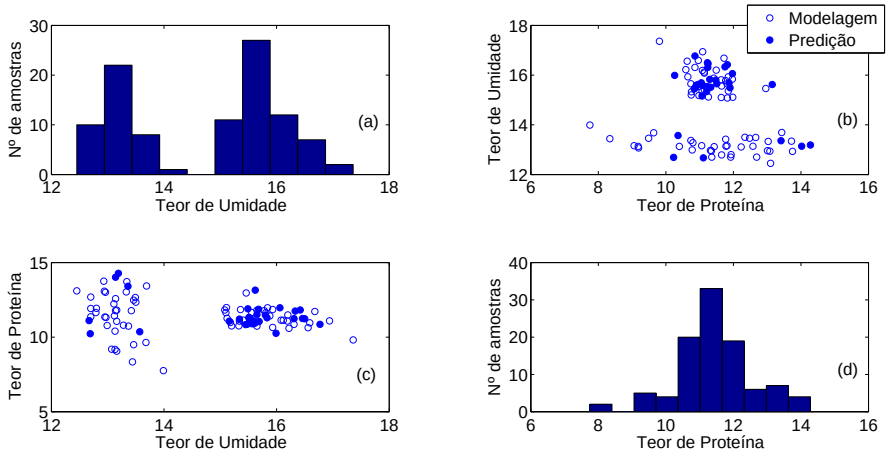
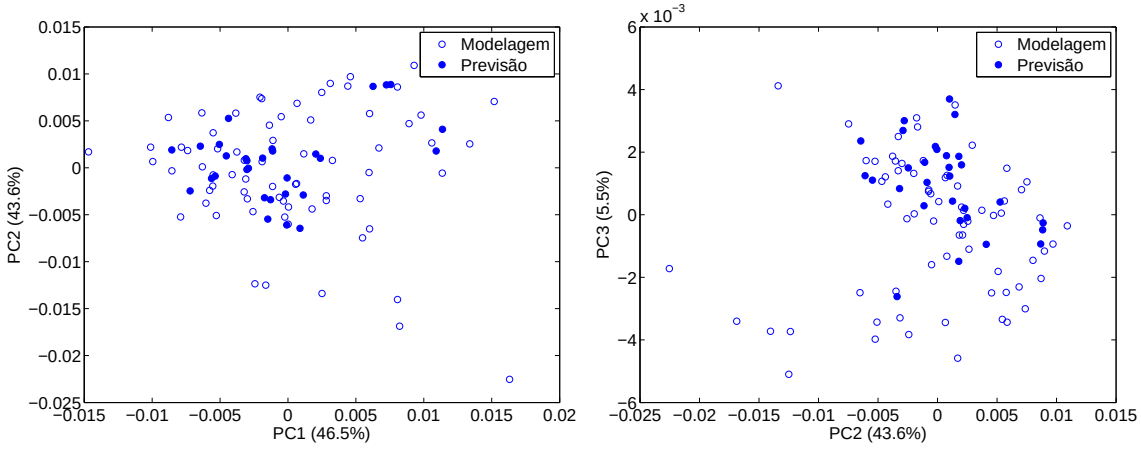


FIGURA 4.4 – Teores de umidade e proteína com indicação dos objetos selecionados para modelagem e predição.

Pode-se notar que os valores de proteína encontram-se agrupados em torno de 11 %m/m, em contra partida os valores de umidade apresentam uma separação em dois gru-

pos. Poderia-se analisar essa heterogeneidade da umidade, gerando modelos separados, mas optou-se neste trabalho por tratar o conjunto como um só. Também não há evidencias de *outliers* que necessitassem serem descartados do conjunto de dados. Observa-se também que os objetos de predição representam adequadamente o conjunto de dados (tanto na proteína quanto nos dois grupos da umidade) permitindo então uma forma apropriada de avaliação da capacidade preditiva dos modelos.

Tais considerações são corroboradas pela Figura 4.5 que mostra os gráficos dos escores obtidos pela análise por Componentes Principais (PC) aplicada aos 100 espectros pré-processados (FERREIRA *et al.*, 1999). Mais uma vez não há indícios de *outliers* que nitidamente devam ser excluídos do conjunto de dados, e novamente observa-se que o conjunto de predição se distribui de maneira adequada dentro do conjunto de dados.



(a) PC1 × PC2 (b) PC2 × PC3

FIGURA 4.5 – Gráfico dos escores de (a) PC1 × PC2 e (b) PC2 × PC3 para as 100 amostras de trigo.

Os valores entre parênteses no gráfico mostram a variância explicada em cada PC. As três variâcias explicadas somadas explicam $95,6\%$ dos dados, o que é suficiente para a realização dessa análise exploratória.

4.2 Algoritmos empregados

4.2.1 APS e APS-VC

Para a aplicação do APS, os 70 objetos de modelagem foram divididos em um conjunto de calibração e um de validação contendo 49 e 21 objetos, respectivamente. Para esse propósito, aplicou-se novamente o algoritmo de Kennard-Stone, como ilustrado na Figura 4.6.

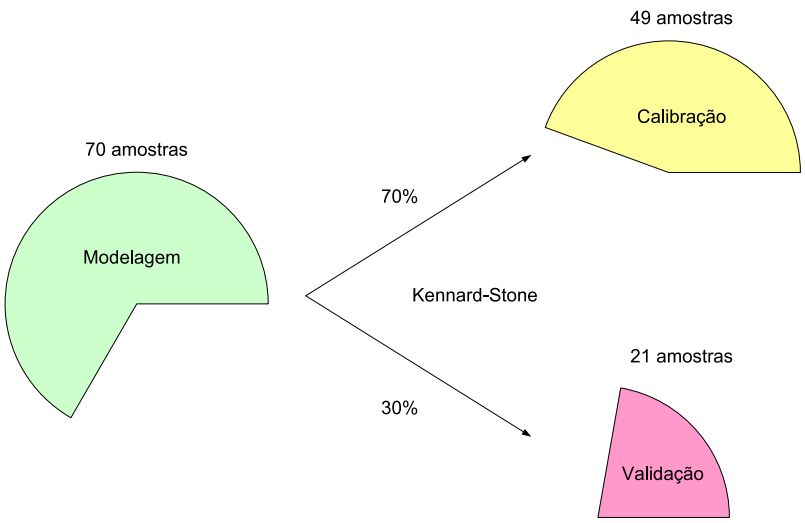


FIGURA 4.6 – Divisão dos dados de trigo em modelagem e previsão.

Vale salientar que tal separação não é necessária para aplicação do APS-VC. Nesse caso, utilizam-se diretamente os dados de modelagem, como indicado na Figura 4.7*a*.

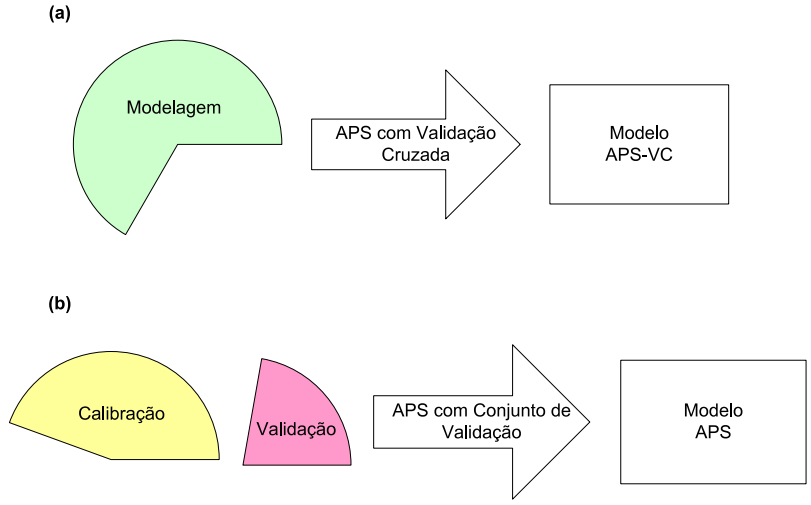


FIGURA 4.7 – Obtenção dos modelos (a) APS-VC e (b) APS.

O APS foi configurado para utilizar todas as $M = 48$ variáveis disponíveis, em contra partida os modelos APS-VC que utilizavam todas as M variáveis selecionaram um número excessivo de variáveis resultando em modelos menos parcimoniosos. Por esse motivo a quantidade máxima de variáveis utilizadas no APS-VC seja $\overline{M} = 35$ (metade do número de objetos de modelagem). Para ambos os algoritmos, a quantidade de objetos de calibração foi fixado em $N_c = 49$ (70% do conjunto de modelagem).

4.2.2 APS-*Subagging* e APS-VC-*Subagging*

O APS-*Subagging* foi implementado como mostra a Figura 4.9. Neste caso, ao contrario da aplicação do APS descrita na Seção 4.2.1, a separação entre os conjuntos de calibração e validação foi realizada de forma aleatória repetindo-se o procedimento P vezes obtendo-se modelos que posteriormente foram promediados resultando no modelo APS-*Subagging*.

Como descrito em (GALVãO *et al.*, 2006) para valores de $P > 50$ não houve ganhos consideráveis. Dessa forma neste trabalho investigou-se o efeito de se variar o número de

iterações ate o máximo de $P = 50$.

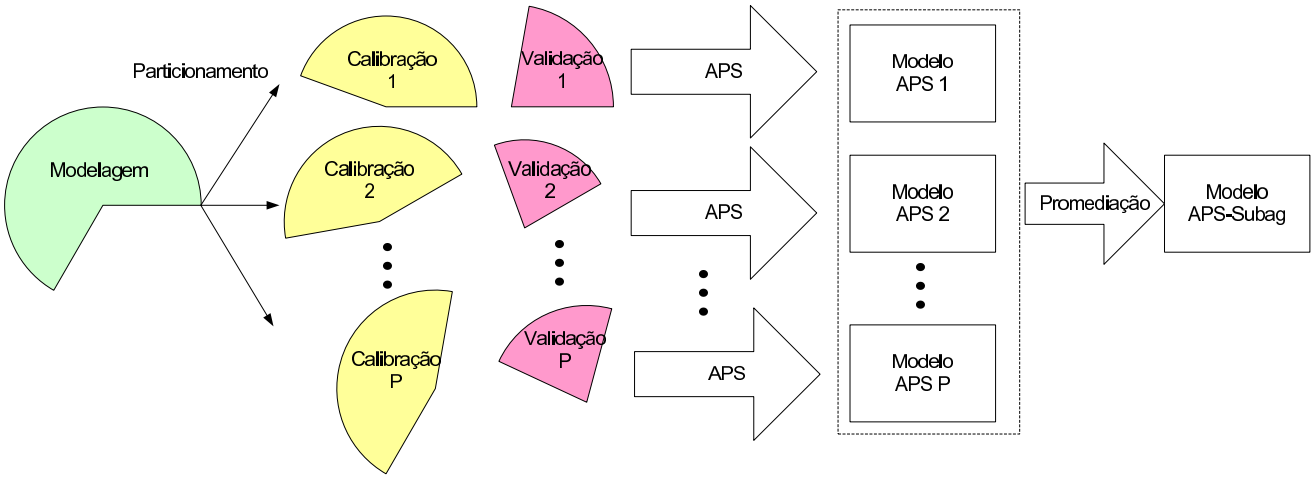


FIGURA 4.8 – Obtenção do modelo APS-*Subagging*.

Na implementação do APS-VC-*Subagging* conjuntos de calibração foram aleatóriamente extraídos do conjunto de modelagem de modo a gerar modelos que posteriormente foram promediados resultando no modelo APS-VC-*Subagging*, como ilustrado na Figura 4.9.

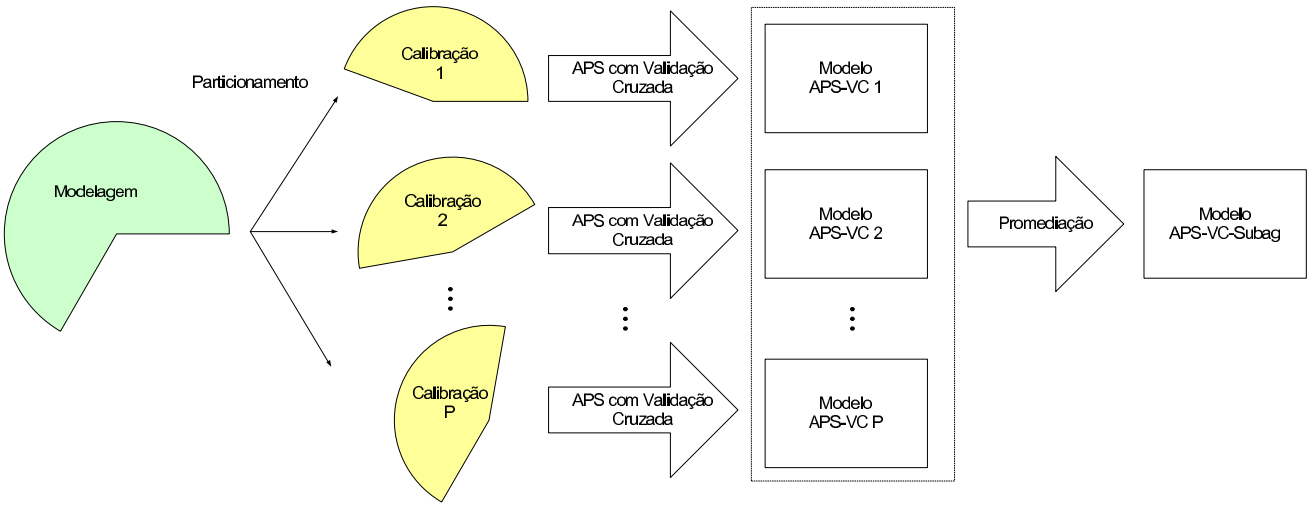


FIGURA 4.9 – Obtenção do modelo APS-VC-*Subagging*.

Vale salientar que ambos algoritmos serão testados variando-se a quantidade de objetos no conjunto de calibração (N_c) e se o algoritmo segue o mesmo comportamento se as

sementes do gerador de números aleatórios utilizadas forem alteradas.

4.3 Forma de Análise dos Resultados

Os resultados obtidos foram avaliados por dois parâmetros relacionado à capacidade de previsão e à sensibilidade ao ruído. A capacidade de previsão dos modelos foi aferida com base na raiz quadrada do erro quadrático médio de predição (*Root Mean Square Error of Prediction*, RMSEP) obtido com o conjunto de predição. O RMSEP é definido como:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N_p}(y_i - \hat{y}_i)^2}{N_p}} \tag{4.1}$$

em que y_i e $\hat{y}_i$ são o valor de referência e o dos valor predito da propriedade em consideração para o i-ésimo objeto de predição, sendo N_p o número de amostras de predição (30 neste caso).

Já a sensibilidade ao ruído foi avaliada com base na norma 2 do vetor de coeficientes de regressão ($||\mathbf{b}||$), que pode ser definida como:

$$||\mathbf{b}|| = \sqrt{\sum_{k=1}^{K}(\mathbf{b}(k))^2} \tag{4.2}$$

sendo $K = 652$ a quantidade de variáveis em questão.

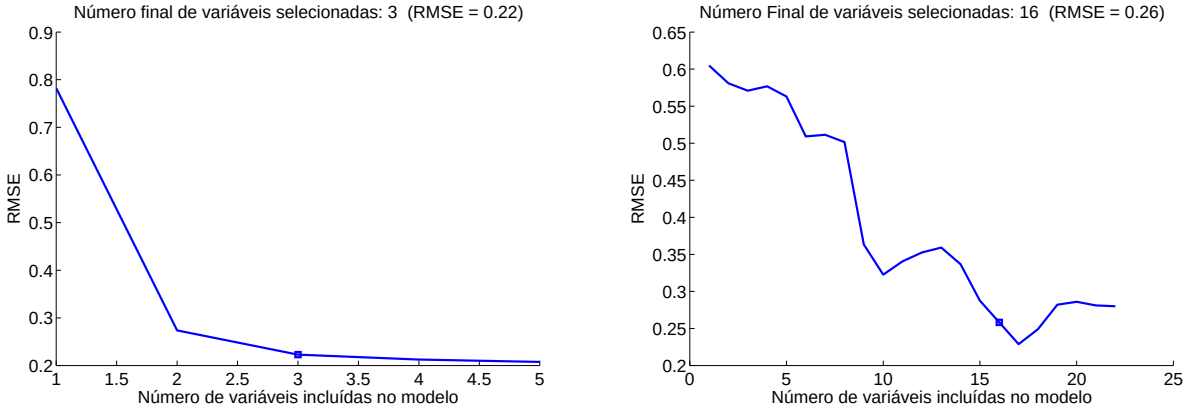
5 Resultados

Neste capítulo apresentam-se inicialmente os resultados obtidos com o algoritmo das projeções sucessivas sem uso de *subagging*, considerando as alternativas de validação por série de teste (APS) e validação cruzada (APS-VC). Em seguida, tais resultados são comparados com os obtidos mediante o uso de *subagging*, levando-se em conta os critérios de capacidade preditiva e sensibilidade a ruído. Em particular, avalia-se o efeito de se variar a fração de subamostragem empregada. Por fim, os principais resultados são resumidos, apresentando-se uma análise comparativa das técnicas estudadas.

5.1 APS e APS-VC

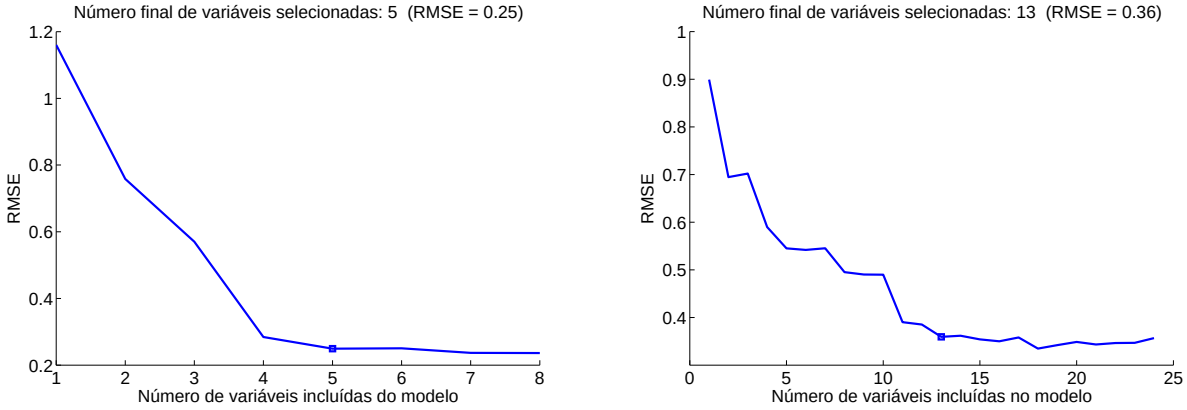
Conforme indicado na Figura 5.1, a aplicação do APS resultou na seleção de 3 variáveis para de umidade e de 16 variáveis para proteína. Os gráficos apresentados nesta figura trazem os valores de RMSE (*Root Mean Square Error*) obtidos para o conjunto de validação na fase 3 do algoritmo. Tal métrica é definida como $RMSE = \sqrt{PRESS/N_v}$. Como se pode notar, o número de variáveis selecionadas não corresponde ao valor mínimo de RMSE. Com efeito, o teste F empregado na fase 3 indica que é possível tomar um ponto anterior da curva sem que o RMSE seja significativamente maior que o mínimo encontrado.

Vale salientar que o número máximo de variáveis considerado nesta avaliação é diferente para umidade e proteína, tendo em vista que a fase 3 opera sobre o resultado da seleção realizada na fase 2 (5 variáveis para umidade e 22 variáveis para proteína). Resultados similares foram obtidos com o uso de validação cruzada, como apresentado na Figura 5.2.



(a) Umidade (b) Proteína

FIGURA 5.1 – Valores de RMSE ($\%m/m$) obtidos na fase 3 do APS.



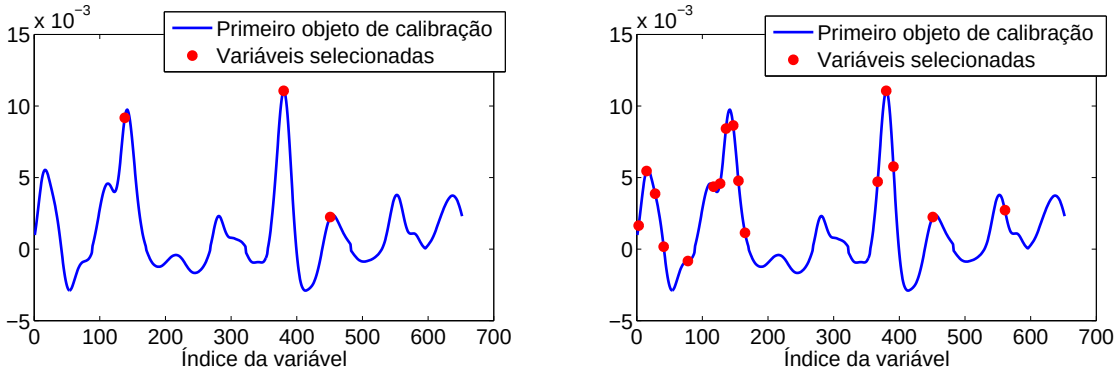
(a) Umidade (b) Proteína

FIGURA 5.2 – Valores de RMSE ($\%m/m$) obtidos na fase 3 do APS-VC.

A Figura 5.3 apresenta as variáveis selecionadas ao final do APS. Tais variáveis encontram-se bem espalhas pelo espectro não havendo concentração em regiões específicas. Este resultado era esperado, tendo em vista a característica do APS de não selecionar variáveis redundantes. Vale salientar que as três variáveis de umidade na Figura 5.3(a)

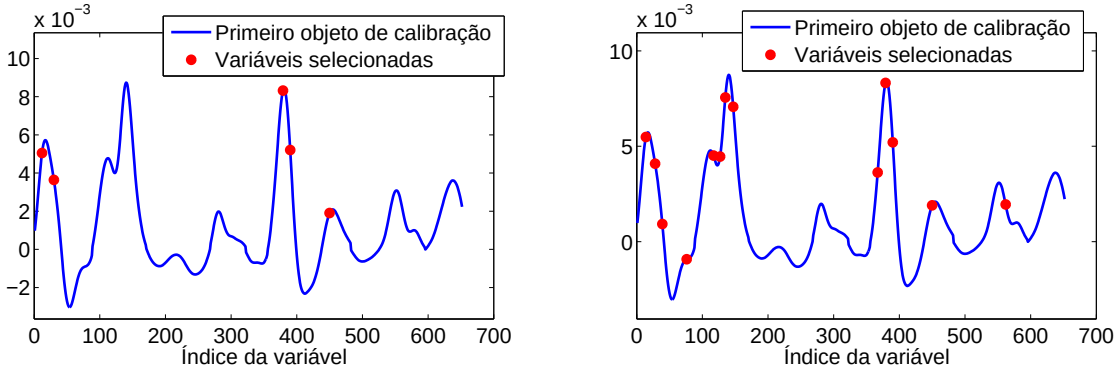
estão próximas de variáveis selecionadas para proteína, como visto em 5.3(b). Contudo para proteína observa-se que foi necessária a inclusão de variáveis adicionais.

Boa parte das variáveis selecionadas pelo APS-VC encontram-se próximas às variáveis selecionadas pelo APS, como mostra a Figura 5.4. Isso indica uma certa "robustez" do algoritmo com respeito ao procedimento de validação empregado. Contudo algumas diferenças podem ser apontadas. Por exemplo, a primeira variável selecionada (contando da esquerda para direita) da Figura 5.3(a) não foi incluída no modelo pelo APS-VC como mostra a Figura 5.4(a). Isso mostra que a forma como o conjunto de modelagem é utilizado para guiar o processo de seleção de variáveis influencia o resultado, em certo grau, o que é uma motivação para se explorar técnicas de reamostragem.



(a) Umidade (b) Proteína

FIGURA 5.3 – Variáveis selecionadas pelo APS.



(a) Umidade (b) Proteína

FIGURA 5.4 – Variáveis selecionadas pelo APS-VC.

A Tabela 5.1 compara os modelos obtidos por APS e APS-VC em termos da norma dos coeficientes de regressão ($||\mathbf{b}||$) e do valor de RMSEP obtido no conjunto de predição. Tanto a $||\mathbf{b}||$ quanto o RMSEP foram ligeiramente menores para o APS-VC, o que corrobora a importância de se incluir o uso de validação cruzada no estudo aqui realizado. Contudo, vale salientar que o custo computacional demandado é maior para o APS-VC devido ao fato de se realizar N_{model} vezes a RLM para a avaliação de cada subconjunto de variáveis.

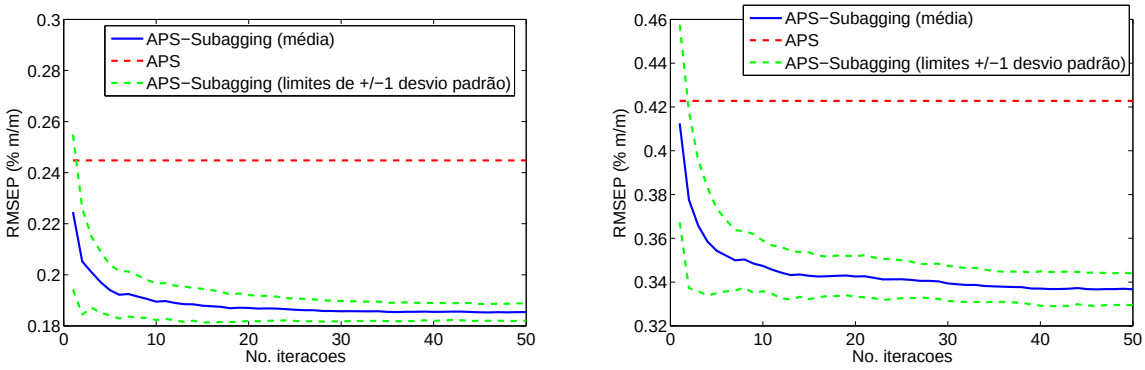
TABELA 5.1 – Valores de RMSEP e $||\mathbf{b}||$ obtidos com APS e APS-VC para Umidade e Proteína (%m/m).

	Umidade		**Proteína**	
	RMSEP	$\|\|\mathbf{b}\|\|$	*RMSEP*	$\|\|\mathbf{b}\|\|$
APS	0,24	4452	0,42	30880
APS-VC	0,20	4084	0,39	29656

5.2 Uso do *subagging*

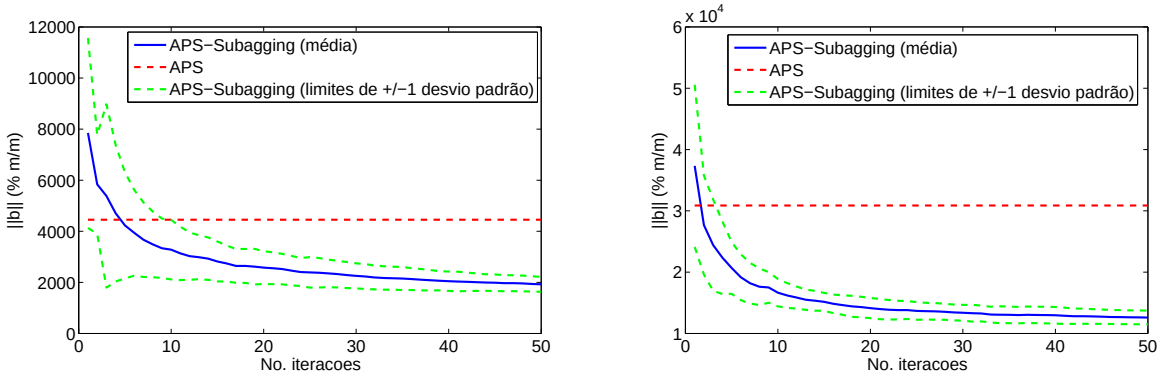
Neste estudo, fixou-se inicialmente $N_c = 49$ para o APS-*Subagging* e APS-VC-*Subagging*. A Figura 5.5 apresenta o RMSEP obtido com o APS-*Subagging*, para (a) umidade e (b) proteína, em função do número de iterações de subamostragem. A linha cheia representa o resultado médio de 30 realizações do *subagging*, obtidas variando-se a semente do gerador de números aleatórios empregado na subamostragem. As linhas tracejadas verdes mostram os limites (superior e inferior) de ± 1 desvio-padrão, enquanto a linha tracejada vermelha indica o resultado obtido pelo APS. Observa-se uma melhora para ambas propriedades, mesmo considerando o limite de $+1$ desvio-padrão. Vale salientar também que não há melhora expressiva para valores de $P > 30$. Da mesma forma, a Figura 5.6 mostra o resultado médio das 30 realizações, porém agora da $||\mathbf{b}||$ para as duas propriedades em

questão. Verifica-se os mesmo resultados, em que a média e grande parte da região de ± 1 desvio-padrão encontram-se abaixo dos resultados obtidos sem o uso do *subagging*. E também não é observado nenhuma melhora expressiva para $P > 30$.



(a) Umidade (b) Proteína

FIGURA 5.5 – RMSEP obtido com 30 realizações do APS-*Subagging*.

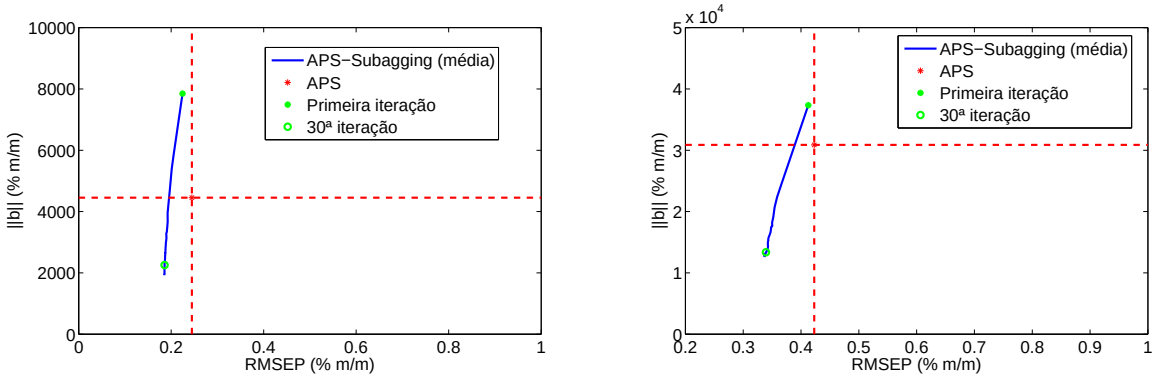


(a) Umidade (b) Proteína

FIGURA 5.6 – $||\mathbf{b}||$ obtida com 30 realizações do APS-*Subagging*.

A Figura 5.7 mostra uma avaliação multiobjetivo para o resultado médio das 30 iterações do *subagging*. A linha cheia apresenta a evolução desde a primeira iteração até a ultima. Como se pode notar, dado um número suficientemente grande de iterações, é possível obter um resultado que domina aquele obtido sem o *subagging* (representado pela linhas tracejadas vermelhas). Vale ainda salientar que há pouca alteração entre $P = 30$ e $P = 50$ iterações, como comentado anteriormente.

A linha cheia mostra a evolução da primeira iteração até a 30ª trespassando as li-

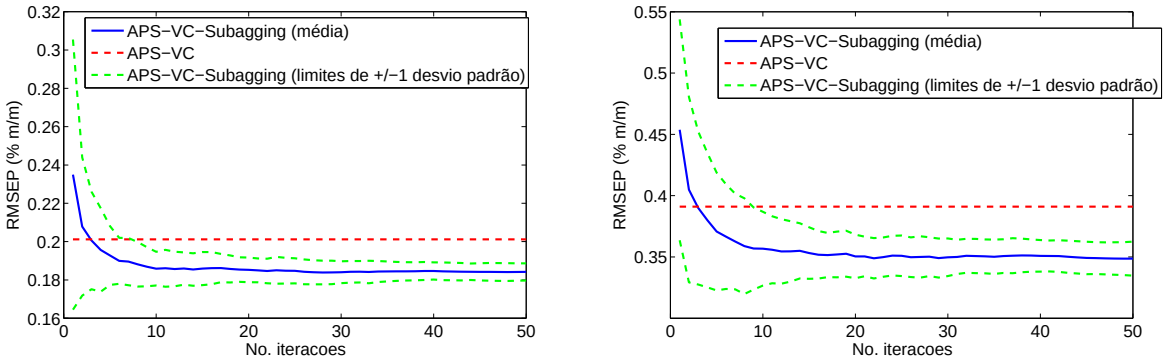


(a) Umidade (b) Proteína

FIGURA 5.7 – Avaliação multiobjetivo do APS-*Subagging* (média de 30 realizações).

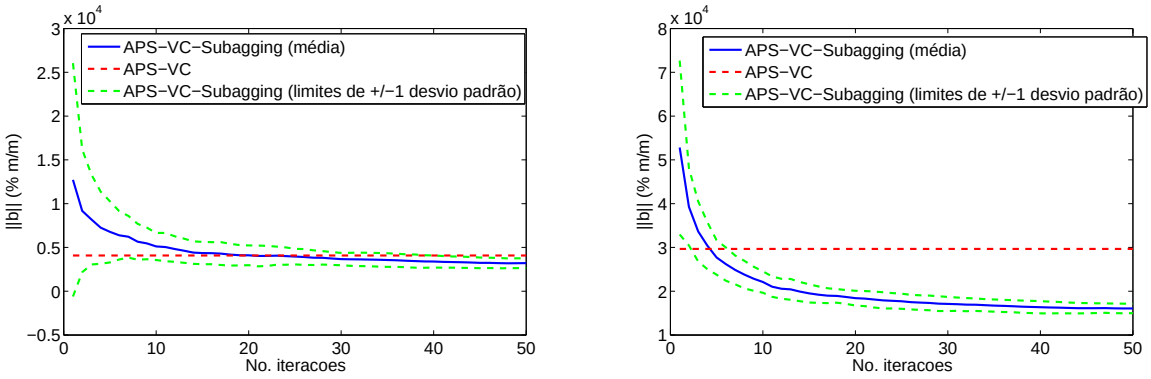
nhas tracejadas vermelhas, que simbolizam o modelos sem o *subagging*, evidenciando a convergência para a melhora em ambos os parâmetros.

Os mesmos procedimentos foram repetidos para o APS-VC-*Subagging* e resultaram em um comportamento semelhante, como mostram as Figuras 5.8, 5.9 e 5.10.



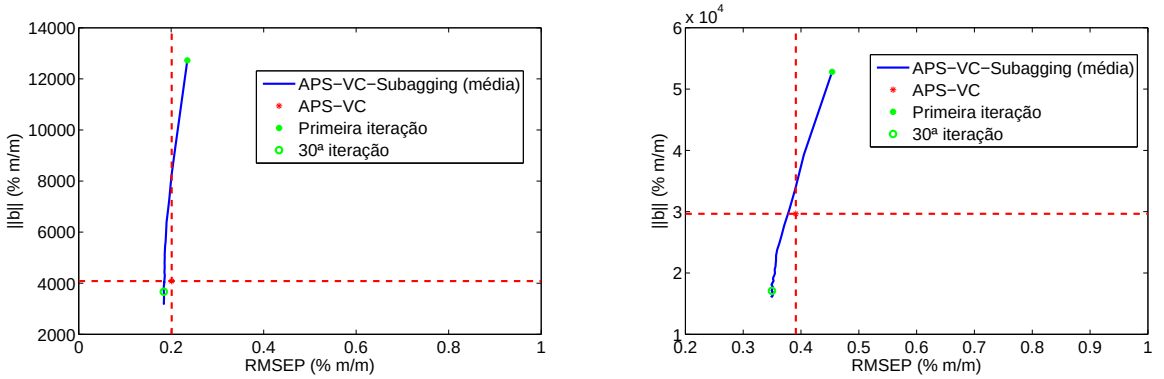
(a) Umidade (b) Proteína

FIGURA 5.8 – RMSEP obtido com 30 realizações do APS-VC-*Subagging*.



(a) Umidade (b) Proteína

FIGURA 5.9 – $||\mathbf{b}||$ obtida com 30 realizações do APS-VC-*Subagging*.



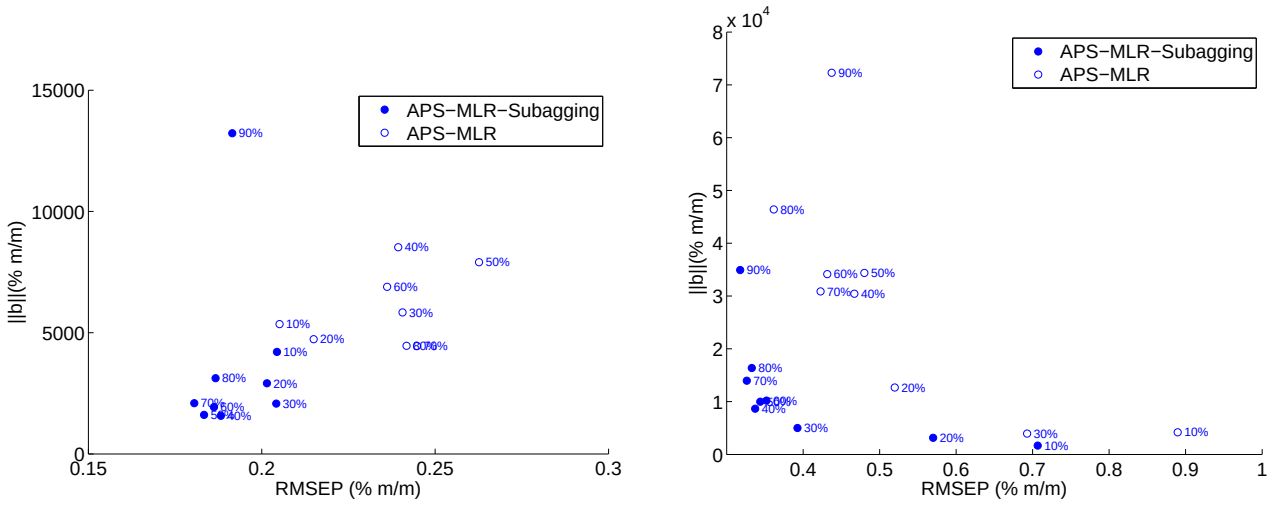
(a) Umidade (b) Proteína

FIGURA 5.10 – Avaliação multiobjetivo do APS-VC-*Subagging* (média de 30 realizações).

5.3 Variação da fração de subamostragem

Como discutido na Seção anterior, os resultados do APS-*Subagging* e APS-VC-*Subagging* não variam expressivamente para um número de iterações maior de que $P = 30$. Desse modo, os resultados apresentados adiante serão referentes a $P = 30$ iterações. Adicionalmente, como a dispersão observada entre as 30 realizações foi pequena, serão mostrados os resultados de uma única realização.

Avaliou-se o resultado do APS-*Subagging* ao variar o tamanho dos subconjuntos de calibração N_c, alterando-os de 10% a 90% dos 49 objetos de calibração. Para efeito de comparação, o conjunto de modelagem foi dividido nas mesmas proporções utilizando o algoritmo Kennard-Stone, para posterior aplicação do APS. Os resultados obtidos estão apresentados na Figura 5.11. O uso do *subagging* mais uma vez mostrou-se favorável em todas as frações de subamostragem para as duas propriedades analisadas, tanto para o RMSEP quanto para a $||\mathbf{b}||$.

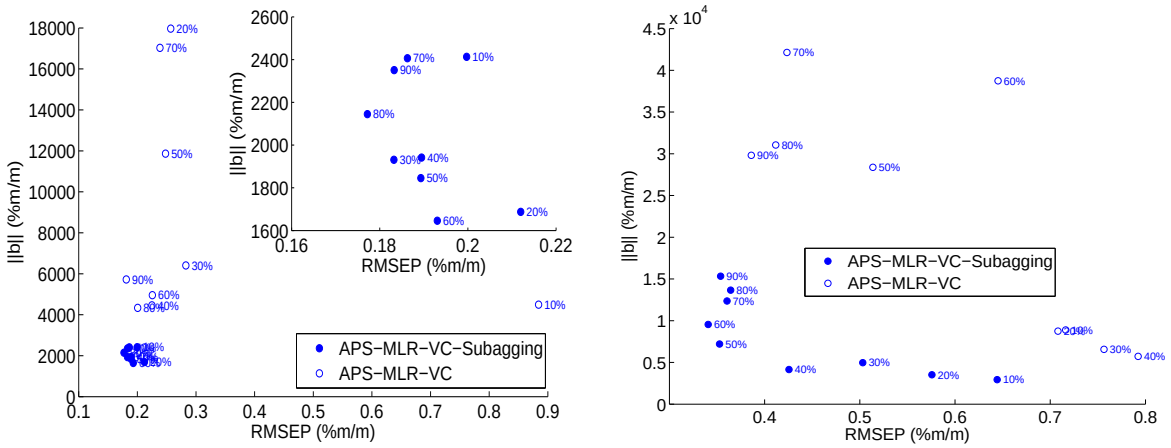


(a) Umidade (b) Proteína

FIGURA 5.11 – Avaliação multiobjetivo do APS-*Subagging* variando a fração de subamostragem.

Repetindo esse procedimento para o APS-VC-*Subagging* novamente é observada uma

melhora ao utilizar-se do *Subagging* em ambos os parâmetros para todas as frações.



(a) Umidade (b) Proteína

FIGURA 5.12 – Avaliação multiobjetivo do APS-VC-*Subagging* variando a fração de subamostragem.

De um modo geral, conclui-se que com a fração de 50% são obtidos bons resultados de RMSEP e $||\mathbf{b}||$ para ambas as propriedades, tanto para o APS quanto para o APS-VC.

5.4 Comparação geral

Conforme mostra a Tabela 5.2, o APS-*Subagging* apresenta resultados similares ao APS-VC-*Subagging* em termos de RMSEP, porém melhor em relação a sensibilidade a ruído ($||\mathbf{b}||$). Este comportamento é observado para as duas propriedades analisadas (umidade e proteína).

TABELA 5.2 – Valores de RMSEP e $||\mathbf{b}||$ obtidos com APS-*Subagging* e APS-VC-*Subagging* (%m/m) e a melhora (%) obtida em comparação com APS e APS-VC.

	Umidade				**Proteína**			
	RMSEP	Melhora	$\|\|\mathbf{b}\|\|$	Melhora	RMSEP	Melhora	$\|\|\mathbf{b}\|\|$	Melhora
APS-*Subagging*	0,18	25%	2254	49%	0,34	19%	13370	57%
APS-VC-*Subagging*	0,18	10%	3667	10%	0,35	10%	17089	42%

A melhora obtida com o *subagging* é graficamente apresentada nas Figuras 5.13 e

5.14 comparando das predições obtidas com (a) APS e (b) APS-*Subagging* para as duas propriedades em questão.

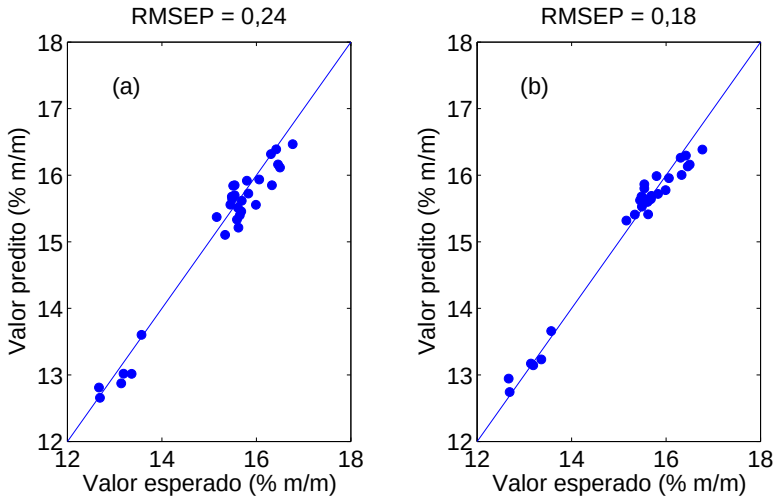


FIGURA 5.13 – Comparação entre o valor esperado e o predito com (a) APS e (b) APS-*Subagging* para umidade ($P = 30$ iterações).

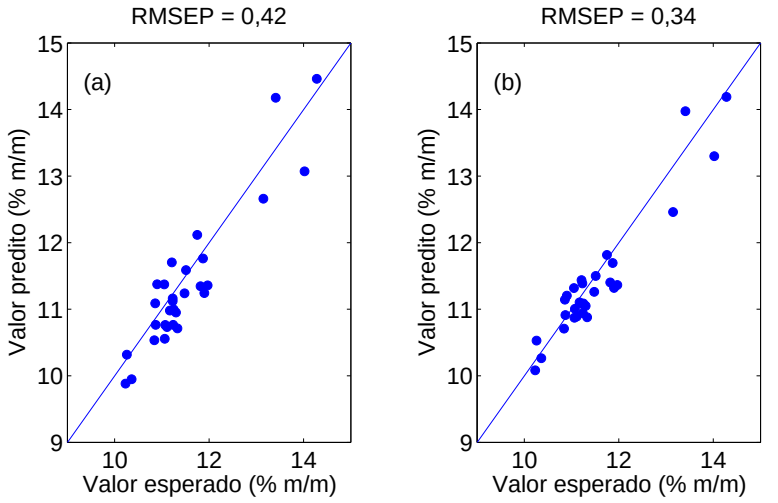


FIGURA 5.14 – Comparação entre o valor esperado e o predito com (a) APS e (b) APS-*Subagging* para proteína ($P = 30$ iterações).

Verifica-se que tanto para umidade quanto para proteína as predições obtidas com APS-*Subagging* encontram-se mais agrupadas em relação a bissetriz de que as obtidas com o APS. Vale salientar que não há existência de erro sistemático, pois os pontos do

gráficos se distribuem em ambos os lados da reta bissetriz.

As Figuras 5.15 e 5.16 mostram os coeficientes de regressão obtidos com APS e com APS-*Subagging* para (a) umidade e (b) proteína. Vale notar que com a subamostragem houve um aumento do número de variáveis incluídas no modelo, mas em contrapartida houve uma redução na magnitude dos coeficientes de regressão.

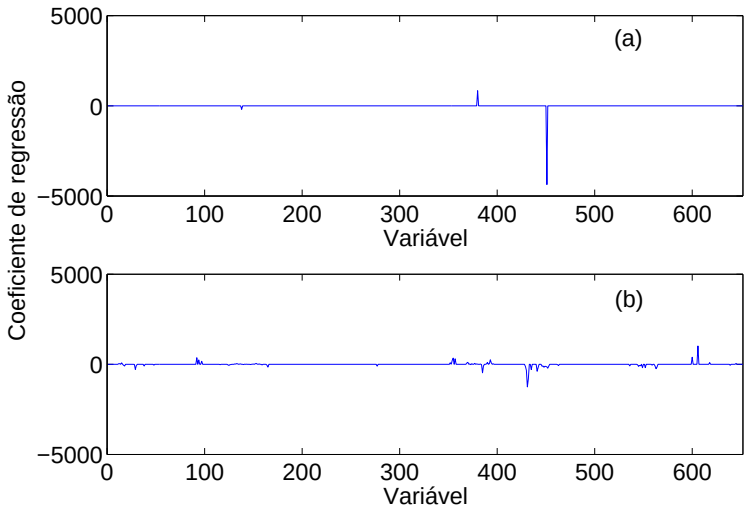


FIGURA 5.15 – Comparação dos coeficientes de regressão do modelo obtido com (a) APS e (b) APS-*Subagging* para umidade ($P = 30$ iterações).

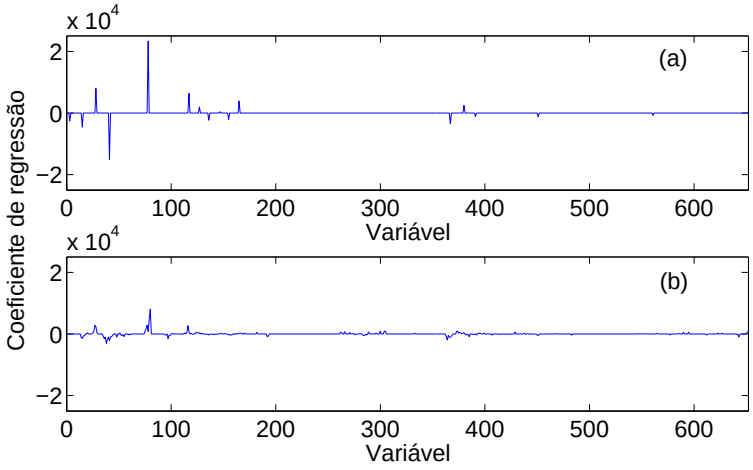


FIGURA 5.16 – Comparação dos coeficientes de regressão do modelo obtido com (a) APS e (b) APS-*Subagging* para proteína ($P = 30$ iterações).

O mesmo comportamento pode ser observado para as comparações dos modelos obtidos com APS-VC e APS-VC-*Subagging*. As predições são apresentadas nas Figuras 5.17 e 5.18, juntamente com os coeficientes de regressão nas Figuras 5.19 e 5.20. Novamente nota-se uma melhora do uso do *subagging*.

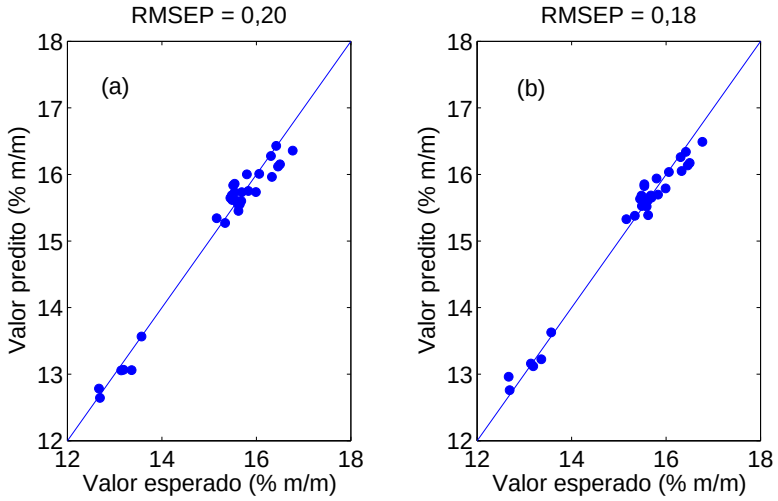


FIGURA 5.17 – Comparação entre o valor esperado e o predito com (a) APS-VC e (b) APS-VC-*Subagging* para umidade ($P = 30$ iterações).

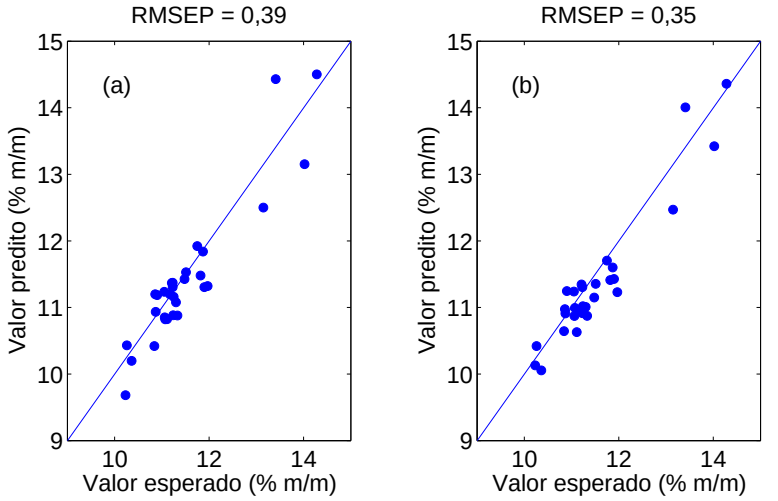


FIGURA 5.18 – Comparação entre o valor esperado e o predito com (a) APS-VC e (b) APS-VC-*Subagging* para proteína ($P = 30$ iterações).

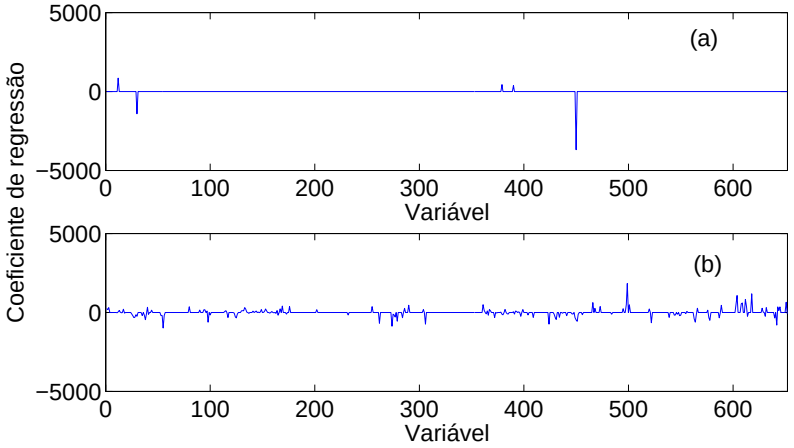


FIGURA 5.19 – Comparação dos coeficientes de regressão do modelo obtido com (a) APS-VC e (b) APS-VC-*Subagging* para umidade ($P = 30$ iterações).

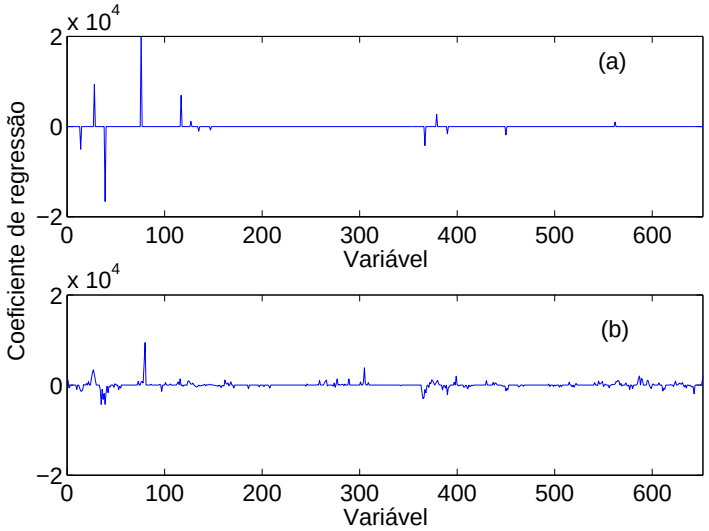
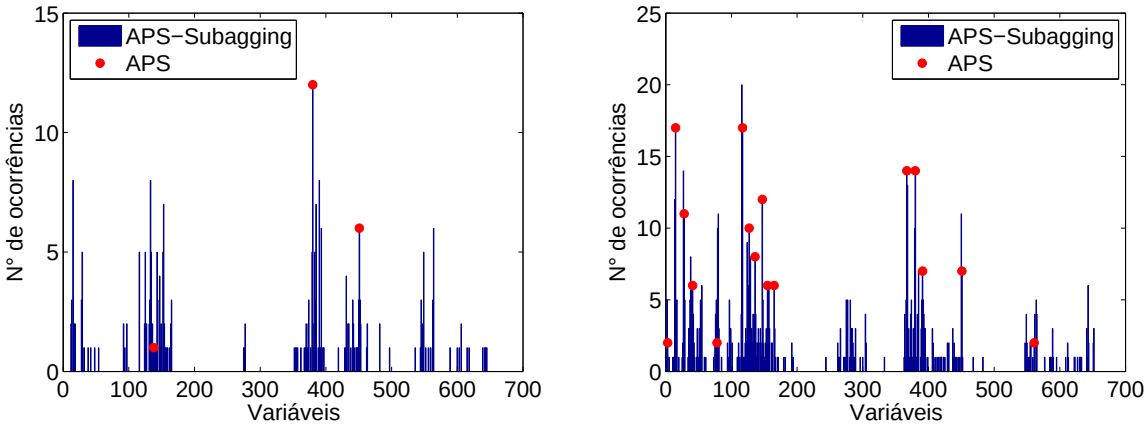


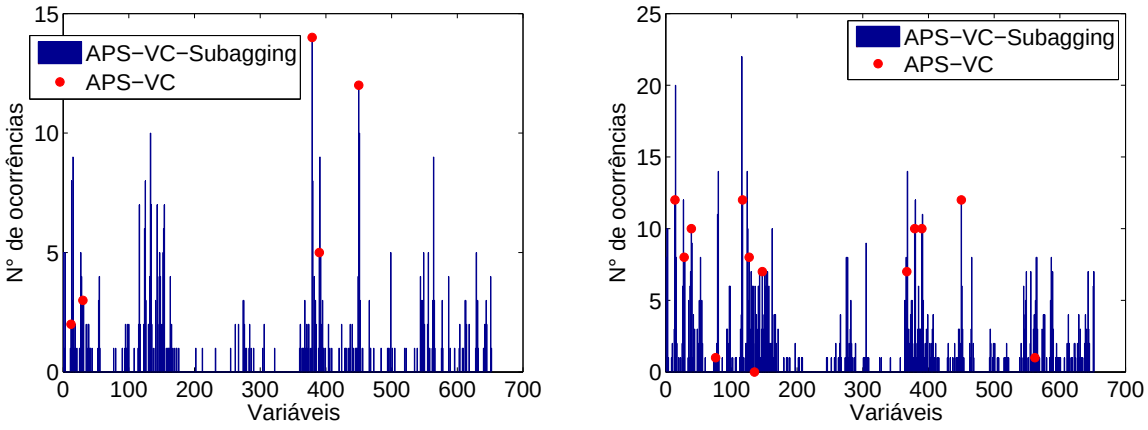
FIGURA 5.20 – Comparação dos coeficientes de regressão do modelo obtido com (a) APS-VC e (b) APS-VC-*Subagging* para proteína ($P = 30$ iterações).

As Figuras 5.21 e 5.22 mostram os histogramas contendo as variáveis selecionadas no decorrer de 50 iterações do APS e APS-VC, respectivamente. Em azul estão acumuladas as variáveis selecionadas no *subagging*, enquanto em vermelho são mostradas as sem o uso da subamostragem.



(a) Umidade (b) Proteína

FIGURA 5.21 – Histograma das variáveis selecionadas com APS-*Subagging* para 50 iterações.



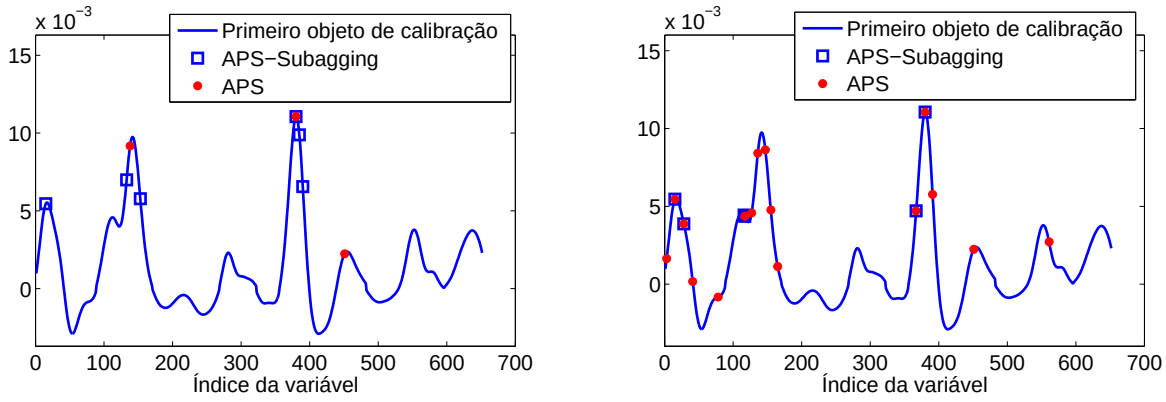
(a) Umidade (b) Proteína

FIGURA 5.22 – Histograma das variáveis selecionadas com APS-VC-*Subagging* para 50 iterações.

Observa-se que dentre as variáveis selecionadas pelo APS e APS-VC, a maior parte delas foram também selecionadas pelo *Subagging*, mostrando que são variáveis importantes para modelagem. Também é observado que o uso do *subagging* permite contemplar no modelo um número maior de variáveis, evitando que determinadas informações sejam

descartadas do processo.

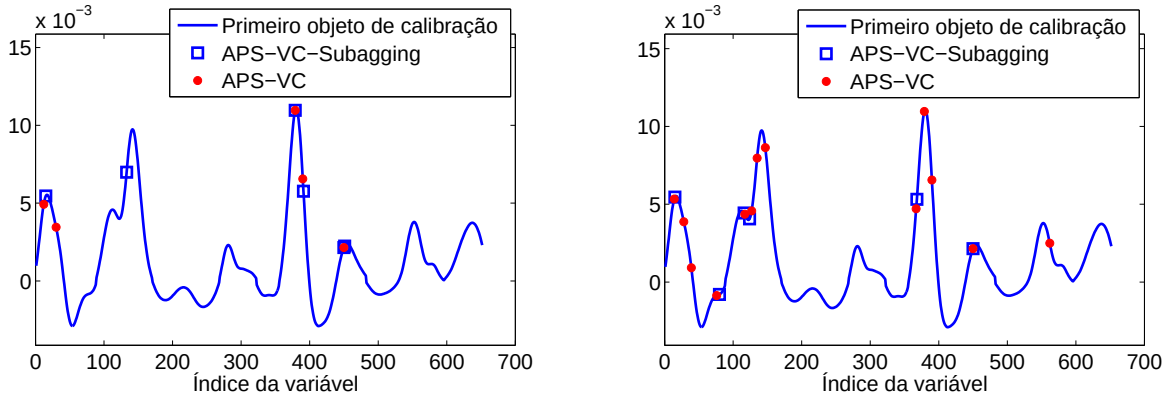
As Figuras 5.23 e 5.24 mostram o processo contrário, ou seja, as variáveis mais selecionadas com o uso do *subagging* comparadas com as variáveis selecionadas sem o seu uso. Os quadrados azuis representam os valores referentes às variáveis mais vezes selecionadas no *subagging* (maiores picos apresentados nos histogramas anteriores). Já as bolas vermelhas continuam representando as variáveis selecionadas pelos algoritmos sem o uso da subamostragem. Em vários casos as variáveis mais selecionadas no *subagging* fazem parte do conjunto selecionado sem o *subagging*.



(a) Umidade

(b) Proteína

FIGURA 5.23 – Variáveis mais selecionadas com APS-*Subagging* comparadas com as variáveis selecionadas pelo APS para 50 iterações.



(a) Umidade

(b) Proteína

FIGURA 5.24 – Variáveis mais selecionadas com APS-VC-*Subagging* comparadas com as variáveis selecionadas pelo APS-VC para 50 iterações.

6 Conclusão

Neste trabalho investigou-se o uso de uma técnica de reamostragem conhecida como *subagging* em conjunto com o APS como inicialmente proposto em (GALVãO *et al.*, 2006). Os resultados foram avaliados em um estudo de caso para determinação de umidade e proteína em amostras de trigo empregando espectrofotometria no infravermelho próximo. A seguir as contribuições deste trabalho são apresentadas, juntamente com as principais conclusões do estudo de caso e sugestões para pesquisas futuras.

6.1 Contribuições

- Melhor formalização da notação e descrição algorítmica do APS e do APS-VC com respeito a trabalhos anteriores (ARAúJO *et al.*, 2001; GALVãO *et al.*, 2007; GALVãO *et al.*, 2008).

- Avaliação da sensibilidade a ruído de medida (associada à norma do vetor de coeficiente de regressão), complementando a análise de capacidade de predição efetuada em (GALVãO *et al.*, 2006).

- Implementação do *subagging* em conjunto com o algoritmo APS-VC, contribuindo para o estudo comparativo de modelos APS-*Subagging* e APS-VC-*Subagging*.

- Avaliação do APS-*Subagging* e APS-VC-*Subagging* por meio de diversas realizações, possibilitando uma análise dos resultados em termos de média e desvio-padrão.

- Investigação do efeito de se variar a fração de subamostragem no APS-*Subagging* e APS-VC-*Subagging*.

- Implementação de um esquema baseado em regressão sequenciais para o APS, de modo a reduzir o tempo gasto na fase 2 de algoritmo. Tal esquema foi publicado em (SOARES *et al.*, 2010).

6.2 Conclusões do estudo de caso

- O APS-VC apresentou resultados ligeiramente melhores que o APS, porém demandou maior esforço computacional.

- O uso de *subagging* proporcionou melhorias com respeito aos modelos APS e APS-VC em termos das duas métricas consideradas (RMSEP e $||\mathbf{b}||$), mesmo tomando o limite superior dado pelo resultado médio mais um desvio-padrão. Constatou-se que os resultados não eram substancialmente alterados para um número de iterações de subamostragem maior do que $P = 30$, corroborando o resultado de (GALVÃO *et al.*, 2006).

- Os melhores resultados foram obtidos para frações de subamostragem em torno 50%. Contudo, os modelos obtidos mediante *subagging* revelaram-se superiores aos modelos APS e APS-VC tradicionais para qualquer fração de subamostragem aplicada ao conjunto de modelagem.

- O APS-*Subagging* mostrou-se similar ao APS-VC-*Subagging* em termos de RMSEP

para as duas propriedades (umidade e proteína) analisadas, porém revelou-se melhor em termos de sensibilidade a ruído (menor valor para a $||\mathbf{b}||$). Tendo em vista ainda a carga computacional demandada, pode-se concluir no estudo de caso realizado que o APS-*Subagging* seria a melhor alternativa.

6.3 Trabalhos futuros

- Realizar estudos de caso adicionais utilizando outros conjuntos de dados.
- Investigar o uso de *subagging* em conjunto com outros algoritmos de seleção de variáveis para RLM tais como algoritmo genético e *stepwise regression*, entre outros.
- Comparar os resultados obtidos através de *subagging* com os de outras técnicas de reamostragem e combinação de modelos, como por exemplo *bagging* e *boosting* (DRUCKER *et al.*, 2008).
- Propor extensões do APS-*Subagging* para problemas de classificação.

Referências Bibliográficas

ARAúJO, M. C. U.; SALDANHA, T. C. B.; GALVãO, R. K. H.; YONEYAMA, T.; CHAME, H. C.; VISANI, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 57, p. 65–73, 2001.

BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics: A Practical Guide**. New York, NY: John Wiley, 1998.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996.

BREIMAN, L. Technical note: Some properties of splitting criteria. **Machine Learning**, v. 24, p. 41–47, 1996.

BREITKREITZ, M. C.; RAIMUNDO, I. M.; ROHWEDDER, J. J. R.; PASQUINI, C.; FILHO, H. A. D.; JOSé, G. E.; ARAúJO, M. C. U. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. **The Analyst**, v. 128, p. 1204–1207, 2003.

BROADHURSTA, D.; GOODACREA, R.; JONESA, A.; ROWLANDB, J. J.; KELLA, D. B. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. **Analytica Chimica Acta**, v. 348, p. 71–86, 1997.

BÜHLMANN, P.; YU, B. Analyzing bagging. **Annals of Statistics**, v. 30, p. 927–961, 2002.

CHARNET, R.; FREIRE, C. A. de L.; CHARNET, E. M. R.; BONVINO, H. **Análise de modelos de regressão linear com aplicações**. Campinas, São Paulo: Unicamp, 2008.

COELHO, C. J.; GALVãO, R. K. H.; ARAUJO, M. C. U.; PIMENTEL, M. F.; SILVA, E. C. da. A linear semi-infinite programming strategy for constructing optimal wavelet transforms in multivariate calibration problems. **Journal of Chemical Information and Modeling**, v. 43, p. 928–933, 2003.

COELHO, C. J.; GALVãO, R. K. H.; ARAUJO, M. C. U.; PIMENTEL, M. F.; SILVA, E. C. da. A solution to the wavelet transform optimization problem in multicomponent analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 66, p. 205–217, 2003.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. New York, NY: Jonh Wiley & Sons, 1998.

DRUCKER, H.; CORTES, C.; JACKEL, L. D.; LECUN, Y.; VAPNIK, V. Boosting and other ensemble methods. **Neural Computation**, v. 6, n. 6, p. 1289–1301, 2008.

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. Quimiometria I: calibração multivariada, um tutorial. **Química Nova**, v. 22, n. 5, 1999.

FORINA, M.; LANTERI, S.; CASALE, M.; OLIVEIROS, M. C. C. Stepwise orthogonalization of predictors in classification and regression techniques: An "old"technique revisited. **Chemometries and Intelligent Laboratory Systems**, v. 87, p. 252–261, 2007.

GALVãO, R. K. H.; ARAúJO, M. C. U.; FRAGOSO, W. D.; SILVA, E. C.; JOSé, G. E.; SOARES, S. F. C.; PAIVA, H. M. A variable elimination method to improve the parsiminy of MLR models using the successive projections algorithm. **Chemometrics and Intelligent Laboratory Systems**, v. 83, p. 83–91, 2008.

GALVãO, R. K. H.; ARAúJO, M. C. U.; MARTINS, M. do N.; JOSé, G. E.; PONTES, M. J. C.; SILVA, E. C.; SALDANHA, T. C. B. An application of subagging for the improvement of prediction accuracy of multivariate calibration models. **Chemometrics and Intelligent Laboratory Systems**, n. 81, p. 60–67, 2006.

GALVãO, R. K. H.; ARAúJO, M. C. U.; SILVA, E. C.; JOSé, G. E.; SOARES, S. F. C.; PAIVA, H. M. Cross-validation for the selection of spectral variables using the successive projections algorithm. **Journal of The Brazilian Chemical Society**, v. 18, n. 8, p. 1580–1584, 2007.

GALVãO, R. K. H.; JOSé, G. E.; FILHO, H. A. D.; ARAUJO, M. C. U.; ; SILVA, E. C. da; PAIVA, H. M.; SALDANHA, T. C. B.; SOUZA, Ê. S. O. N. de. Optimal wavelet filter construction using X and Y data. **Chemometrics and Intelligent Laboratory Systems**, v. 70, p. 1–10, 2004.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Boston, MA: Addison-Wesley, 1989.

GUSNANTO, A.; PAWITAN, Y.; HUANG, J.; LANE, B. Variable selection in random calibration of near-infrared instruments: ridge regression and partial least squares regression settings. **Journal of Chemometrics**, v. 17, p. 174–185, 2003.

HAALAND, D. M.; THOMAS, E. V. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. **Analytical Chemistry**, v. 60, n. 11, 1988.

HOLLER, F. J.; SKOOG, D. A.; CROUCH, S. R. **Princípios de Análise Instrumental.** Porto Alegre, RS: Bookman, 1998.

HONORATO, F. A.; GALVãO, R. K. H.; PIMENTEL, M. F.; NETO, B. de B.; ARAúJO, M. C. U.; CARVALHO, F. R. de. Robust modeling for multivariate calibration transfer by the successive projections algorithm. **Chemometrics and Intelligent Laboratory Systems**, v. 76, p. 65–72, 2005.

KENNARD, R. W.; STONE, L. A. Computer-aided design of experiments. **Technometrics**, v. 11, n. 1, 1969.

KOKALY, R. F.; CLARK, R. N. Spectroscopic determination of leaf biochemistry using band-depth analysis of absorption features and stepwise multiple linear regression. **Elsevier**, New York, NY, v. 67, p. 267–287, 1999.

LAWSON, C. L.; HANSON, R. J. **Solving Least Squares Problems**. Englewood Cliffs, NJ: Prentice-Hall, 1974.

LEARDI, R. Genetic algorithms in chemometrics and chemistry: a review. **Journal of Chemometrics**, v. 15, p. 559–569, 2001.

LI, B.; WANG, D.; XU, C.; ZHANG, Z. Flow-injection simultaneous chemiluminescence determination of ascorbic acid and l-cysteine with partial least squares calibration. **Microchimica Acta**, v. 149, n. 3, p. 205–212, 2005.

LI, L. na; LI, Q. bo; ZHANG, G. jun. A weak signal extraction method for human blood glucose noninvasive measurement using near infrared spectroscopy. **Journal of Infrared, Millimeter and Terahertz Waves**, p. 1191–1204, 2009.

LIU, F.; HE, Y. Application of successive projections algorithm for variable selection to determine organic acids of plum vinegar. **Food Chemistry**, v. 115, p. 1430–1436, 2009.

LIU, F.; JIANG, Y.; HE, Y. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer. **Analytica Chimica Acta**, v. 635, p. 45–52, 2009.

NEZIO, M. S. D.; PISTONESI, M. F.; FRAGOSO, W. D.; PONTES, M. J. C.; GOICOECHEA, H. C.; ARAUJO, M. C. U.; BAND, B. S. F. Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water. **Microchemical Journal**, v. 85, p. 194–200, 2007.

NæS, T.; MEVIK, B. H. Understanding the collinearity problem in regression and discriminant analysis. **Journal of Chemometrics**, v. 15, p. 413–426, 2001.

PEREIRA, A. F. C.; PONTES, M. J. C.; NETO, F. F. G.; SANTOS, S. R. B.; GALVãO, R. K. H.; ARAúJO, M. C. U. NIR spectrometric determination of quality parameters in vegetable oils using ipls and variable selection. **Food Research International**, v. 41, p. 341–348, 2008.

PINO-MEJIAS, R.; VEGA, M. D. C. de la; COELLO, M. L.; RAMIREZ, E. L. S.; GAMERO, M. D. J. Bagging classification models with reduced bootstraps. **Lecture Notes in Computer Science**, n. 3138, p. 966–973, 2004.

PONTES, M. J. C.; GALVãO, R. K. H.; ARAúJO, M. C. U.; MOREIRA, P. N. T.; NETO, O. D. P.; JOSé, G. E.; SALDANHA, T. C. B. The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems**, v. 78, p. 11–18, 2005.

RIMBAUD, D. J.; MASSART, D. L.; LEARDI, R.; NOORD, O. E. D. Genetic algorithms as a tool for wavelength selection in multivariate calibration. **Analytical Chemistry**, v. 67, p. 4295–4301, 1995.

SOARES, A. S.; FILHO, A. R. G.; GALVãO, R. K. H.; ARAúJO, M. C. U. Improving the computational efficiency of the successive projections algorithmby using a sequential regression implementation: A case study involving NIR spectrometric analysis of wheat samples. **Journal of Brazilian Chemical Society**, v. 21, n. 4, p. 760–763, 2010.

SOUTO, U. T. C. P.; PONTES, M. J. C.; SILVA, E. C.; GALVãO, R. K. H.; ARAúJO, M. C. U.; SANCHES, F. A. C.; CUNHA, F. A. S.; OLIVEIRA, M. S. R. UV-VIS spectrometric classification of coffees by spa-lda. **Food Chemistry**, v. 119, p. 368–371, 2010.

WALCZAK, B.; FERRé, R. T.; BROWN, S. **Comprehensive Chemometrics: Chemical and Biochemical Data Analysis**. [S.l.]: Elsevier, 2009. 233-283 p.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo