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

Marcelo do Nascimento Martins

TRANSFERÊNCIA DE CALIBRAÇÃO DE INSTRUMENTOS PARA ANÁLISE ESPECTROMÉTRICA EMPREGANDO SELEÇÃO DE VARIÁVEIS, REAMOSTRAGEM E COMBINAÇÃO DE MODELOS

Tese aprovada em sua versão final pelos abaixo assinados:

Prof. Roberto Kawakami Harrop Galvão

Orientador

Prof. Homero Santiago Maciel

Pró-Reitor de Pós-Graduação e Pesquisa

Campo Montenegro

São José dos Campos, SP - Brasil

2006

Dados Internacionais de Catalogação-na-Publicação (CIP)

Divisão Biblioteca Central do ITA/CTA

Martins, Marcelo do Nascimento

Transferência de Calibração de Instrumentos Para Análise Espectrométrica Empregando Seleção de Variáveis, Reamostragem e Combinação de Modelos / Marcelo do Nascimento Martins. São José dos Campos, 2006.

112f.

Tese de Mestrado – Curso de Engenharia Eletrônica e Computação. Área de Sistemas e Controle – Instituto Tecnológico de Aeronáutica, 2006. Orientador: Prof. Roberto Kawakami Harrop Galvão. .

1. Qualidade de Combustíveis. 2. Transformada Wavelet. 3. Espectroscopia no Infravermelho. 4. Transferência de Calibração. 5. Gasolina. I. Centro Técnico Aeroespacial. Instituto Tecnológico de Aeronáutica. Divisão de Engenharia Eletrônica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

MARTINS, Marcelo do Nascimento. **Transferência de Calibração de Instrumentos Para Análise Espectrométrica Empregando Seleção de Variáveis, Reamostragem e Combinação de Modelos**. 2006. 112f. Tese de Mestrado – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

NOME DO AUTOR: Marcelo do Nascimento Martins

TITULO DO TRABALHO: Transferência de Calibração de Instrumentos Para Análise Espectrométrica Empregando Seleção de Variáveis, Reamostragem e Combinação de Modelos.

TIPO DO TRABALHO/ANO: Tese / 2006

Marcelo do Nascimento Martins

Rua Manoel Ricardo Júnior, n 72, ap 12

CEP 12.245-560 – São José dos Campos–SP

TRANSFERÊNCIA DE CALIBRAÇÃO DE INSTRUMENTOS PARA ANÁLISE ESPECTROMÉTRICA EMPREGANDO SELEÇÃO DE VARIÁVEIS, REAMOSTRAGEM E COMBINAÇÃO DE MODELOS

Marcelo do Nascimento Martins

Composição da Banca Examinadora:

Prof.	Karl Heinz Kienitz	Presidente	-	ITA
Prof.	Roberto Kawakami Harrop Galvão	Orientador	-	ITA
Profª.	Maria Fernanda Pimentel	Membro Externo	-	UFPE
Prof.	Cairo Lúcio Nascimento Júnior	Membro	-	ITA
Profª.	Mischel Carmen Neyra Belderrain	Membro	-	ITA

ITA

Aos meus pais, Enilda e Sergio,
e meu irmão, Lucas, pelo incentivo e apoio incessantes.

Agradecimentos

Ao CNPq pelo apoio financeiro a parte da pesquisa aqui relatada. Ainda na linha de apoio financeiro, não poderia deixar de agradecer à minha mãe e ao meu pai, meus "pais-trocinadores" de toda hora e ocasião.

Gostaria de agradecer ao meu orientador, Prof. Roberto Kawakami Harrop Galvão, por sua dedicação e pelo exemplo que tem sido para mim.

Ao grupo auto-intitulado de "Irmãos Resenha", do qual faço parte, nas figuras de Hugo Dias Filho, Rafael Breckenfeld e Ricardo Carrera. Sem o apoio e amizade deles, esse período não teria sido tão agradável.

À minha família joseense ("tio" Marco e Marlene, "tia" Sônia e Rubens e seus filhos) que prontamente me acolheu.

Aos meus colegas de curso, em especial aos meus conterrâneos.

Ao grupo ITALUS (http://www.comp.ita.br/italus) pela criação e manutenção da classe LaTeX usada neste trabalho. O trabalho deles, sem dúvidas, economizou-me muitas horas na preparação desta tese.

Os conjuntos de dados para avaliação das metodologias de transferência de calibração desenvolvidas foram obtidos através de colaboração com o Departamento de Química da Universidade Federal da Paraíba (UFPB) (Laboratório de Instrumentação e Automação em Química Analítica/Quimiometria - Prof. Mário César Ugulino de Araújo e Prof. Edvan Cirino da Silva) e o Departamento de Engenharia Química da Universidade Federal de Pernambuco (UFPE) (Laboratório de Combustíveis - Profª. Maria Fernanda Pimentel). Gostaria de agradecê-los pela colaboração e, em especial, à Drª Fernanda Araújo Honorato (UFPE) pela aquisição dos espectros de gasolina.

"É verdade que não podemos encontrar a pedra filosofal, mas é bom que ela seja procurada; procurando-a, descobrem-se muitos bons segredos que se não procuravam."

— Bernard Fontenelle

Resumo

A espectroscopia no infravermelho tem se mostrado uma ferramenta de valia para o monitoramento da qualidade de combustíveis. Contudo, tal técnica requer a calibração de modelos empíricos para relacionar medidas espectrais com parâmetros físico-químicos de interesse. Neste trabalho propõe-se um método que permite explorar um conjunto de dados já adquirido por um espectrômetro (instrumento Primário) na construção de um modelo para um segundo instrumento (Secundário). Tal método evita a duplicação de custo e esforço experimental no processo de calibração do modelo. Para isso, emprega-se o Algoritmo das Projeções Sucessivas para selecionar variáveis que sejam minimamente redundantes e portem informação relevante nos dois instrumentos. Adicionalmente, é empregado um método de reamostragem e combinação de modelos conhecido como *subagging*. Para validação do método proposto, apresenta-se um estudo de caso envolvendo a determinação de densidade e temperaturas para 10% e 90% de evaporados em amostras de gasolina, assim como um referente a determinação do teor de umidade em amostras de milho. É apresentado, também, um estudo comparativo com técnicas de transferência baseadas em padronização. Os resultados da técnica proposta se mostraram superiores aos obtidos através da técnica clássica de Mínimos-Quadrados Parciais empregando Padronização Direta por Partes. Em particular, verificou-se que o *subagging* propicia uma melhora expressiva na capacidade preditiva dos modelos obtidos por regressão linear múltipla. As técnicas propostas se dividem em abordagens no domínio original de comprimento de onda e no domínio da transformada wavelet. O desenvolvimento no domínio wavelet proporcionou uma redução no esforço computacional.

Palavras–Chave: qualidade de combustíveis; gasolina; espectroscopia no infravermelho; transferência de calibração; transformada wavelet.

Abstract

Infrared spectroscopy can be a useful tool for fuel quality monitoring. However, such a technique requires the calibration of empirical models to relate spectral measurements with physical and chemical parameters of interest. The present work proposes a method that allows the use of a data set previously acquired by a spectrometer (Primary instrument) in the construction of a model for a second instrument (Secondary). Such a method avoids the duplication of cost and experimental effort in the model calibration process. For this purpose, the Successive Projections Algorithm is employed to select variables that are minimally redundant and convey relevant information in both instruments. In addition, a method of re-sampling and model combination known as subagging is employed. The proposed method is validated in a case study concerning the determination of density and 10%, 90% distillation temperatures for gasoline samples and, also, in a case study concerning the determination of humidity in corn samples. A comparative study of calibration transfer techniques based on spectral response standardization is also presented. The results of the proposed technique were superior to those obtained by using the classic technique of Partial Least-Squares employing Piecewise Direct Standardization. In particular, the use of subagging led to a significant improvement in the predictive ability of the models obtained by multiple linear regression. The proposed techniques are divided into two approaches, the first one uses spectra in the original wavelength domain whereas the other uses spectra in the wavelet domain. The use of the wavelet domain provided a reduction in the computational effort.

Keywords: fuel quality; gasoline; infrared spectroscopy; calibration transfer; wavelet transform.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

%m/m	Porcentagem massa/massa, unidade de umidade
^{o}C	Graus Celsius, unidade de temperatura
ANP	Agência Nacional de Petróleo, Gás Natural e Biocombustíveis
APS	Algoritmo das Projeções Sucessivas
APSO	Algoritmo que usa o APS no domínio original de comprimento de onda para transferência de calibração
APSO-BAG	Algoritmo que faz uso do subagging em conjunto com o APSO
APSO/W	Ambas técnicas de transferência de calibração, APSO e APSW.
APSW	Algoritmo que usa o APS no domínio da transformada wavelet para transferência de calibração
APSW-BAG	Algoritmo que faz uso do subagging em conjunto com o APSW
Bagging	***B**ootstrap **agg**regat**ing*** (agregação de modelos por *bootstrap*)
CWT	*Continuous Wavelet Transform* (Transformada Wavelet Contínua)
DS	*Direct Standardization* (Padronização Direta)
DWT	*Discrete Wavelet Transform* (Transformada Wavelet Discreta)
FIR	*Finite Impulse Response* (Resposta a Impulso Finita)
kg/m^3	Kilograma por metro cúbico, unidade de massa específica
LV	*Latent Variable* (Variável Latente)

ME	Massa específica
MSC	*Multiplicative Signal Correction* (Correção Multiplicativa de Sinal)
MSEV	*Mean Square Error of Validation* (Média do Erro Quadrático de Validação) - média do PRESS
OSC	*Orthogonal Signal Correction* (Correção de Sinal Ortogonal)
PC	*Principal Component* (Componente Principal)
PCR	*Principal Component Regression* (Regressão em Componentes Principais)
PDS	*Piecewise Direct Standardization* (Padronização Direta por Partes)
PLS	*Partial Least-Squares* (Mínimos-Quadrados Parciais)
PRESS	*PRediction Error Sum of Squares* (Soma Quadrática dos Resíduos de Predição)
RLM	Regressão Linear Múltipla
RMSEP	*Root Mean Square Error of Prediction* (Raiz Quadrada do Erro Quadrático Médio de Predição)
$\mathrm{RMSEP}^{\mathrm{B}}_{\mathrm{A}}$	RMSEP obtido no instrumento B usando-se o modelo calibrado no instrumento A
$\mathrm{RMSEP}^{\mathrm{B}}_{\mathrm{A\text{-}T}}$	RMSEP obtido no instrumento B após transferência pelo método T usando-se o modelo calibrado no instrumento A
RMSET	RMSEP no conjunto de transferência
RMSEV	RMSEP no conjunto de validação
SBC	*Slope and Bias Correction* (Correção de Inclinação e Tendência)
SG	Algoritmo de Savitzky-Golay

Subagging	***Sub****sample* ***agg****regat****ing*** (agregação de modelos por subamostragem)
T10	Temperatura para 10% de gasolina evaporada
T90	Temperatura para 90% de gasolina evaporada
WHDS	*Wavelet Hybrid Direct Standardization* (Padronização Direta Híbrida Wavelet)
WT	*Wavelet Transform* (Transformada Wavelet)

Lista de Símbolos

$\|\mathbf{x}\|$	Norma Euclidiana de $\mathbf{x}$
α	Nível de significância para escolha do número de LV's em um modelo
A	Valor da absorbância de uma substância
b	Caminho ótico através do meio
$\mathbf{b}$	Matriz que contém os coeficientes do modelo
c	Concentração da espécie absorvente
$\mathbf{c}$(k)	Coeficientes wavelet de aproximação no k-ésimo nível de decomposição
$\mathbf{d}$(k)	Coeficientes wavelet de detalhe no k-ésimo nível de decomposição
$\mathbf{f}_i$	Vetor de transformação para o i-ésimo comprimento de onda
$\mathbf{F}$	Matriz de transformação dos métodos de padronização
$\mathbf{L}$	Vetor com o número das variáveis escolhidas pelo APS
m	Número de pontos do espectro medido
M	Número de amostras do conjunto de modelagem no *subagging*
Mc	Número de amostras do conjunto de calibração no *subagging*
Mv	Número de amostras do conjunto de validação no *subagging*
n	Número de amostras em análise
N	Número de variáveis a serem escolhidas pelo APS
nc	Número de amostras de calibração
$ndec$	Número de decomposições realizadas na DWT

nv	Número de amostras de validação
$Ntrans$	Número de amostras de transferência
p	Número de propriedades em análise
$\mathbf{p}$	Vetor de *loadings* do PLS
P	Potência da radiação de saída
Po	Potência da radiação incidente
$\mathbf{q}$	Vetor de regressão de $\mathbf{y}$ no PLS
$\mathbf{S}_1$	Espectros das amostras de transferência medidos no Primário
$\mathbf{S}_2$	Espectros das amostras de transferência medidos no Secundário
$\mathbf{S}^\dagger$	Matriz pseudo-inversa de Moore-Penrose da matriz $\mathbf{S}$
$\mathbf{t}$	Vetor de *scores* do PLS
$\mathbf{u}$	Resíduo não explicado pelo modelo
$\mathbf{v}$	Vetor com o resultado da DWT
$\mathbf{w}$	Vetor de pesos do PLS
$\mathbf{X}$	Matriz que contém os valores dos espectros medidos
$\mathbf{y}$	Matriz que contém os valores das propriedades em análise
$\hat{\mathbf{z}}$	Valor estimado de $\mathbf{z}$

1 Introdução

Em 1999, foi iniciado no Brasil o Programa de Monitoramento da Qualidade de Combustíveis, coordenado pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP), com o objetivo de fazer uma avaliação permanente da qualidade de combustíveis no país (DANTAS FILHO, 2003). A importância de tal monitoramento se deve tanto a aspectos de cunho econômico, quanto a questões relacionadas ao meio ambiente e à saúde pública. Com efeito, a utilização de combustíveis não-conformes às especificações de uso pode intensificar a emissão de material particulado, compostos aromáticos e óxidos de nitrogênio e enxofre, que têm sido ligados a uma maior incidência de doenças respiratórias e cardíacas, bem como a certos tipos de câncer (DAVIES; RUSZNAK; DEVALIA, 1998; DOCKERY *et al.*, 1989; POPE *et al.*, 2002).

Nesse contexto, a análise espectrométrica no Infravermelho tem se mostrado uma ferramenta de grande valia para a determinação rápida e não-destrutiva de propriedades de combustíveis (BOHACS; OVADI; SALGO, 1998; WANG *et al.*, 1999), podendo ser uma alternativa de baixo custo para os métodos atualmente em uso (DANTAS FILHO, 2003). Contudo, tal análise requer a construção de modelos matemáticos que relacionem propriedades físicas e/ou químicas da amostra com medidas espectrais registradas em diferentes comprimentos de onda (MARTENS; NAES, 1989; SKOOG; HOLLER; NIEMAN, 1998). O processo de obtenção de tais modelos a partir de amostras que tenham sido previamente caracte-

rizadas por métodos de referência é conhecido como Calibração Multivariada (FERREIRA *et al.*, 1999).

Uma vez que um modelo tenha sido desenvolvido para uma dada análise espectrométrica, seria desejável que o mesmo pudesse ser aplicado também em conjunto com instrumentos (espectrômetros) diferentes daquele utilizado na etapa de modelagem. Desse modo, poderia ser evitada a duplicação de custo e esforços envolvidos no processo de calibração multivariada. Tal necessidade pode surgir, por exemplo, em empresas que possuam vários espectrômetros (possivelmente de modelos e/ou fabricantes diferentes) distribuídos entre suas instalações (DESPAGNE *et al.*, 2000). Outra aplicação típica surge quando se deseja empregar um instrumento de laboratório de alta qualidade (em termos de resolução e relação sinal-ruído, por exemplo) para construir um modelo que posteriormente venha a ser empregado com instrumentos de campo de menor custo (WANG; LYSAGHT; KOWALSKI, 1992).

A área conhecida como "Transferência de Calibração" tem por objetivo o desenvolvimento de técnicas de tratamento de dados e modelagem para abordar problemas como os exemplificados acima (FEUDALE *et al.*, 2002; FEARN, 2001). Entre os métodos empregados para tanto podem-se distinguir duas grandes vertentes (SWIERENGA *et al.*, 1998b). A primeira busca a adaptação do modelo, ou dos dados espectrais, ao novo instrumento ("métodos de padronização") (BOUVERESSE; MASSART; DARDENNE, 1995; DREASSI *et al.*, 1998). A segunda, ("métodos sem padronização"), envolve o uso de técnicas que reduzam a influência das diferenças instrumentais e/ou ambientais presentes (BLANK; SUM; BROWN, 1996; GELADI; MACDOUGALL; MARTENS, 1985; WOLD *et al.*, 1998), ou o desenvolvimento de modelos robustos[1] a tais diferenças com base no emprego de características ("*features*")

[1] O termo robusto, sempre que usado neste texto, se refere a robustez em relação a variações instrumentais.

que apresentem pequenas variações entre os instrumentos considerados (MARK; JR., 1988; OZDEMIR; MOSLEY; WILLIAMS, 1998; SWIERENGA *et al.*, 1998a).

Seguindo a abordagem de construção de modelos robustos, recentemente foi proposto um esquema baseado na escolha de variáveis que sejam minimamente redundantes e portem informação relevante nos dois instrumentos em análise (HONORATO *et al.*, 2005). A escolha das variáveis, nesse método, é realizada no domínio de comprimento de onda e o modelo é construído usando-se regressão linear múltipla (RLM). Tal artigo foi usado como ponto de partida neste trabalho em que diversas alterações da técnica inicial foram propostas.

Nesse contexto, o instrumento de referência, para o qual o modelo foi inicialmente desenvolvido, é denominado "Primário". O instrumento para o qual se pretende transferir o modelo é denominado "Secundário". Um procedimento simples de padronização consiste em compensar diferenças de intensidade entre os espectros do Secundário e o do Primário multiplicando cada variável espectral do Secundário por um fator de correção (SHENK; WESTERHAUS; TEMPLETON, 1985). Tal fator é obtido por regressão linear com base em espectros registrados nos dois instrumentos para um mesmo subconjunto representativo de amostras (ditas "amostras de transferência"). Contudo, esse procedimento univariado (ou seja, baseado na relação entre variáveis espectrais individuais) não é eficaz em caso de alargamento de bandas (FEUDALE *et al.*, 2002), o que motivou o desenvolvimento de métodos multivariados.

Os métodos de padronização multivariada mais utilizados são os de Padronização Direta (*Direct Standardization*, DS) e Padronização Direta por Partes (*Piecewise Direct Standardization*, PDS) (WANG; VELTKAMP; KOWALSKI, 1991). Tais métodos têm por objetivo construir uma transformação linear que modifique o espectro registrado no ins-

trumento Secundário, de modo a torná-lo mais semelhante ao do Primário. No método DS, realiza-se uma transformação global, na qual cada variável espectral registrada no instrumento Secundário é corrigida com base nos valores de todas as demais variáveis. Já no PDS, a transformação é local, no sentido de que cada variável é corrigida com base nos valores de suas vizinhas mais próximas. O PDS, por ter se mostrado eficiente em uma grande variedade de aplicações, tem sido freqüentemente utilizado como técnica de referência para a avaliação de novas propostas para transferência de calibração (FEUDALE *et al.*, 2002).

Para fins de transferência de calibração, o uso da Transformada Wavelet têm se mostrado uma ferramenta de grande potencial (FEUDALE *et al.*, 2002; WALCZAK; BOUVERESSE; MASSART, 1997; PARK *et al.*, 2001; YOON; LEE; HAN, 2002), tanto para fins de pré-processamento (correção de variações na linha de base dos espectros, por exemplo) (TAN; BROWN, 2002), quanto para fins de padronização (WALCZAK; BOUVERESSE; MASSART, 1997; TAN; BROWN, 2001). O presente estudo apresenta uma abordagem um pouco distinta, pois explora a extração de características com o auxílio da transformada wavelet para a construção de modelos robustos. Como hipótese de trabalho, supõe-se que essas características sejam mais facilmente identificadas no plano tempo-escala do domínio wavelet do que no domínio original de comprimentos de onda.

De acordo com a Resolução ANP Nº 6, de 24.2.2005 - DOU 25.2.2005 há alguns parâmetros de qualidade que a gasolina automotiva deve atender de modo a poder ser comercializada. Os principais parâmetros descritos nas tabelas de especificações dos Regulamentos Técnicos ANP No. 02/2005 (anexo da Resolução supracitada) são divididos em análises regulares e análises complementares. Dentro das análises regulares tem-se aspecto, massa específica (ME), teor de AEAC (Álcool Etílico Anidro Combustível), des-

tilação (temperaturas (T) para 10%, 50%, 90% de evaporados, pontos final e inicial de destilação e resíduo), número de octano motor (*Motor Octane Number*, MON), número de octano pesquisa (*Research Octane Number*, RON) e teor de hidrocarbonetos olefínicos e aromáticos. As análises complementares são: pressão de vapor, goma atual lavada, período de indução, corrosividade ao cobre, teor de enxofre e de chumbo.

Os parâmetros que serão avaliados neste trabalho são descritos a seguir, juntamente com sua relevância para a qualidade da gasolina (ANP, 2005). A Tabela 1.1 mostra os níveis de variação aceitáveis para cada uma dessas propriedades, de acordo ainda com a mesma Resolução.

ME a 20 °C: Indica possíveis adulterações, com produtos mais leves ou mais pesados.

T10: Garante que a gasolina possua uma quantidade mínima de frações leves que se vaporizem e queimem com facilidade, na temperatura de partida a frio do motor, facilitando o início de funcionamento do veículo. Uma concentração muito alta pode dificultar a partida a quente e prejudicar a dirigibilidade do veículo devido à geração de bolhas na linha de combustível.

T90: A limitação dessa temperatura visa minimizar a formação de depósitos na câmara de combustão e nas velas de ignição, o que ocorre se a temperatura for muita elevada. A porção do combustível que não queima tende a vazar para o cárter do motor, "lavando" o cilindro e contaminando o óleo lubrificante. Hidrocarbonetos pesados proporcionam potência e contribuem para economia de combustível, mas a presença destes deve ser limitada, pois são difíceis de vaporizar e queimar.

No trabalho aqui apresentado, é proposto um método que permite explorar um conjunto de dados já adquirido por um espectrômetro (instrumento Primário) na construção

TABELA 1.1 – Especificações da gasolina padrão para ensaios de consumo e emissões.

Propriedade	Unidade	Limites
ME	kg/m^3	735,0 a 765,0
T10	°C	45,0 a 60,0
T90	°C	160,0 a 190,0

FONTE: Resolução ANP Nº6 de 24.02.2005

de um modelo para um segundo instrumento (Secundário). Tal método evita a duplicação de custo e esforço experimental no processo de calibração do modelo. Para isso, emprega-se o Algoritmo das Projeções Sucessivas (APS) para selecionar variáveis que sejam minimamente redundantes e portem informação relevante nos dois instrumentos, como proposto em (HONORATO *et al.*, 2005). Os modelos construídos no domínio original serão denominados de APSO, ao passo que os construídos no domínio wavelet são designados por APSW.

Adicionalmente, é empregado um método de reamostragem e combinação de modelos conhecido como *subagging*. O uso de tal técnica para calibração espectrométrica foi originalmente proposto por Galvão *et al.* (2006). Neste trabalho o *subagging* será usado para fins de transferência de calibração de modelos construídos tanto no domínio original de comprimento de onda (APSO-BAG) quanto no domínio da transformada wavelet (APSW-BAG). Espera-se com o uso do *subagging* poder suavizar o processo de seleção "dura"[2] de variáveis tal como é conduzido no APS. O *subagging* permite a atribuição pesos não-nulos a cada variável mantendo, adicionalmente, a simplicidade da RLM. Desse modo as técnicas de transferência apresentariam um comportamento mais sistemático, dado que não estariam sujeitas a variações bruscas causadas pela escolha de diferentes variáveis para cada novo conjunto de validação.

Neste trabalho, a ênfase será em aplicações de Espectrometria no Infravermelho para

[2]No sentido de uma variável estar presente ou não. Neste caso os possíveis pesos atribuídos a uma variável seriam zero ou um.

determinação de propriedades de gasolina automotiva. Vale ressaltar que, nesse campo, a transferência de calibração seria fundamental para a implementação de instrumentos simples e baratos que possam ser instalados nos próprios postos de combustível (DANTAS FILHO, 2003).

Para avaliação das metodologias de transferência de calibração desenvolvidas foram usados espectros de amostras de gasolina no Infravermelho Médio obtidos através de colaboração com o Departamento de Química da Universidade Federal da Paraíba (Laboratório de Instrumentação e Automação em Química Analítica/Quimiometria) e o Departamento de Engenharia Química da Universidade Federal de Pernambuco (Laboratório de Combustíveis).

Para fins de validação da técnica proposta, será usado ainda um conjunto de dados adicional constituído de espectros de amostras de milho que estão publicamente disponíveis em http://www.eigenvector.com/Data/Corn/.

2 Apresentação do Problema

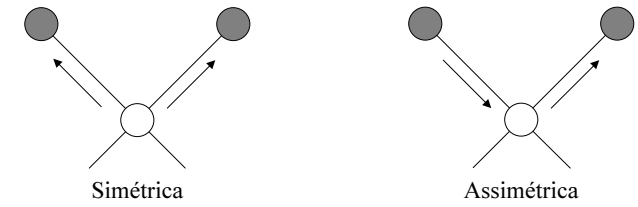
Neste capítulo será apresentado o problema de transferência de calibração. Para tanto será apresentada uma breve introdução sobre espectrometria e os fenômenos físicos envolvidos na análise espectroscópica. Em seguida será abordado o problema da calibração de modelos para uso na espectrometria. Em especial, será dada uma maior atenção a modelos de calibração multivariados. Logo após será tratada a questão de como se utilizar um modelo calibrado num determinado instrumento (Primário) em um outro instrumento (Secundário), processo esse denominado de transferência de calibração. Neste capítulo, serão dados maiores detalhes de técnicas de padronização dos espectros obtidos no instrumento Secundário, uma vez que estas servirão de comparação para a técnica proposta. Por fim, as técnicas apresentadas serão analisadas sob a ótica da análise no domínio wavelet.

2.1 Espectrometria

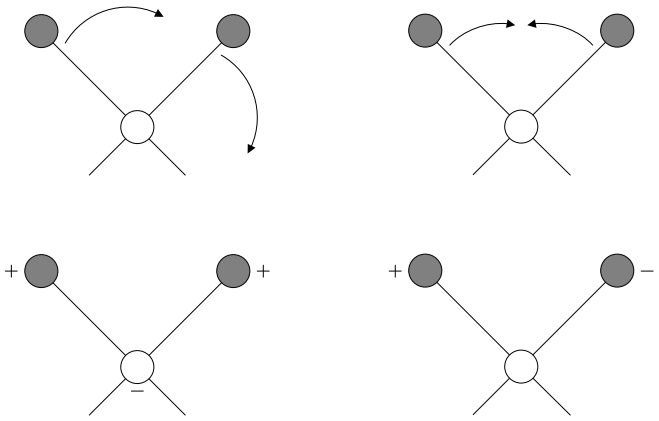
Espectroscopia é o nome dado à ciência que estuda as diversas interações entre radiação e matéria. Espectrometria, por sua vez, é o nome dado à área da Química que estuda métodos analíticos de análise baseados na espectroscopia atômica e molecular (SKOOG; HOLLER; NIEMAN, 1998). Neste trabalho são usadas técnicas de espectroscopia de absor-

ção no infravermelho [1].

Resumidamente, a teoria de espectroscopia de absorção no Infravermelho se baseia no fato de que as moléculas, a depender de suas estruturas, possuem modos vibracionais provenientes das interações entre os diferentes átomos que os compõem. Na Figura 2.1 estão ilustrados os modos de vibração de estiramento e de flexão de uma molécula com mais de três átomos (SKOOG; HOLLER; NIEMAN, 1998).



(a) Vibrações de estiramento



(b) Vibrações de flexão

FIGURA 2.1 – Tipos de vibrações moleculares. Nota: + indica um movimento saindo do plano da página em direção ao leitor; - indica um movimento saindo do plano da página se afastando do leitor. (SKOOG; HOLLER; NIEMAN, 1998)

Cada um desses modos vibracionais ocorre com uma determinada freqüência de oscilação. Se alguma radiação com freqüência igual à de um desses modos incidir nessa molécula poderá causar a ativação desse tipo de vibração com conseqüente absorção de energia[2]. Na espectroscopia no infravermelho faz-se a freqüência da radiação incidente

[1] O termo infravermelho se refere à faixa de freqüência em que as medidas são feitas. O infravermelho corresponde a freqüências entre 3×10^{11} a 3×10^{14} Hertz.

[2] Pode-se fazer uma analogia com um sistema mecânico massa-mola. Esse sistema possui uma freqüên-

variar numa determinada faixa de modo a se poder observar a taxa de absorção ao longo de toda a faixa de interesse, obtendo-se assim um espectro da amostra em análise. O espectro de uma amostra de gasolina pode ser visto na Figura 2.2.

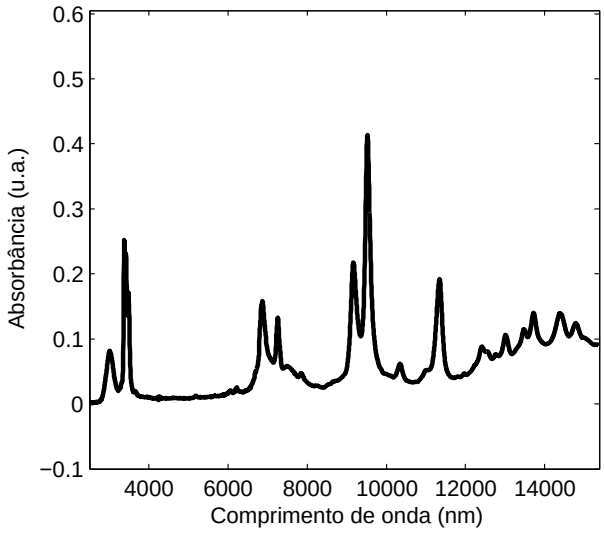


FIGURA 2.2 – Espectro de absorbância no infravermelho de uma amostra de gasolina.

Na espectroscopia de absorção há uma relação física conhecida como *lei de Beer* que diz que, para uma radiação monocromática (apenas uma freqüência[3]), a absorbância A é diretamente proporcional ao caminho ótico b através do meio e à concentração c da espécie absorvente (SKOOG; HOLLER; NIEMAN, 1998). A absorbância é definida através da seguinte equação matemática que relaciona a potência da radiação incidente Po com a potência da radiação de saída P

$$A = \log \frac{P_0}{P} \tag{2.1}$$

cia de oscilação natural que depende das relações entre a massa e a mola. Se alguma força externa excitar tal sistema com freqüência igual à de oscilação natural do sistema, a transferência de energia para o sistema será máxima.

[3]Os termos *freqüência* e *comprimento de onda* serão usados intercambialmente uma vez que estão relacionados, deterministicamente, por

$$c = f\nu$$

onde c é a velocidade da luz no vácuo e vale aproximadamente 3×10^8 m/s, f é a freqüência e ν é o comprimento de onda da radiação.

e a *lei de Beer* é expressa como

$$A = abc \tag{2.2}$$

em que a é uma constante de proporcionalidade denominada de absortividade. A Figura 2.3 ilustra os elementos definidos acima.

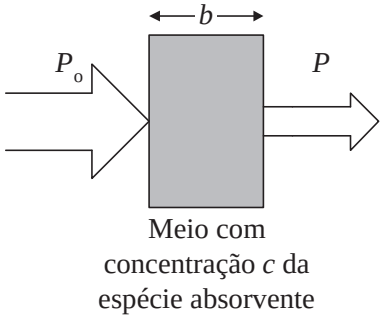


FIGURA 2.3 – Atenuação de uma radiação ao passar por um meio absorvente.

Pode-se generalizar a Equação 2.2 para $A(\lambda) = a(\lambda)bc$ e se construir um modelo matemático que relacione a concentração de uma espécie absorvente c a medidas da absorbância em diversos comprimentos de onda. Ainda mais, pode-se construir um modelo que relacione as concentrações de diversas espécies absorventes ao espectro de absorbância dessa mistura (BEEBE; PELL; SEASHOLTZ, 1998). A construção de tais modelos matemáticos será tratada na próxima seção.

Em se tratando da determinação de propriedades físicas, não há uma lei geral como a de Beer que justifique o uso de modelos lineares. Contudo, em diversas aplicações, modelos lineares têm sido empregados de forma heurística com resultados satisfatórios (THOMAS; HAALAND, 1990; BOHACS; OVADI; SALGO, 1998). Há, entretanto, também abordagens com modelos não-lineares (SEKULIC *et al.*, 1993).

2.2 Calibração Multivariada

O processo de calibração multivariada em espectrometria consiste, basicamente, em se construir um modelo matemático que relacione a resposta de um instrumento (no presente caso, o espectro de absorbância) a uma determinada propriedade da espécie em análise (BEEBE; PELL; SEASHOLTZ, 1998). O intuito é que o modelo construído seja usado para predição da propriedade de interesse em novas amostras de tal espécie. À luz da teoria da seção anterior, há indícios de que um modelo linear seja adequado para a devida representação da relação existente entre a resposta instrumental e a propriedade de interesse. Ou seja, em termos matemáticos

$$\mathbf{Y}_{[n\times p]} = \mathbf{X}_{[n\times m]} \cdot \mathbf{B}_{[m\times p]} + \mathbf{u}_{[n\times p]} \tag{2.3}$$

em que a matriz $\mathbf{Y}$ contém os valores das p propriedades de interesse das n amostras em análise, $\mathbf{X}$ contém os valores de absorbância medidos ao longo dos m comprimentos de onda, $\mathbf{b}$ é uma matriz que contém os coeficientes do modelo e $\mathbf{u}$ é o resíduo não explicado pelo modelo (seja por ruído de medida ou não-linearidades existentes). Ao longo deste trabalho, os modelos lineares utilizados serão construídos para apenas uma propriedade de interesse ($p = 1$), portanto as matrizes $\mathbf{Y}$ e $\mathbf{B}$ serão substituidas por suas notações vetoriais $\mathbf{y}$ e $\mathbf{b}$, respectivamente. Portanto, os modelos utilizados de fato são dados pela Equação abaixo.

$$\mathbf{y}_{[n\times 1]} = \mathbf{X}_{[n\times m]} \cdot \mathbf{b}_{[m\times 1]} + \mathbf{u}_{[n\times 1]} \tag{2.4}$$

2.2.1 RLM

Uma forma simples de se obter os coeficientes $\mathbf{b}$ para o modelo seria através de Regressão Linear Múltipla (RLM) por minimização do erro quadrático entre os valores reais e os preditos. Nesse caso, a estimativa $\hat{\mathbf{b}}$ pode ser obtida através do seguinte procedimento:

$$\begin{aligned}
\mathbf{E} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} \\
\mathbf{J} &= \mathbf{E}^{\mathbf{T}}\mathbf{E} = \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right)^{\mathbf{T}} \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \\
\frac{\partial \mathbf{J}}{\partial \hat{\mathbf{b}}} &= -2\mathbf{X}^{\mathbf{T}} \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) = 0 \\
\mathbf{X}^{\mathbf{T}}\mathbf{y} &= \mathbf{X}^{\mathbf{T}}\mathbf{X}\hat{\mathbf{b}} \\
\hat{\mathbf{b}} &= \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-\mathbf{1}} \mathbf{X}^{\mathbf{T}}\mathbf{y}
\end{aligned} \tag{2.5}$$

e uma nova predição é feita da seguinte forma

$$\hat{\mathbf{y}}_{\text{nova}} = \mathbf{X}_{\text{nova}}\hat{\mathbf{b}} \tag{2.6}$$

em que o símbolo $\hat{(\,)}$ representa um valor estimado.

Uma vantagem da RLM é que os modelos obtidos são simples e fáceis de se interpretar, uma vez que são constituídos de pesos para as diferentes variáveis espectrais (ver Equação 2.6). Porém, para poder se aplicar a RLM é necessário que a matriz de dados $\mathbf{X}$ tenha posto m de modo que $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ seja inversível (ver Equação 2.5). Entretanto, os dados utilizados nesse trabalho são de espectros com mais variáveis (comprimentos de onda) que amostras (ou seja, $m > n$) e, portanto, com posto menor que m. Logo, algum método de seleção de variáveis se torna imprescindível caso se queira usar a RLM.

O Algoritmo das Projeções Sucessivas (APS) (ARAÚJO *et al.*, 2001) (ver Seção 3.2) e

outros descritos na literatura como algoritmo genético, otimização combinatória (SASAKI; KAWATA; MINAMI, 1986), *generalized simulated annealing* (KALIVAS; ROBERTS; SUTTER, 1989) e redes neurais artificiais (DESPAGNE *et al.*, 2000) se propõem a selecionar variáveis que se mostrem mais apropriadas, em termos de capacidade preditiva, para a construção do modelo de RLM.

Uma característica atrativa do APS é que as variáveis selecionadas por esse método apresentam baixa multicolinearidade. Tal fato, como apontado em (NAES; MEVIK, 2001), diminui a propagação de ruídos instrumentais. Mais detalhes a respeito do APS serão apresentados na Seção 3.2.

2.2.2 PLS

Como alternativa ao uso da RLM, podem ser usados, como apontado em (FERREIRA *et al.*, 1999; BEEBE; PELL; SEASHOLTZ, 1998), a Regressão em Componentes Principais (*Principal Component Regression*, PCR) ou o método dos Mínimos-Quadrados Parciais (*Partial Least Squares*, PLS). O princípio básico desses métodos se baseia na construção de uma nova base ortogonal para o espaço gerado pelas amostras. O PCR cria essa base de modo que cada novo eixo coordenado, chamado de componente principal (PC), descreva a maior variação possível restante na matriz $\mathbf{X}$. Já no PLS, cada novo eixo coordenado, chamado de fator ou variável latente (LV), tenta descrever a maior variação possível de $\mathbf{X}$ correlacionada com $\mathbf{y}$ (FERREIRA *et al.*, 1999). As coordenadas das amostras nessa nova base são chamadas de *scores* e os cossenos dos ângulos dessa nova base com os antigos eixos são chamados de *loadings* (BEEBE; PELL; SEASHOLTZ, 1998).

O PLS tem sido amplamente utilizado em trabalhos relacionados a espectroscopia no

Infravermelho (WOLD; SJOSTROM; ERIKSSON, 2001), portanto uma maior atenção será dada a ele e o mesmo será adotado como método de calibração de referência para comparação com os demais métodos analisados nesse trabalho. Na construção do modelo PLS assume-se uma relação linear entre as grandezas envolvidas, como a dada pela Equação 2.4. O modelo PLS pode ser levantado para todos os parâmetros em conjunto ou para cada um em separado. Neste trabalho é adotada essa última abordagem, que tem sido adotada em um grande número de trabalhos recentes (BÜRCK *et al.*, 2006; GALVÃO *et al.*, 2006; LI *et al.*, 2006; VALVERDE *et al.*, 2006) apresentando bons resultados.

A primeira etapa na construção do modelo PLS consiste no pré-tratamento dos dados. De modo que cada novo eixo descreva a maior variação de um conjunto de dados é necessário que esses eixos tenham origem no centro de massa desses dados. Portanto é necessário se retirar a média de cada variável, de modo que a nova média seja nula. Tal procedimento é chamado de centrar os dados na média e é realizado tanto para os dados em $\mathbf{X}$ quanto em $\mathbf{y}$. Um exemplo didático da utilidade dessa etapa é mostrado na Figura 2.4. Pode-se notar que o ajuste dos eixos aos dados é melhor no caso em que os dados estão centrados na média, como representado na Figura 2.4a. Vale lembrar que, para fins de predição, tal média deve ser reincorporada aos dados ao fim do processo.

Na versão mais comum do PLS, a primeira variável latente $\mathbf{t_1} = \mathbf{X}\mathbf{w_1}$ é formada de modo a maximizar a covariância entre ela e o espaço formado por $\mathbf{y}$. Para isso, o vetor de pesos $\mathbf{w_1}$ é definido como o primeiro autovetor (i. e., associado ao maior autovalor) da matriz de covariância $\mathbf{X^T y y^T X}$, que no presente caso (modelo para único parâmetro) é dado por $\mathbf{w_1} = \mathbf{X^T y}/\|\mathbf{X^T y}\|$. Depois disso, as colunas de $\mathbf{X}$ e $\mathbf{y}$ são regredidas contra $\mathbf{t_1}$ para gerar os vetores de regressão $\mathbf{p_1} = \mathbf{X^T t_1}/(\mathbf{t_1^T t_1})$ e $\mathbf{q_1} = \mathbf{y^T t_1}/(\mathbf{t_1^T t_1})$. Em seguida, preparam-se $\mathbf{X}$ e $\mathbf{y}$ para a próxima iteração fazendo-se $\mathbf{X_2} = \mathbf{X} - \mathbf{t_1 p_1^T}$, $\mathbf{y_2} = \mathbf{y} - \mathbf{t_1 q_1^T}$. O

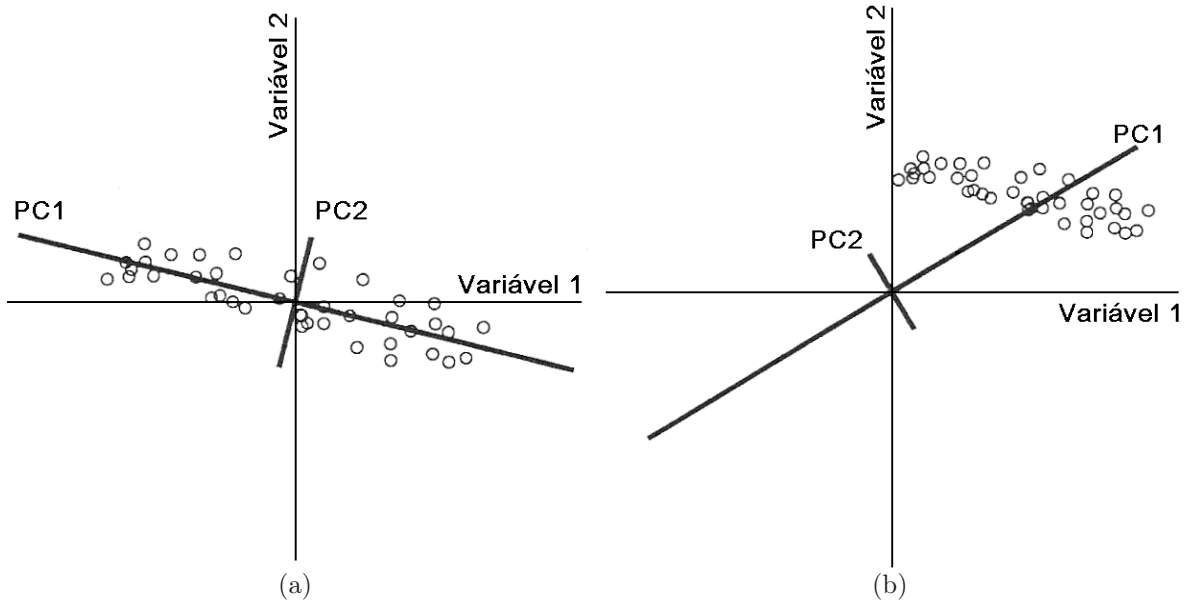


FIGURA 2.4 – Eixos com dados (a) centrados e (b) não centrados na média. (BEEBE; PELL; SEASHOLTZ, 1998)

processo é repetido até um número adequado de LV's. Ao término obtém-se o vetor $\mathbf{b}$ (ver Equação 2.4) através de $\mathbf{b} = \mathbf{W}(\mathbf{P^T W})^{-1}\mathbf{Q^T}$, onde as matrizes $\mathbf{W}$, $\mathbf{P}$ e $\mathbf{Q}$ são obtidas da concatenação dos vetores $\mathbf{w_i}$, $\mathbf{p_i}$ e $\mathbf{q_i}$, respectivamente (WOLD; SJOSTROM; ERIKSSON, 2001).

Uma dificuldade comum a ambos os métodos é determinar o número ideal de PC's (PCR), ou LV's (PLS), que devem ser acrescentados ao modelo. Não deve causar surpresa o fato de que, no conjunto de dados em que o modelo foi levantado, o resíduo $\mathbf{u}$ tende a diminuir [4] com o aumento do número de PC's, ou LV's. Corre-se, então, o risco de se criar um modelo que esteja "sobreajustado"[5] para descrever aquele conjunto de dados específico. Essa adversidade é, em parte, contornada separando-se o conjunto de amostras disponíveis para calibração do modelo em dois subconjuntos, um para a calibração e outro para o teste (ou validação) do modelo levantado no primeiro subconjunto. Porém, ainda assim, não é trivial se estabelecer um critério automático para a seleção do número de

[4] A rigor, seu valor pode também ficar constante, porém não aumenta com o acréscimo de PC's/LV's.
[5] Do inglês *overfit*.

fatores de um modelo. Esse será o assunto tratado na próxima subseção.

2.2.2.1 Critério de Seleção do Número de LV's

Existem diversos critérios que podem ser aplicados para a seleção do número de LV's, ou PC's, em um modelo. Como exemplos de alguns desses critérios pode-se citar o proposto por Wold (1978), que posteriormente recebeu o nome de critério R de Wold, em que o número ótimo de LV's do modelo é escolhido quando ocorre o primeiro mínimo na curva do PRESS (ver definição no próximo parágrafo) em função do número de LV's. Há uma variante desse critério, chamado critério R de Wold ajustado, em que a próxima LV só será acrescentada se a diminuição no PRESS em relação ao modelo sem essa nova LV for significativa. Em geral essa significância varia entre 1% e 10% e é definida pelo usuário. Outro método é o proposto por Krzanowski (1987), em que se considera a diminuição percentual do PRESS em função do número de LV's, porém ponderado pelo número de amostras, de variáveis e da quantidade de LV's já acrescentados.

O critério aqui adotado é o sugerido por Haaland e Thomas (1988), que tem sido adotado em vários trabalhos recentes (LI *et al.*, 2005; MOROS *et al.*, 2005; ZAREI; ATABATI; MALEKSHABANI, 2006). Para se aplicar tal critério é necessário calcular a soma quadrática dos resíduos da predição (*PRediction Error Sum of Squares*, PRESS) no conjunto de validação (PRESS_v), como definido na Equação 2.7:

$$\text{PRESS}_\text{v}(h) = \sum_{k=1}^{Nval} (y_k - \hat{y}_k(h))^2 \tag{2.7}$$

em que y_k e $\hat{y}_k(h)$ são os valores de referência e o predito usando h fatores da k-ésima amostra de validação para o parâmetro em questão. O valor de $\text{PRESS}_\text{v}(h)$ deve ser

calculado para h variando de 1 até um máximo de $Ncal$.

O modelo que leva a um menor PRESS é usado como referência e o número de fatores associados a esse modelo é denotado por h^*. Todos os modelos com menos fatores ($h < h^*$) são comparados com essa referência. O propósito desse critério é achar o modelo com menor número de fatores tal que o PRESS não seja significamente maior que o PRESS do modelo de referência com h^* fatores. O teste de significância é feito com base na estatística F de Snedecor (PAPOULIS, 1991).

Os passos para a aplicação do critério são os seguintes:

Passo 1. Calcular

$$F(h) = \frac{\text{PRESS (modelo com } h \text{ fatores)}}{\text{PRESS (modelo com } h^* \text{ fatores)} \qquad \text{para } h = 1, 2, \ldots, h^*$$

Passo 2. Escolher o número ótimo de fatores como sendo o menor h tal que $F(h) < F_{\alpha;Ncal,Ncal}$ onde $F_{\alpha;Ncal,Ncal}$ é o (1-α) percentil da distribuição F de Snedecor com $Ncal$ e $Ncal$ graus de liberdade.

Assumindo que os erros de predição têm média zero, são mutuamente independentes e normalmente distribuídos, $\text{Prob}\{F(h) > F_{\alpha;Ncal,Ncal} |\ \sigma_h^2 = \sigma_{h^*}^2\} = 2\alpha$. Aqui σ_h^2 e $\sigma_{h^*}^2$ representam a variância do erro de predição do modelo com h e h^* fatores, respectivamente, e α é um valor de probabilidade a ser escolhido. Vale notar que a probabilidade é 2α e não α. Isso porque o denominador da estatística F foi escolhido como sendo PRESS(h^*) no lugar de se escolher aleatoriamente entre PRESS(h^*) e PRESS(h). Portanto, a estatística F calculada é sempre maior ou igual a 1. Vale ressaltar também que, em aplicações práticas, $\text{Prob}\{F(h) > F_{\alpha;Ncal,Ncal}\} < 2\alpha$ por causa da correlação positiva entre os erros de

predição entre os modelos com h e h^* fatores. O efeito potencial dessa correlação positiva é o de se escolher um modelo com poucos fatores, ou seja, um modelo "subajustado"[6].

Em geral, o número de fatores em um modelo aumenta com α. Se α for muito pequeno, há chances de se ter problemas de subajuste, enquanto que se for muito grande resultará em sobreajuste do modelo. Em (HAALAND; THOMAS, 1988) os autores apontam que um valor de $\alpha = 0.25$, embora arbitrário, é um bom compromisso em aplicações práticas.

2.3 Transferência de Calibração Multivariada

Uma vez construído um modelo para uma dada análise espectrométrica deve-se tomar certos cuidados ao aplicá-lo para predição em novas amostras, pois há situações que podem tornar o modelo inadequado para uso. São três as principais situações (WANG; VELTKAMP; KOWALSKI, 1991; FEUDALE *et al.*, 2002):

1. mudanças na constituição física e/ou química de amostras de diferentes lotes. Essas mudanças podem resultar de diferenças na viscosidade, tamanho das partículas, textura superficial, entre outras, que podem ocorrer entre diferentes lotes de produção;

2. mudanças nas condições ambientais de operação do instrumento. Exemplos são flutuações na temperatura e variação na umidade; e

3. mudanças na resposta instrumental. Essas mudanças podem ocorrer se as amostras forem coletadas em um instrumento diferente daquele usado para a calibração do modelo, ou se alguma parte do aparelho foi trocada, ou em razão do próprio envelhecimento de seus componentes ópticos, eletrônicos e mecânicos.

[6] Do inglês *underfit*.

Os métodos que tentam contornar os problemas acima são denominados de métodos de transferência de calibração. Embora o termo transferência de calibração se refira, de forma geral, à transferência de um modelo entre diferentes condições, neste trabalho será assumido que se trata da transferência entre dois instrumentos distintos.

Após a calibração de um modelo em um dado instrumento, chamado de "Primário", é natural o desejo de que o mesmo possa ser aplicado a instrumentos (espectrômetros) diversos. O instrumento para o qual se deseja transferir a calibração é denominado de "Secundário".

Como mencionado na Introdução há métodos de transferência com e sem padronização. Ambas as abordagens possuem seus prós e contras. Métodos baseados em padronização têm seus méritos por poder corrigir, através de manipulações matemáticas, diversas diferenças entre os instrumentos Primário e Secundário como, por exemplo, intensidade da resposta, deslocamento espectral e alargamento de picos (FEUDALE *et al.*, 2002). A principal desvantagem dos métodos de padronização é que é necessário que um conjunto de amostras seja medido em ambos os instrumentos (essas amostras são ditas de transferência) de modo que se possa levantar um modelo adequado de correção das diferenças instrumentais.

Porém, muitas vezes é inviável medir a resposta da mesma amostra em ambos os instrumentos, seja por distância física, ou por que a amostra é quimicamente instável e não pode ser transportada a longas distâncias, ou por que é uma substância perigosa e envolveria custos de transporte elevados. Neste argumento reside a principal vantagem dos métodos baseados no desenvolvimento de modelos robustos a variações instrumentais. Em tais técnicas não há a obrigatoriedade do uso de amostras de transferência, pois as mesmas são, em sua maioria, baseadas em técnicas de pré-processamento dos dados de

modo a se extrair características que apresentem pequenas variações nos instrumentos em questão. Em contrapartida, tais métodos possuem reduzida capacidade de correção de variações instrumentais, sendo restritos, basicamente, a correção na intensidade da resposta (FEUDALE *et al.*, 2002).

Os dados disponíveis para a realização deste trabalho tornam viável o uso de técnicas de padronização. Tais técnicas serão, portanto, utilizadas para comparação com a proposta apresentada de construção de modelos robustos.

2.3.1 Métodos de Padronização

O objetivo dos métodos de padronização é fazer com que a resposta medida no instrumento Secundário emule a resposta que seria obtida se a medição fosse realizada no instrumento Primário. Técnicas de padronização se dividem, basicamente, em três classes. Na primeira, busca-se padronizar os coeficientes **b** do modelo (ver Equação 2.4) obtido no Primário, fazendo com que o modelo se torne adequado para uso no novo instrumento (Secundário). Tal abordagem foi proposta por Wang *et al.* (1991), porém mostrou-se mais tarde que tal proposta equivale a uma recalibração do modelo no instrumento Secundário (FEUDALE *et al.*, 2002).

Outra opção é se padronizar os valores preditos. Um método bastante utilizado para correção dos valores preditos é a correção de inclinação e tendência (*Slope and Bias Correction*, SBC). Nessa técnica se assume uma relação linear entre as predições realizadas no instrumento Secundário e as que teriam sido realizadas no Primário. Resumidamente, as amostras de transferência são medidas em ambos os instrumentos e uma reta afim (com coeficientes angular e de interseção) é ajustada de modo a relacionar os valores medidos

no Secundário aos valores medidos no Primário. Novos valores obtidos no Secundário são corrigidos usando-se a equação dessa reta (BOUVERESSE *et al.*, 1996).

Por fim, há os métodos que buscam padronizar a resposta espectral obtida nos dois instrumentos. A idéia essencial desses métodos consiste em se modelar as diferenças instrumentais através de uma regressão linear das respostas espectrais (das amostras de transferência) do instrumento Secundário no instrumento Primário. Após a devida correção das respostas espectrais pode-se aplicar o modelo calibrado no Primário diretamente no Secundário. Há tanto abordagens univariadas (SHENK; WESTERHAUS; TEMPLETON, 1985) quanto multivariadas, sendo que as últimas têm demonstrado melhores resultados que as primeiras (FEUDALE *et al.*, 2002).

Neste trabalho maior atenção é dada a técnicas pertencentes à classe de padronização da resposta instrumental. Dentre essas técnicas, duas têm recebido grande destaque na literatura: a Padronização Direta (*Direct Standardization*, DS) e a Padronização Direta por Partes (*Piecewise Direct Standardization*, PDS). Ambas foram propostas por Wang *et al.* (1991) e serão usadas como técnicas de referência para a técnica proposta neste trabalho.

Tanto o DS como o PDS são métodos de transferência de calibração multivariada que relacionam a resposta espectral do Secundário com a do Primário através de uma matriz de transformação $\mathbf{F}$. Tal transformação, em teoria, captura as variações instrumentais existentes e faz com que uma nova amostra se comporte como se tivesse sido medida no instrumento Primário tornando-se, portanto, apropriada para fins de predição no modelo originalmente levantado (WANG; VELTKAMP; KOWALSKI, 1991). Para se estimar a matriz $\mathbf{F}$ é necessário que um conjunto de amostras seja medido em ambos os instrumentos obtendo-se espectros $\mathbf{S}_1$ no Primário e $\mathbf{S}_2$ no Secundário. Essas amostras são ditas de

transferência e sua quantidade será denotada por $Ntrans$.

2.3.1.1 Padronização Direta

A Padronização Direta (DS) é um método de transferência multivariada que utiliza informação de todo o espectro das amostras de transferência para realizar a padronização. É, portanto, uma técnica de caráter global e com boa rejeição a ruído (TAN; BROWN, 2001). A estimação de $\mathbf{F}$ decorre de

$$\mathbf{S}_{1_{[Ntrans\times m]}} = \mathbf{S}_{2_{[Ntrans\times m]}}\mathbf{F}_{[m\times m]} \tag{2.8}$$

$$\mathbf{F} = \mathbf{S}_2^{\dagger}\mathbf{S}_1 \tag{2.9}$$

em que $\mathbf{S}_2^{\dagger}$ é a matriz pseudo-inversa de Moore-Penrose da matriz $\mathbf{S}_2$. Uma vez que $\mathbf{F}$ tenha sido calculada, a medida de uma nova amostra emulada a resposta do Primário a partir da seguinte transformação

$$\hat{\mathbf{x}}_{1,\text{nova}_{[1\times m]}} = \mathbf{x}_{2,\text{nova}_{[1\times m]}}\mathbf{F}_{[m\times m]} \tag{2.10}$$

Como se nota na Equação 2.9, a transformação é feita assumindo-se que a relação entre as diferenças instrumentais é linear, entretanto não-linearidades pequenas podem ser toleradas pelo método. Outro ponto relevante é que cada comprimento de onda do espectro do Primário se relaciona com todos os comprimentos de onda do espectro do Secundário (ver Equação 2.9).

2.3.1.2 Padronização Direta por Partes

A motivação para a aplicação do PDS advém do fato que, como apontado em (WANG; VELTKAMP; KOWALSKI, 1991), variações espectrais geralmente são limitadas a uma pequena região do espectro. Portanto, cada comprimento de onda do instrumento Primário seria melhor relacionado a uma região, chamada janela, em volta do mesmo comprimento de onda no Secundário. Essa abordagem por partes se propõe a corrigir eventuais deslocamentos espectral e ainda diferenças locais nos espectros como, por exemplo, alargamento de picos.

Dado um comprimento de onda i dos espectros das amostras de transferência medidos no Primário $\mathbf{r}_{1,i}$, o método PDS busca estabelecer uma relação linear com uma região em torno desse comprimento de onda no espectro do Secundário denotada por $\mathbf{R}_i = [\mathbf{r}_{2,i-j},\ \mathbf{r}_{2,i-j+1},\ \ldots,\ \mathbf{r}_{2,i+k-1},\ \mathbf{r}_{2,i+k}]$. Então

$$\mathbf{r}_{1,i_{[Ntrans\times 1]}} = \mathbf{R}_{i_{[Ntrans\times(j+k+1)]}}\mathbf{f}_{i_{[(j+k+1)\times 1]}} \tag{2.11}$$

em que i é o i-ésimo comprimento de onda, j é o limite à esquerda da janela, k é o limite à direita e $\mathbf{f}_i$ é o vetor de transformação para o i-ésimo comprimento de onda. O tamanho da janela usada é dado por $w = j + k + 1$. Quando o índice i está perto de uma das extremidades pode acontecer de a janela ultrapassar o limite do espectro. Neste caso pode-se estender o espectro com zeros, o que equivale a reduzir um dos lados da janela. Outra alternativa é simplesmente descartar esses comprimentos de onda. A regressão apontada na Equação 2.11 (note a semelhança com a Equação 2.4) pode ser conduzida por vários métodos de calibração multivariada, porém, em geral, se usa ou PCR, ou PLS (WANG; VELTKAMP; KOWALSKI, 1991). Neste trabalho usou-se o PCR com o critério de

se adicionar PC's até que 99% da variação de $\mathbf{R}_i$ tenha sido explicada. Vale ressaltar que o critério adotado para a escolha do número de PC's, nesse caso, não é tão crítico quanto no caso da calibração do modelo (ver Seção 2.2.2), dado que o posto da matriz $\mathbf{R}_i$ em geral é de ordem reduzida (da ordem de unidades) implicando num reduzido número de PC's (em geral, ≤ 3)[7].

Podem-se organizar os vetores $\mathbf{f}_i$ de modo a se construir uma matriz de transformação $\mathbf{F}$ como no caso do DS. Para o PDS tal matriz é da forma

$$\mathbf{F} = \text{diag}(\mathbf{f}_1,\ \mathbf{f}_2,\ \ldots,\ \mathbf{f}_m) \tag{2.12}$$

em que m é o número total de comprimentos de onda. Com a matriz $\mathbf{F}$ dessa forma, pode-se prosseguir como na Equação 2.10 para que o espectro de uma nova amostra medida no Secundário seja, aproximadamente, o que teria sido obtido se a medida fosse realizada no Primário.

O algoritmo PDS foi posteriormente aperfeiçoado para poder corrigir grandes diferenças aditivas entre as respostas instrumentais (WANG; DEAN; KOWALSKI, 1995). Tal técnica consiste em centrar na média as colunas de cada espectro e recebeu o nome de *Additive Background Correction* [8].

O PDS é um dos métodos de transferência de calibração mais usados e é, em geral, usado como referência para novas propostas (FEUDALE *et al.*, 2002). Estudos realizados por Wang *et al.* (1992) mostram que o uso do PDS pode levar a resultados melhores do que a recalibração do modelo no instrumento Secundário caso o Primário seja de qualidade superior.

[7] Tal fato foi constatado durante a análise dos resultados obtidos com os dados disponíveis.
[8] Uma possível tradução seria "correção de tendências aditivas".

2.3.1.3 Técnicas de Padronização no Domínio Wavelet

Observando-se as características do DS e do PDS, nota-se que o primeiro possui um aspecto global, enquanto o PDS apresenta uma característica local e ambos possuem suas vantagens. Tal dualidade leva ao questionamento natural de se uma abordagem multirresolucional, envolvendo aspectos tanto globais quanto locais, não levaria a resultados melhores que os obtidos pelo DS e PDS. Tal abordagem multirresolucional pode ser obtida através da Transformada Wavelet (TW) (ver Seção 3.1).

Entre as técnicas de padronização no domínio wavelet encontradas na literatura, pode-se citar a proposta de Yoon *et al.* (2002) onde se sugere a aplicação do DS nos coeficientes wavelet após o devido descarte de coeficientes abaixo de um certo limiar. Walczak *et al.* (1997) propuseram um esquema de padronização univariada dos coeficientes wavelet que explora tanto os níveis de decomposição da árvore wavelet quanto a escolha de filtros adequados para cada nível.

Dentre as técnicas encontradas, uma que apresenta bons resultados é a proposta por Tan e Brown (2001), neste trabalho os autores propõem a separação do sinal em componentes de baixa e de alta freqüência (no sentido da Transformada de Fourier) através dos filtros wavelet. O espectro de baixas freqüências (associado aos coeficientes wavelet de aproximação) é padronizado através do PDS e o de altas (associado aos coeficientes wavelet de detalhes), por DS. Ao final, os resultados são somados. Tal algoritmo é denominado de *Wavelet Hybrid Direct Standardization* (WHDS). No trabalho supracitado, Tan e Brown reportam que o melhor resultado foi obtido com o filtro "Daubechies1" com apenas um nível de decomposição e PDS com janela simétrica de 3 pontos.

2.3.2 Métodos Sem Padronização

Uma alternativa aos métodos de padronização é o uso de técnicas que diminuam, ou suprimam (no caso ideal), as variações instrumentais e/ou ambientais, ou a construção de modelos com o intuito de serem robustos a tais variações. Tais métodos trazem consigo a grande vantagem de não necessitarem de amostras de transferência, como discutido no início desta seção.

Entre as técnicas que tentam diminuir as variações instrumentais, pode-se citar o método das derivadas, em que se remove o *offset* (1ª derivada) e a inclinação (2ª derivada) da linha de base. Como a operação de diferenciação amplifica ruídos presentes nas medições é comum se suavizar os espectros antes da diferenciação. Para isso, um algoritmo bastante difundido é o Savitzky-Golay (SG) (BEEBE; PELL; SEASHOLTZ, 1998) que realiza essas duas operações. Uma outra técnica é a correção multiplicativa de sinais (*Multiplicative Signal Correction*, MSC) (GELADI; MACDOUGALL; MARTENS, 1985), em que cada espectro medido é regredido contra um espectro de referência, geralmente a média do conjunto de calibração, e a correção se dá pela subtração da interseção e pela divisão da inclinação da reta de regressão. Porém, como apontado em (FEUDALE *et al.*, 2002), tais métodos (SG e MSC) oferecem poucos ganhos para a transferência de modelos espectrométricos, uma vez que eles assumem que as diferenças de *offset* e inclinação da linha de base são constantes em toda a região espectral medida, fato que não se verifica na maioria das aplicações espectrométricas.

Uma variação do MSC é a filtragem com filtros de resposta a impulso finita (*Finite Impulse Response*, FIR) (BLANK; SUM; BROWN, 1996). Esse algoritmo se propõe a corrigir características da linha de base que variem ao longo da região espectral em questão. Sua

abordagem equivale ao uso do MSC por partes através de janelas espectrais móveis [9]. Porém, ainda assim, o MSC e a filtragem FIR têm utilidade limitada para transferência dado que essas técnicas forçam os espectros medidos a se parecerem com um único espectro (o médio), sendo suscetível, portanto, a perdas de informações químicas das amostras.

De modo a prevenir perda de informações durante o processamento dos espectros foi proposta a correção de sinal ortogonal (*Orthogonal Signal Correction*, OSC) (WOLD *et al.*, 1998). Originalmente proposto como um método de pré-processamento dos dados, a idéia geral do OSC é reduzir a variação em $\mathbf{X}$ que não está correlacionada, ou seja, que é (matematicamente) ortogonal, a $\mathbf{y}$. No contexto de transferência de calibração, vetores (LV's) que são ortogonais a $\mathbf{y}$ e comuns aos equipamentos envolvidos são removidos de modo a tornar o modelo mais transferível. Entretanto alguns trabalhos recentes vêm questionando a validade do OSC para fins de processamento dos dados e transferência (PIERNA *et al.*, 2001; SVENSSON; KOURTI; MACGREGOR, 2002) *apud* (FEUDALE *et al.*, 2002).

Na abordagem de construção de modelos robustos a variações ambientais/instrumentais é comum o uso de técnicas que tentem extrair características (*"features"*) que apresentem pequenas variações entre os instrumentos em questão. Os modelos são, então, construídos tomando como base essas características. Alguns trabalhos abordando tais aspectos foram desenvolvidos em (MARK; JR., 1988; OZDEMIR; MOSLEY; WILLIAMS, 1998; SWIERENGA *et al.*, 1998a). Em (HONORATO *et al.*, 2005) foi proposto o uso do APS para escolha de variáveis (comprimentos de onda) que sejam apropriados para construção de modelos RLM com boa capacidade preditiva tanto no Primário quanto no Secundário.

Neste trabalho será feita uma extensão do método proposto em (HONORATO *et al.*, 2005) para modelos de regressão linear no domínio wavelet (ver Seção 3.1), o APSW.

[9] O conceito de janelas espectrais móveis já foi discutido na ocasião da introdução ao PDS (ver Seção 2.3.1.2).

Adicionalmente, os modelos finais serão obtidos através de um procedimento conhecido como *subagging* (ver Seção 3.3), APSW-BAG. A utilidade do *subagging* em conjunto com o APS já foi mostrada para fins de calibração multivariada em (GALVÃO *et al.*, 2006). Espera-se que tal junção seja interessante, também, para fins de transferência, seja com modelos construídos no domínio original de comprimento de onda, seja no domínio da transformada wavelet.

3 Técnicas Utilizadas

Este capítulo traz uma breve introdução sobre a teoria da Transformada Wavelet Discreta (DWT), assim como a descrição do Algoritmo das Projeções Sucessivas (APS) e do método *subagging* de reamostragem e combinação de modelos. Serão ainda discutidos algoritmos que fazem uso dessas técnicas, a saber o WHDS, APSO, APSW e APSW-BAG.

3.1 Transformada Wavelet Discreta

A Transformada Wavelet (*Wavelet Transform*, WT) é uma transformação linear que mapeia um sinal unidimensional (geralmente em função do tempo) em um espaço bidimensional (geralmente um plano tempo-escala[1]). Através desse mapeamento é possível realizar uma análise mais detalhada das características do sinal de interesse (DAUBECHIES, 1992; VETTERLI; KOVACEVIC, 1995; STRANG; NGUYEN, 1996; TAN; BROWN, 2001). No caso de espectrometria, o sinal (espectro) é função do comprimento de onda e o plano no domínio wavelet seria definido em termos de comprimento de onda versus taxa de variação do sinal. Contudo, neste trabalho não se faz necessária uma denominação desse plano e pode-se dizer simplesmente domínio wavelet.

[1] É comum se encontrar também a denominação de plano tempo-freqüência, uma vez que escala e freqüência são idéias semelhantes com relação à taxa de variação de algum sinal.

Apesar de haver a versão contínua da WT (CWT), a mesma é redundante e computacionalmente custosa. Uma alternativa sem redundância na análise é a Transformada Wavelet Discreta (DWT). A DWT de um sinal pode ser obtida de forma computacionalmente eficiente através de um banco de filtros (VETTERLI; KOVACEVIC, 1995) estruturado em forma de árvore, como ilustrado na Figura 3.1. A estrutura básica dessa árvore corresponde a um par de filtros passa-baixas/passa-altas (H e G, respectivamente) seguido por operadores de dizimação diádica ($\downarrow 2$)[2]. Tais filtros podem ser escolhidos de modo que a transformação como um todo seja ortonormal, não havendo perda de informação. A saída do canal passa-baixas pode ser sucessivamente decomposta através do uso de pares de filtros adicionais, até um número $ndec$ pré-estabelecido de iterações. Caso o sinal de entrada do banco de filtros seja expresso na forma de um vetor $\mathbf{x} = [x[0]\ x[1]\ \ldots\ x[m-1]]$, o resultado da decomposição será um vetor $\mathbf{v} = [\mathbf{c}(ndec) \mid \mathbf{d}(ndec) \mid \mathbf{d}(ndec-1) \mid \ldots \mid \mathbf{d}(1)]$, em que os elementos dos vetores $\mathbf{c}(k)$ e $\mathbf{d}(k)$ são denominados, respectivamente, coeficientes wavelet de aproximação e de detalhe no k-ésimo nível de decomposição. No contexto de análises espectrométricas, os valores $\{x[0]\ x[1]\ \ldots\ x[m-1]\}$ correspondem a intensidades de sinal registradas em m diferentes comprimentos de onda, formando um espectro.

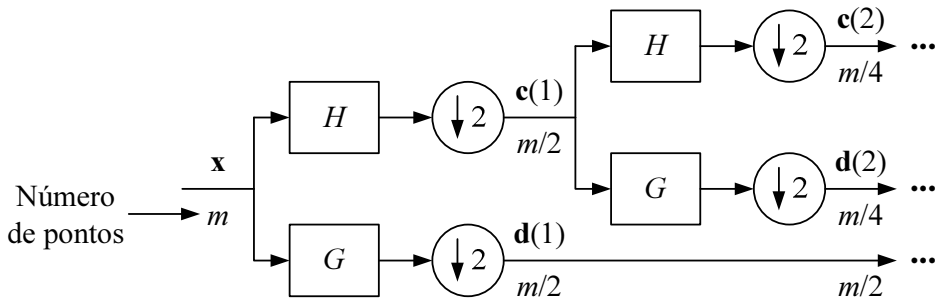


FIGURA 3.1 – Implementação da Transformada Wavelet Discreta através de um banco de filtros. H e G representam, respectivamente, filtros passa-baixas e passa-altas.

Para fins de transferência de calibração, o uso da Transformada Wavelet tem se mos-

[2]Dizimação é o nome dado ao procedimento de se reduzir o número de elementos de uma seqüência, no presente caso tomando um elemento a cada dois.

trado uma ferramenta de grande potencial (FEUDALE *et al.*, 2002; WALCZAK; BOUVERESSE; MASSART, 1997; PARK *et al.*, 2001; YOON; LEE; HAN, 2002), tanto para fins de pré-processamento (correção de variações na linha de base dos espectros, por exemplo) (TAN; BROWN, 2002), quanto para fins de padronização (WALCZAK; BOUVERESSE; MASSART, 1997).

No caso de se realizar a padronização com base em coeficientes wavelet, vale ressaltar que o caráter multirresolucional da análise permite extrair características do sinal através de janelas móveis de tamanho variável (DAUBECHIES, 1992), ao contrário do método PDS, que adota uma janela de largura fixa. Desse modo, a transferência de calibração empregando wavelets pode considerar tanto diferenças locais (pequena escala) quanto globais (grande escala) entre os espectros dos instrumentos Primário e Secundário (YOON; LEE; HAN, 2002).

3.2 APS

O Algoritmo das Projeções Sucessivas (APS) proposto por Araújo *et al.* (2001) é um método de seleção de variáveis que opera usando um conjunto de amostras de calibração (cal) e um de validação (val) contendo as respostas espectrais ($\mathbf{X}$) e os valores do parâmetro de interesse determinados pelo método de referência ($\mathbf{y}$). As principais operações no APS consistem de manipulações algébricas feitas na matriz $\mathbf{X}_{\text{cal}_{[nc\times m]}}$ com nc amostras de calibração e m variáveis espectrais. Partindo de uma coluna $\mathbf{x_0}$ (associada à variável inicial da seleção), a primeira operação do APS determina qual das colunas restantes tem a maior projeção no subespaço S_0 ortogonal a $\mathbf{x_0}$. Essa coluna, denotada por $\mathbf{x_1}$, pode ser interpretada como a que contém a maior quantidade de informação não presente em

$\mathbf{x_0}$. Na próxima iteração, o APS restringe a análise ao subespaço S_0 tomando $\mathbf{x_1}$ como a nova coluna de referência e procede com os passos descritos acima. Portanto, o critério de seleção do APS visa à minimização da colinearidade entre as variáveis. Vale notar que no máximo nc variáveis podem ser selecionadas dessa maneira. De fato, após cada operação de projeção a dimensão do espaço coluna de $\mathbf{X}_{\text{cal}}$ é reduzida por um (i.e., um grau de liberdade é removido). Portanto, após nc operações de projeção todos os vetores coluna de $\mathbf{X}_{\text{cal}}$ seriam projetados na origem do espaço. Dada a regra de construção definida acima, resta determinar a melhor variável inicial (coluna de $\mathbf{X}_{\text{cal}}$) e o número ótimo N de variáveis a serem usadas. Fixando-se N, podem ser montados m subconjuntos de N variáveis, pois cada um desses poderá ter como variável inicial cada uma m variáveis espectrais. Para cada um desses subconjuntos, chamados de "cadeias de variáveis", é calibrado um modelo de Regressão Linear Múltipla (RLM) e calculada a raiz quadrada do erro quadrático médio de previsão no conjunto de validação (RMSEV), dado pela Equação 3.1 abaixo

$$\text{RMSEV} = \sqrt{\frac{1}{nv}\sum_{k=1}^{nv}(\mathbf{y}_{\text{v}}^{k} - \hat{\mathbf{y}}_{\text{v}}^{k})^2} \tag{3.1}$$

em que $\mathbf{y}_{\text{v}}^{k}$ e $\hat{\mathbf{y}}_{\text{v}}^{k}$ são os valores de referência e predito do parâmetro de interesse na k-ésima amostra de validação e nv é o número de amostras de validação. O menor RMSEV obtido é então denotado por RMSEV*(N), em que o "*" é usado para indicar o melhor resultado obtido com as cadeias de N variáveis. Repetindo esse procedimento para N = 1, 2, ..., nc, (note que N não pode ser maior que nc, como explicado anteriormente) o N ótimo pode ser obtido do mínimo da curva de RMSEV*(N).

O tema da próxima subseção será o procedimento matemático envolvido na obtenção do conjunto de variáveis escolhidos pelo APS.

3.2.1 Descrição do Algoritmo das Projeções Sucessivas

Seja $\mathbf{L}$ um vetor que armazena o número das variáveis escolhidas, tal que $\mathbf{L}(0)$ seja pré-definido. Seja ainda N o número de variáveis a serem escolhidas e $\mathbf{x}_j$ a j-ésima coluna de $\mathbf{X}_{\text{cal}}$, $j = 1, \ldots, m$. Os passos que o APS executa são os seguintes:

Passo 0. Faça o contador $i = 1$.

Passo 1. Faça C o conjunto de variáveis que ainda não foram selecionadas. Isto é, $C = \{j \mid 1 \leq j \leq m \text{ e } j \notin \mathbf{L}\}$.

Passo 2. Calcular a projeção de $\mathbf{x}_j$ no subespaço ortogonal a $\mathbf{x}_{\mathbf{L}(i-1)}$ como mostra a Equação 3.2 abaixo

$$\mathbf{P}\mathbf{x}_j = \mathbf{x}_j - (\mathbf{x}_j^T \mathbf{x}_{\mathbf{L}(i-1)})\mathbf{x}_{\mathbf{L}(i-1)}(\mathbf{x}_{\mathbf{L}(i-1)}^T \mathbf{x}_{\mathbf{L}(i-1)})^{-1} \tag{3.2}$$

para todo $j \in C$, onde $\mathbf{P}$ é o operador de projeção.

Passo 3. Faça $\mathbf{L}(i) = \arg(\max \|\mathbf{P}\mathbf{x}_j\|, j \in C)$.

Passo 4. Faça $\mathbf{x}_j = \mathbf{P}\mathbf{x}_j, j \in C$.

Passo 5. Faça $i = i + 1$. Se $i < m$ volte ao Passo 1.

Fim As variáveis escolhidas estão agrupadas em $\mathbf{L}$.

Os passos acima são exemplificados na Figura 3.2 abaixo que ilustra a primeira iteração do APS (ARAÚJO *et al.*, 2001).

Pode-se mostrar que o número de projeções efetuadas durante o processamento do algoritmo é de $(N-1)(m-N/2)$. Vale notar que, embora parecidos, o APS e o procedimento de ortogonalização de Gram-Schmidt têm diferentes objetivos. O último manipula

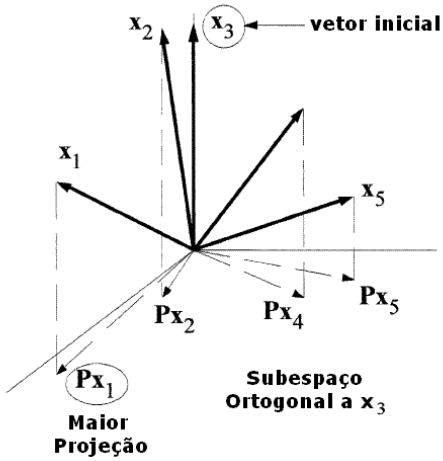


FIGURA 3.2 – Exemplo da aplicação do APS com $m = 5$, $n = 3$ e $\mathbf{L}(0) = 3$. Resultado da primeira iteração: $\mathbf{L}(1) = 1$. (ARAÚJO *et al.*, 2001)

os dados de forma a se obter uma nova base ortogonal que, em geral, não possui significado físico. Já o APS não modifica os dados originais, uma vez que a ortogonalização é usada apenas para fins de seleção. Portanto, permanece, no APS, a relação existente entre as variáveis escolhidas e o espectro original.

3.2.2 Uso do APS para Transferência de Calibração

Neste trabalho, será usada a sigla APSO para o método em que o APS é usado no domínio original da medida dos espectros de absorbância, ou seja, o dos comprimentos de onda. Analogamente, o termo APSW será empregado quando o APS for realizado no domínio da transformada wavelet. Vale ressaltar que o APSW realiza regressão no domínio wavelet (ALSBERG; WOODWARD; KELL, 1997) para a construção dos modelos.

O APS foi originalmente desenvolvido para a seleção de variáveis que levassem a um bom modelo RLM (ARAÚJO *et al.*, 2001). Porém, o mesmo pode ser usado para a construção de modelos a serem usados em mais de um instrumento. Para tanto é desejável que as variáveis escolhidas sejam de certa forma robustas às variações existentes entre

os instrumentos e/ou ambientes em que os instrumentos se encontram. Para empregar o APS na seleção de variáveis robustas a variações instrumentais, adota-se o seguinte índice a ser minimizado (HONORATO *et al.*, 2005):

$$E = \frac{1}{2}(\text{RMSEV} + \text{RMSET}) \tag{3.3}$$

em que RMSET é o erro obtido nas amostras de transferência e possui definição análoga ao RMSEV. Portanto o critério para a seleção das variáveis leva em conta não apenas a capacidade de predição do modelo (medido pelo RMSEV), mas também a robustez do modelo obtido através dessas variáveis (medido pelo RMSET). Na Equação 3.3 o RMSEV e o RMSET possuem o mesmo peso de modo a conferir igual importância a esses dois índices, independentemente do número de amostras de validação e transferência. Portanto, o conjunto de validação no APSO e APSW (ASPO/W) é formado pelas amostras originais de validação e as amostras de transferência, cujos espectros foram medidos no Secundário.

3.3 *Subagging*

Em tarefas de regressão e classificação nas quais o tamanho do conjunto de dados é pequeno, técnicas de reamostragem (GOOD, 1999) têm sido usadas para melhorar a capacidade de generalização dos modelos resultantes e, ainda, testar a sensibilidade do método de modelagem em relação ao conjunto de treinamento ou de teste (DUDA; HART; STORK, 2001). Um método bastante utilizado com esse propósito é o *jacknife* (EFRON, 1982), ou "*leave-one-out*", em que os diferentes modelos são construídos pela remoção e posterior reinserção de um objeto (amostra) por vez do conjunto de treinamento. O erro obtido a cada etapa, dito de validação-cruzada, pode ser usado para apoiar a escolha

de um determinado modelo dentro de uma classe de modelos disponíveis, como é o caso de se escolher o modelo com número adequado de PC's/LV's no PCR/PLS, ou o melhor subconjunto de variáveis para a RLM. Uma técnica similar, conhecida como *bootstrap* (DAVISON; HINKLEY, 1997; SMITH; GEMPERLINE, 2002), forma diferentes conjuntos de calibração através da escolha aleatória, com reposição, de um número fixo de objetos disponíveis do conjunto de dados. Deste modo, diferentes modelos podem ser obtidos como no caso do *jacknife*. Um uso menos comum de tais técnicas consiste em se combinar os diferentes modelos obtidos durante a reamostragem de modo a se gerar um único modelo "conjunto" (OPITZ; MACLIN, 1999). Por exemplo, se estruturas lineares forem utilizadas a combinação pode ser feita utilizando-se a média dos coeficientes dos modelos (SKURICHINA; DUIN, 1998). De modo mais geral, as saídas de classificadores podem ser combinadas por votação de maioria (HANSEN; SALAMON, 1990; BAUER; KOHAVI, 1999) enquanto que em modelos quantitativos pode-se usar a média das saídas (TANIGUCHI; TRESP, 1997). Em muitos casos, o modelo conjunto resultante tem melhor desempenho que seus membros individuais (OPITZ; MACLIN, 1999; HANSEN; SALAMON, 1990).

Tal idéia foi estudada em detalhe por Breiman (1996a, 1996b). Ao processo de se gerar diferentes conjuntos de modelagem por um procedimento de *bootstrap* com posterior combinação desses modelos foi dado o nome de *bagging* (***b**ootstrap **agg**regat**ing***). Breiman mostrou que o *bagging* pode levar a ganhos substanciais de desempenho tanto para classificação quanto para regressão, principalmente quando as alterações no conjunto de treinamento causam mudanças significativas no resultado da etapa de modelagem. Mais especificamente, em tarefas de regressão, nas quais a média quadrática do erro de predição é influenciada pela variância do ruído, pela tendenciosidade[3] do preditor e pela variância

[3]Neste sentido, também se encontram na literatura as palavras viés e polarização.

do preditor, o *bagging* pode ser usado para se reduzir a parcela de variância (BREIMAN, 1996a). Adicionalmente, uma versão modificada do algoritmo do *bagging*, o *iterated bagging*, pode levar a reduções na tendenciosidade do preditor (BREIMAN, 2001). Vale notar que técnicas de *bagging* se tornaram populares para fins de classificação, enquanto que problemas de regressão multivariada têm recebido relativamente pouca atenção, embora já tenha sido apontada sua utilidade em problemas de escolha de variáveis em modelos RLM (BREIMAN, 1996a; BREIMAN, 1996b). Neste caso, mudanças no resultado da seleção causadas por alterações no conjunto de calibração são suavizadas pelo cálculo da média dos modelos resultantes.

Mais recentemente, foi proposto um esquema alternativo de combinação baseado na reamostragem sem reposição (ou "subamostragem") (BÜHLMANN; YU, 2002). Em tal esquema, que recebeu o nome de *subagging* (***sub****sample* ***agg****regati****ng***), cada modelo individual é construído com base em um número reduzido de Mc objetos extraídos de um conjunto de M objetos disponíveis para modelagem ($Mc < M$). Os modelos individuais são, então, combinados para se criar um modelo conjunto, como no *bagging*. Vale ressaltar que o procedimento de *bootstrap* empregado no *bagging* gera conjuntos de calibração com o mesmo tamanho M do conjunto original de modelagem. Uma vantagem imediata do *subagging* é sua menor carga computacional, uma vez que modelos são construídos mais rapidamente se menos objetos forem usados na etapa de calibração. Uma observação interessante é que tal vantagem computacional pode ser obtida a custa de apenas uma pequena diminuição no desempenho do modelo conjunto, como foi demonstrado por exemplos numéricos em (BÜHLMANN; YU, 2002).

No presente trabalho, uma estratégia de *subagging* é proposta com o fim de se melhorar o desempenho, em termos do RMSEP, dos modelos APSO/W calibrados no Primário

para uso no Secundário, ou seja, como uma parte do procedimento de transferência de calibração.

3.3.1 O Método Proposto

O uso do *subagging* requer um procedimento para a obtenção de um modelo para cada reamostragem do conjunto de dados e um método para a combinação dos modelos resultantes. Esta seção descreve a implementação desses procedimentos, assim como algumas considerações a respeito dessa implementação.

O conjunto de dados disponíveis para a construção do modelo será chamado de "conjunto de modelagem", com M amostras. As Mc amostras extraídas por reamostragem do conjunto de modelagem formarão o "conjunto de calibração". Vale ressaltar que a reamostragem é feita sem reposição, como discutido previamente. As Mv amostras restantes no conjunto de modelagem após a extração do conjunto de calibração formarão o "conjunto de validação", usado pelo APS durante o processo de escolha de variáveis (ver Seção 3.2). No caso do APSO/W ainda há as $Ntrans$ amostras de transferência que se juntarão ao conjunto de validação de modo que as variáveis escolhidas possuam, também, boa capacidade preditiva no instrumento Secundário.

Após selecionadas as amostras, a construção do modelo se dá por RLM em conjunto com o APS (APS-RLM). Tal procedimento é repetido por um número $nbag$ pré-estabelecido de iterações do algoritmo. Os modelos gerados nesse esquema são combinados através do cálculo da média das saídas de cada modelo, como ilustrado na Figura 3.3. A combinação resultante é chamada de modelo conjunto.

Assumindo que a relação entre a saída $\hat{\mathbf{y}}$ (parâmetro estimado) e a entrada $\mathbf{X}$ (resposta

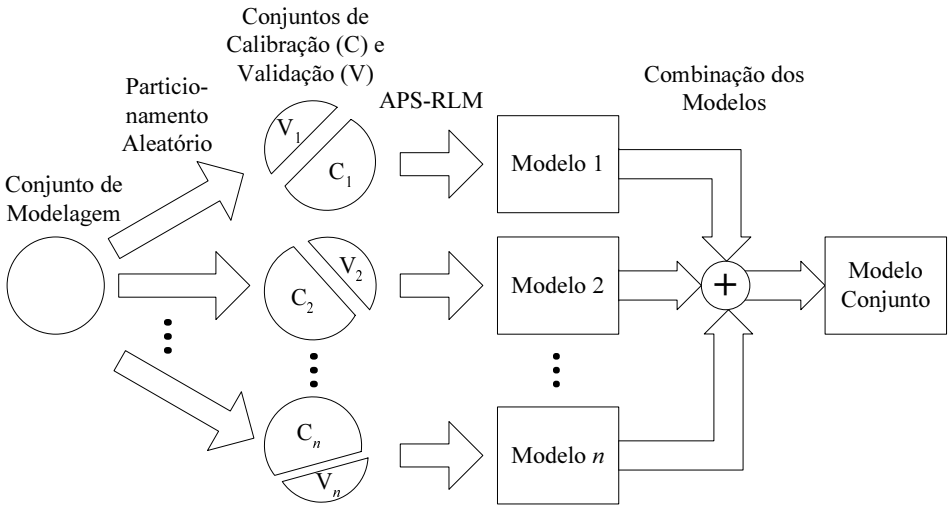


FIGURA 3.3 – Procedimento do *subagging* para geração do modelo conjunto. Os conjuntos de validação são usados no processo de escolha das variáveis.

instrumental) para o i-ésimo modelo individual seja expressa como

$$\hat{\mathbf{y}}_i = f_i(\mathbf{X}) \tag{3.4}$$

então a relação entrada-saída para o modelo conjunto formado a partir de $nbag$ modelos individuais é dada por

$$\hat{\mathbf{y}}_{\text{subag}(nbag)} = f_{\text{subag}(nbag)}(\mathbf{X}) = \frac{1}{nbag}\sum_{i=1}^{nbag}\hat{\mathbf{y}}_i = \frac{1}{nbag}\sum_{i=1}^{nbag} f_i(\mathbf{X}) \tag{3.5}$$

É interessante observar que se $f_i(\mathbf{X})$ é linear, como no caso da RLM, o procedimento acima recai no cálculo da média dos coeficientes dos modelos individuais. Essa interpretação torna mais fácil se notar o efeito suavizador causado pelo *subagging* na seleção de variáveis, uma vez que a cada nova reamostragem diferentes variáveis podem ser escolhidas para a construção dos modelos individuais e ao final elas serão combinadas para a construção do modelo conjunto.

O método implementado nesse trabalho consiste de uma junção do APSO/W e o *su-*

bagging referenciados como APSO/W-BAG. Cada modelo individual é formado de acordo com os algoritmos do APSO/W e o modelo conjunto é formado de acordo com a Equação 3.5. Uma particularidade dos métodos APSO/W-BAG implementados é que as cadeias de variáveis (ver Seção 3.2) são determinadas a partir do conjunto de modelagem e não a partir do conjunto de calibração como preconiza o método *subagging* tradicional implementado em (GALVÃO *et al.*, 2006). Pode-se argumentar que tal particularidade pode causar perda de diversidade das variáveis selecionadas. Porém, por outro lado, a carga computacional é sensivelmente reduzida e, além do mais, a perda de diversidade não é tão pronunciada dado que são formadas m cadeias de N variáveis, em que $m = 3221$ e $N = 30$ para os dados de gasolina, havendo, portanto, uma quantidade de cadeias suficiente para se obter ganhos com o emprego do *subagging* (ver Capítulo 5).

De fato, a veracidade dessa conjectura foi testada e se constatou uma diferença insignificante. Na Figura 3.4 vêem-se as curvas de RMSEP (ver Seção 4.3) para o método implementado neste trabalho (APSO-BAG-v0) e o alternativo implementado em (GALVÃO *et al.*, 2006) (APSO-BAG-v1). Apesar de estar sendo mostrada apenas a propriedade T10, o mesmo comportamento foi observado nas demais propriedades.

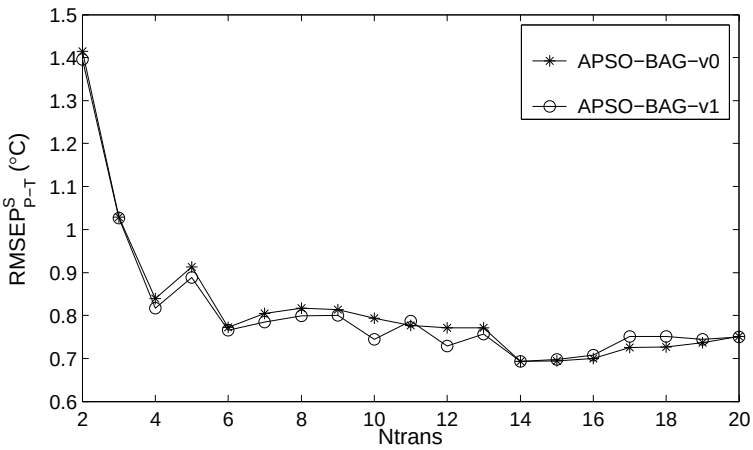


FIGURA 3.4 – Comparação entre as curvas de RMSEP de APSO-BAG-v0 e APSO-BAG-v1 para a propriedade T10.

4 Material e Métodos

4.1 Dados de Gasolina

Foi empregado um banco de dados de 103 amostras de gasolina coletadas nos estados de Pernambuco e Alagoas, tendo sido considerados três parâmetro de qualidade de gasolina, de acordo com os Regulamentos Técnicos da ANP Nº 02/2005, a saber, a massa específica (ME) e dois pontos da curva de destilação (temperaturas para 10% (T10) e 90% (T90) de evaporado). O espectro de cada amostra foi registrado na faixa de 2500 a 15400 nm em dois espectrômetros FT-IR Perkin Elmer Spectrum GX dotados de sonda ATR. O instrumento Primário (no qual a calibração é inicialmente realizada) foi operado em um laboratório conveniado à ANP, sob controle rigoroso de temperatura e umidade. Já o instrumento Secundário (para o qual a calibração deve ser transferida), foi operado em um laboratório multi-usuário com menor controle ambiental.

4.1.1 Partição das Amostras em Conjuntos

As 103 amostras foram divididas em 63 de calibração e 20 de validação para o PDS-PLS e APSO/W, enquanto que essas mesmas amostras formaram o conjunto de modelagem para o APSO/W-BAG. As demais 20 amostras foram mantidas à parte para fins de

previsão. Adotou-se aqui o particionamento dos conjuntos de acordo com o realizado em (HONORATO *et al.*, 2005).

As amostras de validação foram usadas para se escolher o número adequado de fatores do modelo de calibração, no caso do PLS-PDS, e para auxílio (junto com as amostras de transferência) da escolha das variáveis com melhor capacidade de predição no Secundário, no caso do APSO/W. As amostras de predição foram deixadas à parte durante o processo de calibração/validação de modo que se pudesse realizar testes como se fossem realmente novas amostras com parâmetros desconhecidos.

4.1.2 Calibração

Os dados foram pré-processados primeiramente reduzindo-se sua resolução através de uma dizimação diádica, como em (HONORATO *et al.*, 2005). Os espectros nessa etapa, ditos brutos, contêm $m = 3221$ variáveis espectrais e podem ser visualizados na Figura 4.1. Em seguida, foi aplicado um segundo pré-processamento, consistindo no uso do algoritmo de Savitzky-Golay (SG) com um polinômio de 2ª ordem e uma janela de 21 pontos. Tal algoritmo consiste em uma suavização e diferenciação dos espectros analisados. Pretende-se, com tanto, diminuir o efeito do ruído (suavização) e remover as variações da linha de base (diferenciação) dos espectros (BEEBE; PELL; SEASHOLTZ, 1998). Após essa etapa os espectros, ditos derivativos, são reduzidos para $m = 3201$ variáveis espectrais e sua nova forma é ilustrada na Figura 4.2. Na Figura 4.2c vêem-se as diferenças entre os espectros medidos no Primário e no Secundário.

Para calibração dos modelos espectrométricos foi usado o método dos Mínimos-Quadrados Parciais (PLS) (WOLD; SJOSTROM; ERIKSSON, 2001) adotando a versão PLS1 que

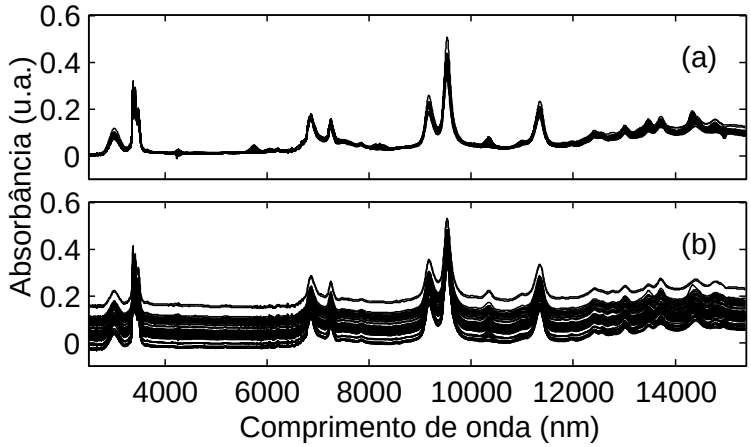


FIGURA 4.1 – Espectros brutos: (a) Primário e (b) Secundário.

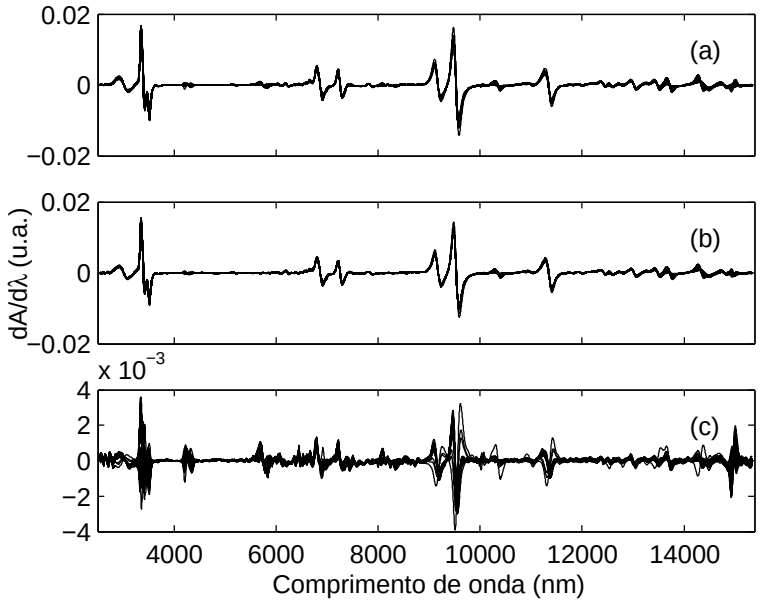


FIGURA 4.2 – Espectros derivativos: (a) Primário, (b) Secundário e (c) diferenças entre o Primário e o Secundário.

considera apenas uma propriedade de interesse. Para escolha do número ótimo de fatores do modelo PLS foi usado o critério $F(0.25)$ de Haaland e Thomas (HAALAND; THOMAS, 1988) (ver Seção 2.2.2.1). As curvas de MSEV[1], sobre a qual o critério F se baseia, para as propriedades em estudo podem ser vistas na Figura 4.3. O número de LV's escolhido para cada modelo foi de 4, 4 e 5 para ME, T10 e T90 respectivamente.

[1]O MSEV é definido como a média do PRESS no conjunto de validação. Portanto, é indiferente aplicar o critério F sobre o PRESS ou MSEV, pois ambos diferem entre si apenas por um fator de escala.

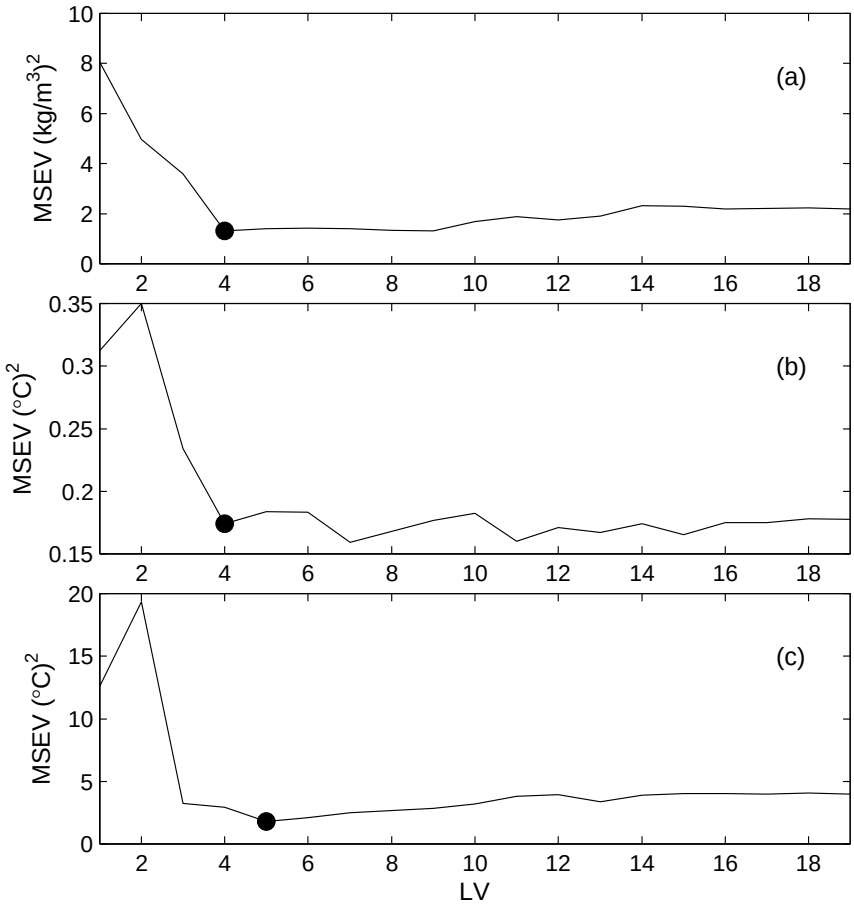


FIGURA 4.3 – Gráfico de MSEV em função do número de LV's. O número de LV's escolhido encontra-se destacado nas figuras. (a) ME, (b) T10 e (c) T90.

Os modelos APSO/W foram construídos, inicialmente, sem o uso de amostras de transferência. O APSW foi realizado usando-se a wavelet "Symlet4" (DAUBECHIES, 1992). A família Symlet foi escolhida com base em trabalhos anteriores na área de espectroscopia no infravermelho (ALSBERG; WOODWARD; KELL, 1997; LIU; BROWN, 2004).

4.1.3 Transferência de Calibração

As técnicas clássicas de transferência de calibração usadas foram as baseadas na padronização da resposta espectral (ver Seção 2.3.1), mais especificamente o DS, o PDS e o WHDS com filtro "Daubechies1"[2] e janela PDS de 3 pontos, como sugerido em (TAN;

[2]Também conhecido como filtro de Haar.

BROWN, 2001). Os algoritmos de DS e PDS foram implementados usando-se correção de tendências aditivas (ver Seção 2.3.1.2).

As abordagens de transferência de calibração por padronização necessitam que os espectros de algumas amostras sejam medidos no instrumento Primário e no Secundário. A escolha desse conjunto deve ser feita de maneira apropriada, pois as amostras devem ser representativas o suficiente para poder descrever as diferenças instrumentais e ainda representar as variações químicas presentes na análise (BOUVERESSE; MASSART, 1996). Portanto, as amostras de transferência foram escolhidas entre as amostras de calibração, pois devido à sua variabilidade, no sentido de varredura do espaço amostral, é razoável supor que estas sejam as mais representativas das diferenças instrumentais (BOUVERESSE; MASSART, 1996). Essas amostras foram escolhidas usando-se o algoritmo Kennard-Stone que busca fazer uma amostragem uniforme do espaço, maximizando as distâncias euclidianas entre os espectros das amostras selecionadas (KENNARD; STONE, 1969). Foi estudado o efeito de se variar o número de amostras de transferência ($Ntrans$) de 2 a 20.

A Figura 4.4 (gráfico dos *scores*) mostra a distribuição das amostras de calibração no espaço formado pelas duas primeiras PC's (PC1 e PC2) que explicam, respectivamente, 49.1% e 22.5% da variância total das amostras de calibração. Na mesma figura ainda é ilustrada a seleção das 10 primeiras amostras de transferência pelo algoritmo de Kennard-Stone. As Figuras 4.5 e 4.6 mostram a porcentagem da máxima variação em $\mathbf{y}$ e nas PC's cobertas pelas amostras de transferência. O intuito dessas Figuras é investigar se as amostras de transferência são, de fato, representativas do conjunto de calibração e se o modelo transferido estaria extrapolando durante as novas predições ou não.

Os resultados de APSO/W e APSO/W-BAG empregando amostras de transferência foram comparados com os obtidos pelos modelos PLS com o uso de PDS (PDS-PLS). No

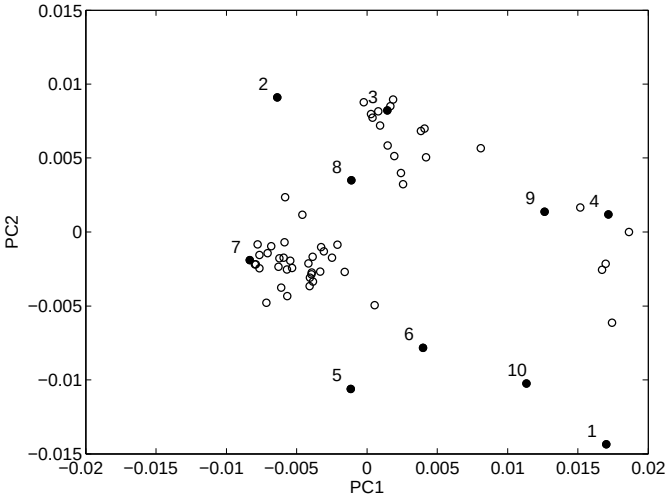


FIGURA 4.4 – Gráfico dos *scores*. Distribuição das amostras no espaço formado pelas duas primeiras PC's e as 10 primeiras amostras de transferência escolhidas pelo algoritmo de Kennard-Stone.

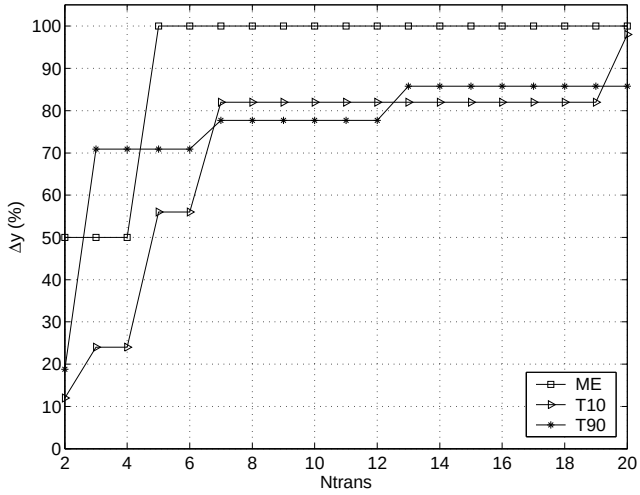


FIGURA 4.5 – Faixa da variação de **y** coberta pela amostras de transferência para as três propriedades.

APSO/W-BAG, o número de iterações do *subagging* foi fixado em 30, pois em um trabalho anterior foi mostrado que a partir desse número a redução no erro de predição deixa de ser significativa (GALVÃO *et al.*, 2006). Ainda nesse trabalho, os autores empregaram uma divisão de 60% das amostras para calibração e 40% para validação, relação que será mantida no presente estudo, ou seja $Mc = 50$ e $Mv = 33$ no procedimento do *subagging*.

Vale ressaltar que o APSO/W não requer que as amostras de transferência sejam

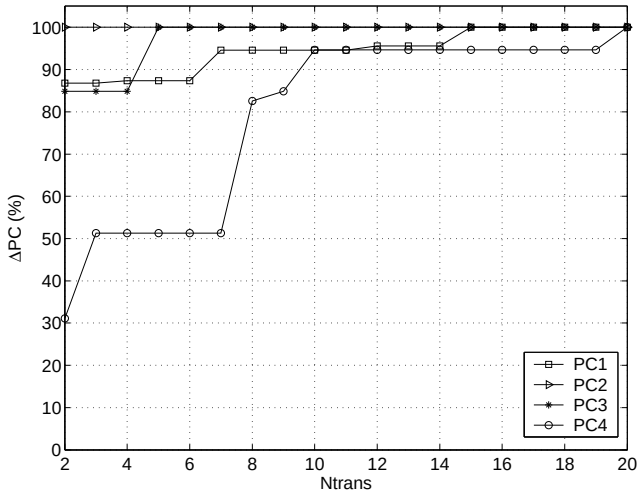


FIGURA 4.6 – Faixa da variação das PC's coberta pela amostras de transferência. As PC's de 1 a 4 explicam, respectivamente, 49.10%, 22.53%, 15.89% e 4.29% da variância total das amostras de calibração.

medidas nos dois instrumentos, ao contrário da técnica clássica PDS-PLS. Porém, foram usadas as mesmas amostras de transferência na avaliação das técnicas em estudo de modo a facilitar a comparação dos resultados.

4.2 Dados de Milho

Com o intuito de se testar a validade das técnicas propostas neste trabalho, foi usado um conjunto de dados adicional. Tal conjunto é constituído de espectros de milho medidos na região de Infravermelho Próximo através de um instrumento de refletância difusa (PASQUINI, 2003). Resumidamente, a refletância difusa é uma técnica relacionada com a espectroscopia de absorção que mede a radiação refletida na substância em análise. Tais dados de milho foram usados em trabalhos recentes como (ANDREW; FEARN, 2004; HONORATO *et al.*, 2005) e estão publicamente disponíveis em http://www.eigenvector.com/Data/Corn/.

Nesta seção será conduzida uma apresentação concisa dos dados de milho.

O particionamento dos conjuntos de calibração, validação e predição adotado foi o seguido em (HONORATO *et al.*, 2005) e detalhado na Seção 4.2.1. A Figura 4.7 mostra os espectros brutos das amostras de milho. A Figura 4.8 ilustra os espectros derivativos e suas diferenças após a aplicação do SG.

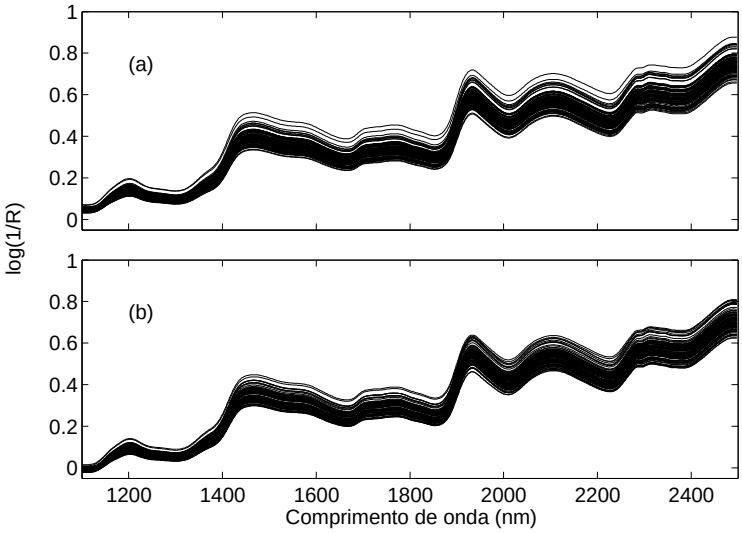


FIGURA 4.7 – Espectros brutos das amostras de milho: (a) Primário e (b) Secundário.

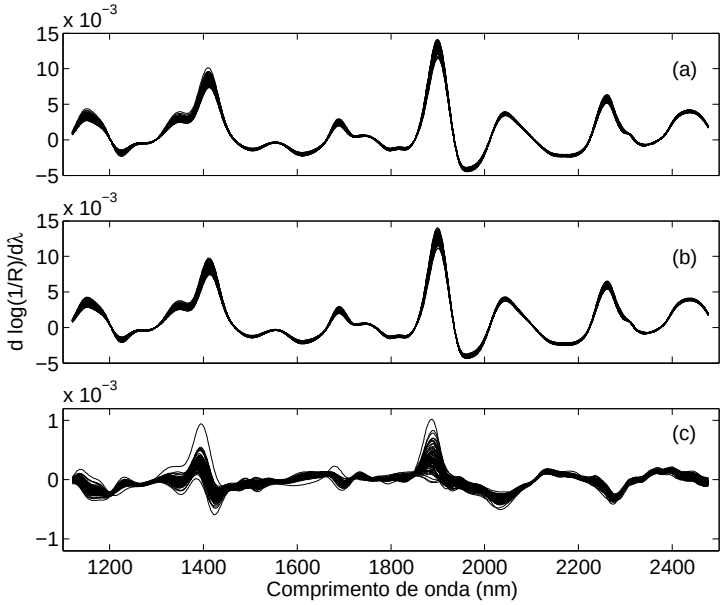


FIGURA 4.8 – Espectros derivativos das amostras de milho: (a) Primário, (b) Secundário e (c) diferenças entre o Primário e o Secundário.

A propriedade escolhida para este estudo é a "umidade" e sua unidade é denotada por

"%m/m" (porcentagem massa/massa). O modelo PLS foi construído usando-se 15 LV's, de acordo com o critério $F(0.25)$ de Haaland e Thomas (1988). A curva do MSEV pode ser vista na Figura 4.9.

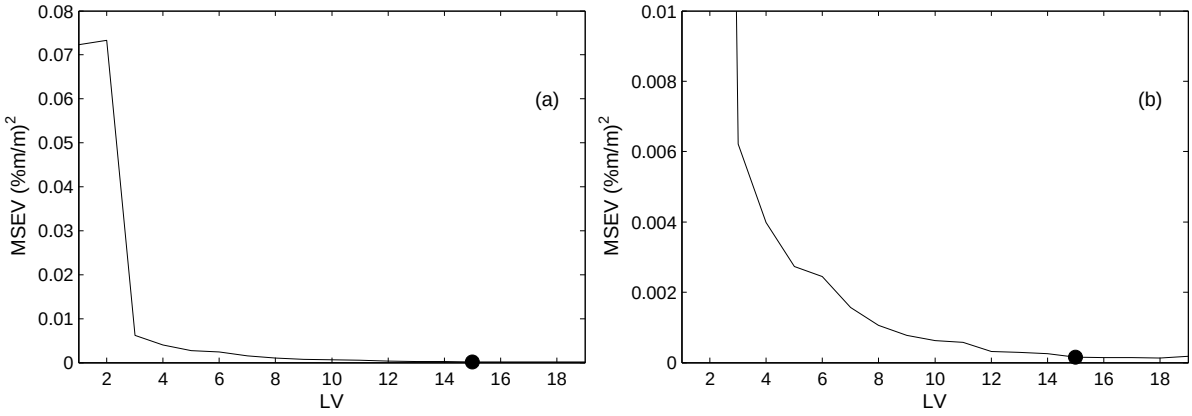


FIGURA 4.9 – Gráfico de MSEV em função do número de LV's para as amostras de milho (a). Em (b) uma ampliação para se visualizar o ponto da curva escolhido. O número de LV's escolhido encontra-se destacado.

Os modelos APSW foram construídos utilizando-se filtros wavelet Symlet4 (mesma escolha adotada para os dados de gasolina).

A Figura 4.10 mostra o gráfico dos *scores*. Também é ilustrada a escolha das 10 primeiras amostras de transferência pelo algoritmo de Kennard-Stone. A Figura 4.11 e 4.12 mostram a porcentagem da máxima variação em $\mathbf{y}$ e nas PC's cobertas pelas amostras de transferência. As PC's de 1 a 4 explicam, respectivamente, 87.10%, 8.91%, 1.23% e 0.86% da variância total das amostras de calibração. O intuito dessas figuras já foi discutido para os dados de gasolina (ver Seção 4.1.3).

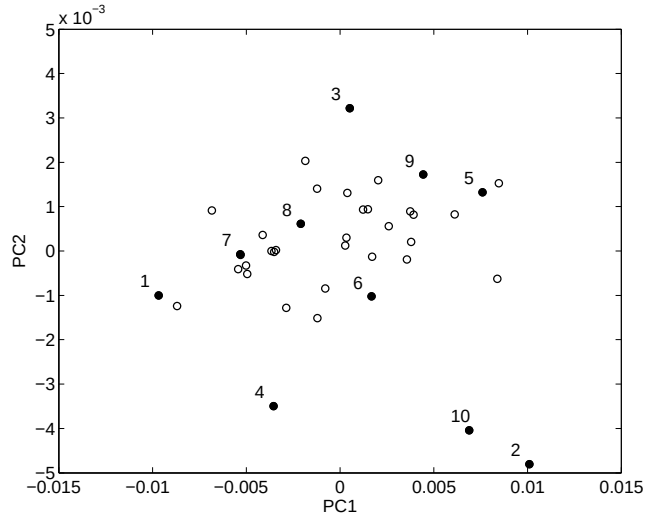


FIGURA 4.10 – Gráfico dos *scores* das amostras de milho. Distribuição das amostras no espaço formado pelas duas primeiras PC's e as 10 primeiras amostras de transferência escolhidas pelo algoritmo de Kennard-Stone.

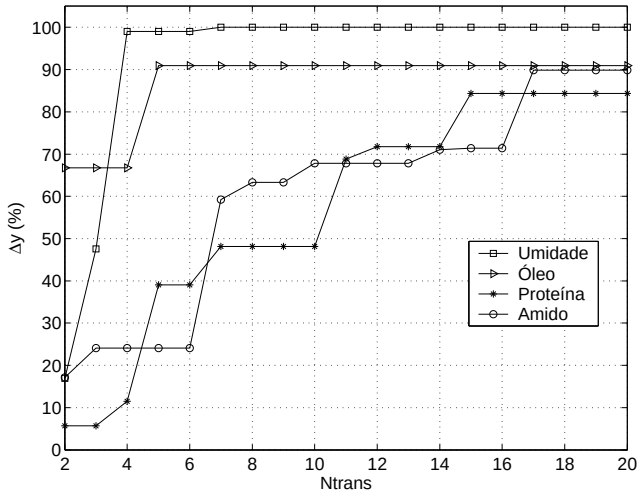


FIGURA 4.11 – Faixa da variação de **y** coberta pela amostras de transferência para os dados de milho.

4.2.1 Particionamento dos Conjuntos

Nesta seção, o particionamento dos conjuntos é detalhado com o intuito de permitir eventual reprodução dos experimentos realizados com os dados de milho. O instrumento Primário foi escolhido como sendo o "m5spec" e nele será conduzido a calibração do modelo. O Secundário é o "mp5spec".

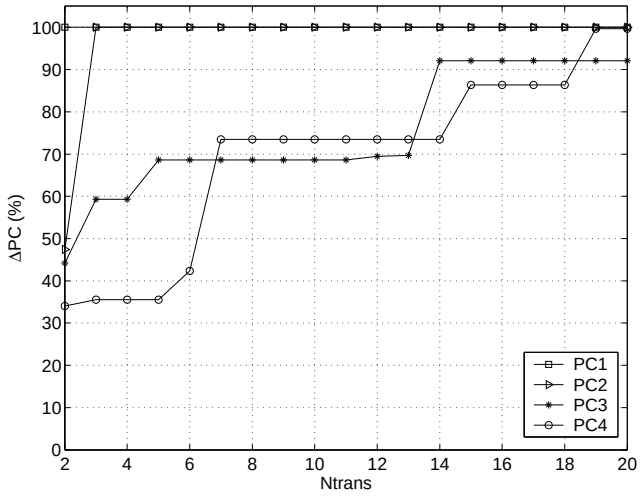


FIGURA 4.12 – Faixa da variação das PC's coberta pela amostras de transferência para os dados de milho.

Conjunto de Validação - formado pelas seguintes amostras:

$[2, 5, 6, 12, 15, 22, 30, 31, 33, 37, 40, 41, 44, 51, 54, 56, 57, 65, 67, 76]$

Conjunto de Predição - formado pelas seguintes amostras:

$[1, 4, 9, 10, 17, 19, 20, 28, 32, 36, 39, 42, 52, 58, 59, 64, 69, 72, 73, 79]$

Conjunto de Calibração - as restantes.

Amostras de Transferência - escolhidas como sendo:

$[55, 75, 71, 80, 27, 25, 66, 46, 34, 77, 78, 7, 43, 47, 14, 48, 11, 35, 63, 62]$

Vale ressaltar que as amostras estão dispostas em ordem de escolha pelo algoritmo de Kennard-Stone. Nota-se que elas fazem parte do conjunto de calibração, como descrito na Seção 4.1.3.

4.3 Forma de Análise dos Resultados

A capacidade de previsão dos modelos foi aferida com base na raiz quadrada do erro quadrático médio de previsão RMSEP (*Root Mean Square Error of Prediction*) obtido no

conjunto de predição. O RMSEP é definido como

$$\text{RMSEP} = \sqrt{\frac{1}{Npred} \sum_{k=1}^{Npred} (y_k - \hat{y}_k)^2} \tag{4.1}$$

em que y_k e $\hat{y}_k$ são os valores de referência e predito na k-ésima amostra de predição para o parâmetro em questão e $Npred$ é o número de amostras de predição (20 no presente caso).

Para cada parâmetro analisado o RMSEP obtido no instrumento Secundário após a transferência foi calculado para $Ntrans$ diferentes números de amostras de transferência com $Ntrans = 2, 3, \ldots, 20$.

4.4 *Hardware* e *Software*

Os algoritmos usados nesse trabalho foram implementados em ambiente MATLAB® versão 6.5 R13 (*The MathWorks, Inc.*).

A configuração do computador usado consiste em um processador Pentium-IV® (*Intel Corp.*), com *clock* de 3.0 GHz e 1GB de memória RAM (*Random Access Memory*), com sistema operacional Windows XP® (*Microsoft Corp.*).

5 Resultados

Neste capítulo serão analisados os resultados da transferência de calibração obtidos com uso de técnicas clássicas, assim como os obtidos com as técnicas propostas. Os resultados, para cada método, são avaliados de acordo com a evolução do RMSEP com o número de amostras de transferência.

A Tabela 5.1 contém as unidades das propriedades analisadas neste trabalho, com o intuito de facilitar a leitura das demais tabelas presentes neste capítulo. Adicionalmente, será introduzida uma notação para auxiliar a apresentação dos resultados.

Notação

RMSEP: Raiz quadrada do erro quadrático médio de previsão.

$\mathbf{RMSEP}^{\mathrm{B}}_{\mathrm{A}}$: RMSEP obtido no instrumento B usando-se o modelo calibrado no instrumento A. A e B podem ser tanto P (Primário) quanto S (Secundário).

$\mathbf{RMSEP}^{\mathrm{B}}_{\mathrm{A\text{-}T}}$: RMSEP obtido no instrumento B após transferência pelo método T usando-se o modelo calibrado no instrumento A.

TABELA 5.1 – Unidades das propriedades em análise.

Propriedade	Unidade
ME	kg/m^3
T10	°C
T90	°C
Umidade	%m/m

5.1 Calibração - Gasolina

Um resumo da etapa de calibração/validação dos modelos de predição PLS é encontrado na Tabela 5.2. Nesta tabela, as colunas $\min(\mathbf{y})$ e $\max(\mathbf{y})$ se referem aos valores mínimos e máximos presentes nas amostras de predição, respectivamente. $\bar{y}$ se refere ao valor médio do parâmetro sendo estimado, a coluna LV's indica com quantas LV's o modelo PLS foi levantado e "S/P" indica a razão entre os RMSEP's no Secundário e no Primário. Como se pode notar, ao se aplicar o modelo calibrado no Primário diretamente no Secundário, sem nenhum procedimento de transferência de calibração, os resultados invariavelmente pioraram. Nas propriedades em estudo, o RMSEP no Secundário se mostrou, no mínimo, maior que o dobro do Primário, cenário que motiva o uso de alguma técnica de transferência de calibração.

TABELA 5.2 – Resumo da etapa de calibração dos modelos PLS.

Prop	$\min(\mathbf{y})$	$\bar{y}$	$\max(\mathbf{y})$	LV's	$\text{RMSEP}_{\text{P}}^{\text{P}}$	$\text{RMSEP}_{\text{P}}^{\text{S}}$	S/P
ME	744.5	755.3	763.3	4	1.4	5.0	3.6
T10	52.2	54.5	56.2	4	0.6	1.6	2.7
T90	154.6	166.8	177.0	5	2.3	9.4	4.1

Os resultados obtidos com o uso de técnicas de transferência de calibração serão analisados na próxima seção.

5.2 Transferência de Calibração - Gasolina

Algumas das técnicas de transferência de calibração descritas na Seção 2.3 são aqui aplicadas ao conjunto de dados de gasolina detalhado na Seção 4.1. Primeiramente, será avaliado o desempenho dos métodos de padronização DS, PDS, WHDS. Na seção subseqüente, serão avaliadas as técnicas APSO/W e APSO/W-BAG. Os resultados desta seção foram relatados de forma preliminar em (MARTINS *et al.*, 2006a) para os métodos de padronização e em (MARTINS *et al.*, 2006b; MARTINS; GALVÃO; PIMENTEL, 2006) para os métodos de seleção de variáveis.

5.2.1 Métodos de Padronização - Gasolina

Uma característica comum às técnicas de padronização espectral é que, em geral, após a padronização os novos espectros (medidos no Secundário) tendem a ser mais parecidos com os que teriam sido medidos no Primário. Algumas ilustrações com os espectros medidos no Primário, no Secundário e após a padronização com 9 amostras de transferência[1] são mostradas nas duas figuras a seguir. O objetivo é se obter uma melhor compreensão das transformações efetuadas nos espectros e o porquê da melhora do $\mathrm{RMSEP}^{\mathrm{S}}_{\mathrm{P\text{-}T}}$ após a padronização. A Figura 5.1 apresenta uma amostra de predição arbitrária medida no Primário e no Secundário e, ainda, o Secundário padronizado por DS. As correções locais que a padronização efetua são bem visíveis. Apenas o DS é mostrado, pois o efeito é semelhante nas outras técnicas. A Figura 5.2 mostra as diferenças entre os espectros derivativos originais e após a padronização com DS, PDS e WHDS. Nota-se que o espectro padronizado, de fato, se assemelha mais ao espectro do Primário e, portanto, seria de se esperar que o modelo de calibração obtivesse melhor desempenho após a padronização do

[1] Essa escolha será justificada no desenvolver desta seção. Por ora, o foco da apresentação é na didática.

espectro do Secundário.

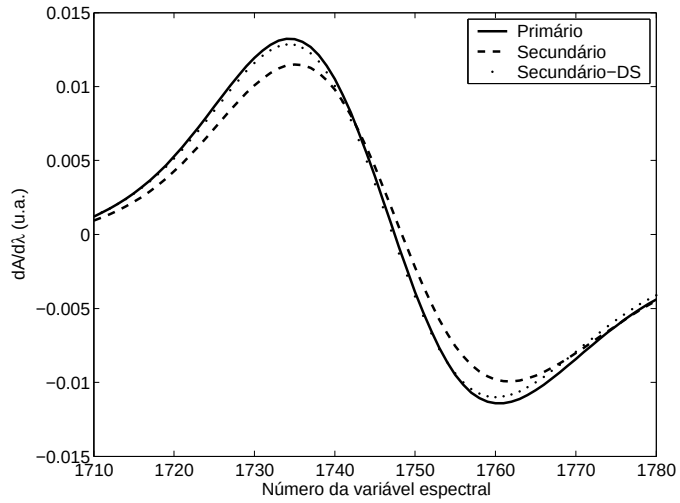


FIGURA 5.1 – Correções locais efetuadas pela padronização DS com 9 amostras de transferência de uma amostra arbitrária.

A seguir uma análise de desempenho mais detalhada para cada método em específico.

5.2.1.1 DS

A Figura 5.3 apresenta o resultado da Padronização Direta (DS), em termos do RMSEP, obtido no escravo ($\text{RMSEP}^{\text{S}}_{\text{P-DS}}$) em função do número de amostras de transferência ($Ntrans$). A linha tracejada indicada por $\text{RMSEP}^{\text{S}}_{\text{P}}$ representa o valor do RMSEP obtido no Secundário sem o uso de técnicas de transferência e a linha indicada por $\text{RMSEP}^{\text{S}}_{\text{P}}$, o valor obtido no Primário.

Pode-se notar uma tendência de diminuição de $\text{RMSEP}^{\text{S}}_{\text{P-DS}}$ com o aumento de $Ntrans$, tendo-se atingido uma estabilização após cerca de 9 amostras de transferência (levando-se em conta o comportamento de todas as propriedades). Com 9 amostras de transferência, todas propriedades já apresentavam resultados melhores do que sem a aplicação do DS. Vale observar que já com $Ntrans = 3$ é possível perceber uma melhora do $\text{RMSEP}^{\text{S}}_{\text{P-DS}}$ em relação ao caso sem transferência, porém esse ponto ainda se apresenta em uma região

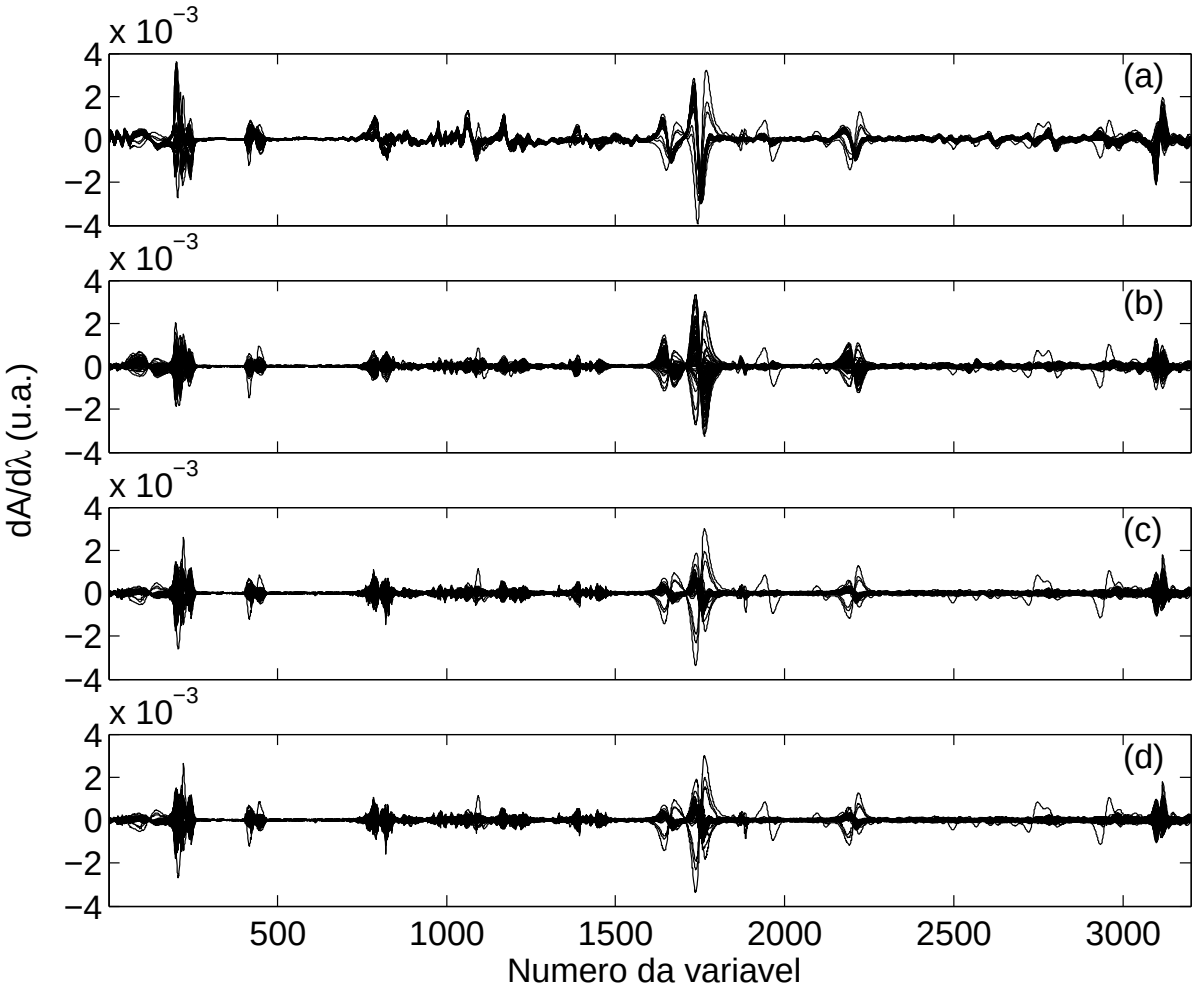


FIGURA 5.2 – Diferenças nos espectros derivativos (a) antes e após a padronização com 9 amostras de transferência por (b) DS, (c) PDS (janela de três pontos) e (d) WHDS.

instável das curvas.

5.2.1.2 PDS

A Figura 5.4 apresenta o resultado da Padronização Direta por Partes (PDS) em termos do RMSEP obtido no escravo ($\mathrm{RMSEP}^{\mathrm{S}}_{\mathrm{P\text{-}PDS}}$) em função de *Ntrans* e do tamanho da janela usada no algoritmo do PDS. A janela é simétrica e possui um número ímpar de pontos variando entre 3 e 15. Os planos $\mathrm{RMSEP}^{\mathrm{S}}_{\mathrm{P}}$ e $\mathrm{RMSEP}^{\mathrm{P}}_{\mathrm{P}}$ são definidos de maneira análoga ao caso do DS. Assim como no DS, observa-se uma diminuição do RMSEP com *Ntrans*, porém a curva do $\mathrm{RMSEP}^{\mathrm{S}}_{\mathrm{P\text{-}PDS}}$ se estabiliza, aproximadamente, com 7 amos-

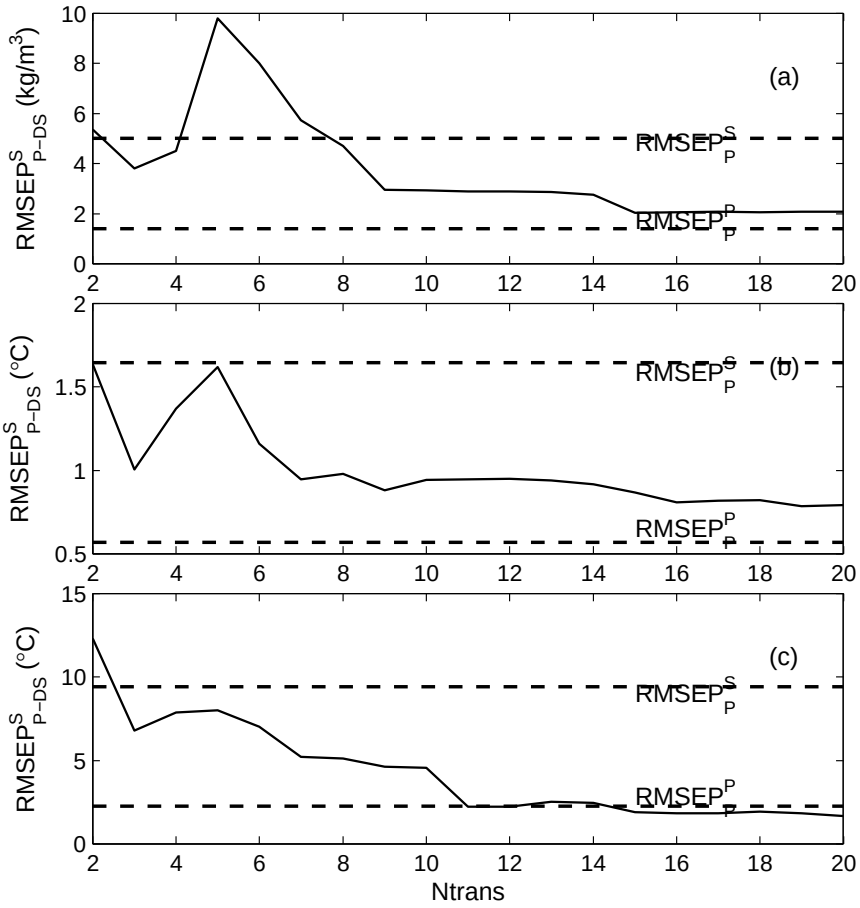


FIGURA 5.3 – RMSEP em função de *Ntrans* obtido com DS. (a) ME, (b) T10 e (c) T90.

tras de transferência. Diferentemente do DS, a partir de 3 amostras de transferência o $\mathrm{RMSEP}^{\mathrm{S}}_{\mathrm{P\text{-}PDS}}$ se mostrou sempre menor que o $\mathrm{RMSEP}^{\mathrm{S}}_{\mathrm{P}}$.

No PDS, além do número de amostras de transferência, deve-se analisar ainda o tamanho da janela usada. O tamanho da janela se mostra um fator crítico quando se possui poucas amostras de transferência, mas torna-se pouco pronunciado quando o número de amostras de transferência empregadas é grande. Observa-se que após 7 amostras de transferência a influência do tamanho da janela é muito pouca, podendo se fixar esse tamanho em 3. A rigor, a janela de tamanho 3 apresentou resultados, em geral, um pouco melhores que as demais janelas.

Os resultados obtidos com o PDS foram melhores que os obtidos pelo DS, porém o

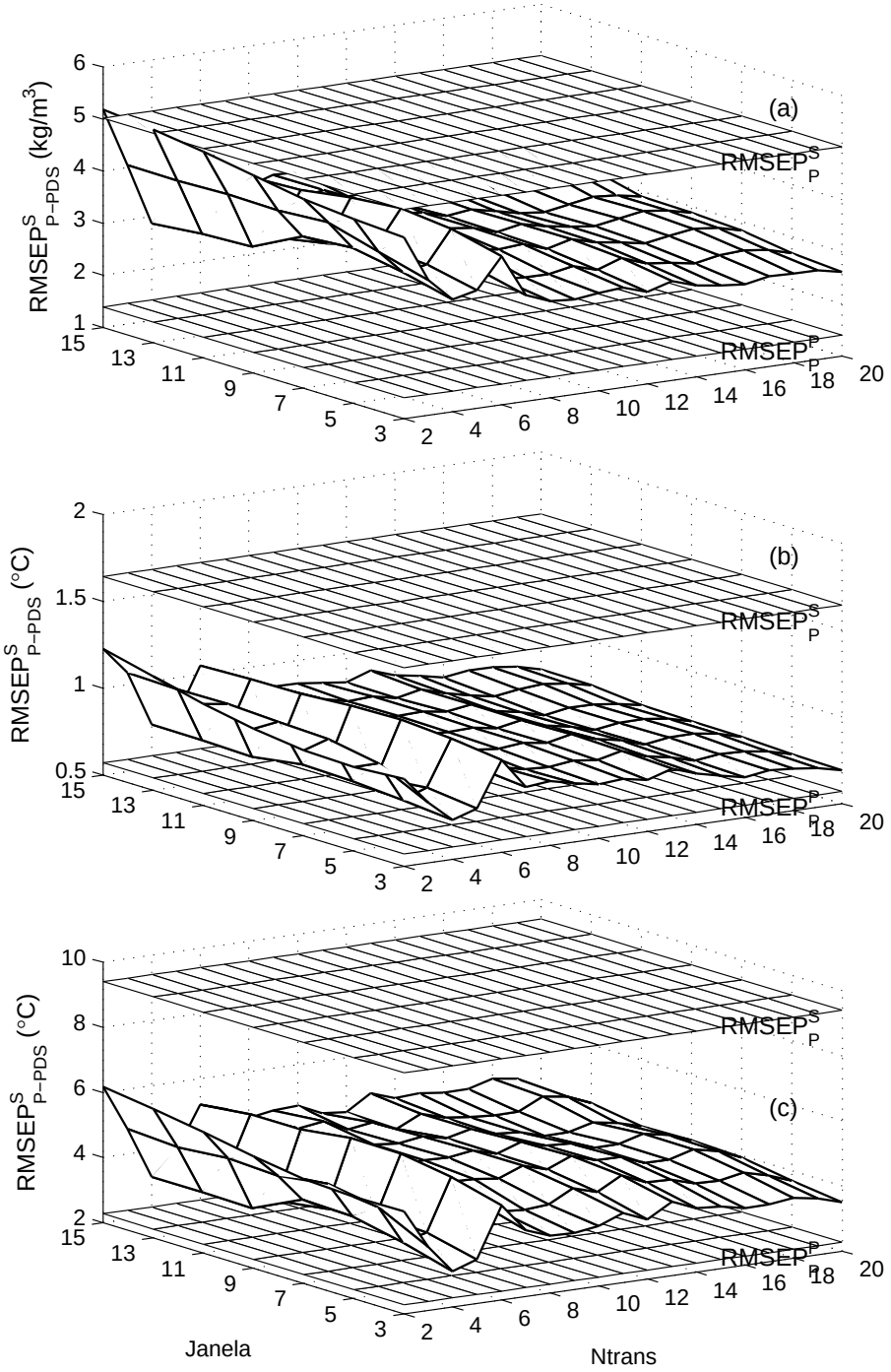


FIGURA 5.4 – RMSEP em função de *Ntrans* obtido com PDS. (a) ME, (b) T10 e (c) T90.

PDS tem uma carga computacional maior e depende do ajuste de um parâmetro adicional, a janela. Apesar dessas desvantagens, o uso do PDS se justifica uma vez que a transferência de calibração é um procedimento realizado apenas esporadicamente em um mesmo instrumento (as ocasiões em que a transferência é necessária são discutidas na

Seção 2.3).

5.2.1.3 WHDS

Como discutido na Seção 2.3.1.3, o WHDS foi implementado usando-se filtros Daubechies1, com um nível de decomposição e janela de 3 pontos para o PDS.

A Figura 5.5 ilustra, para um dos espectros, o resultado das decomposições realizadas pelo algoritmo WHDS. As Figuras 5.5a, 5.5b e 5.5c apresentam o espectro original, o espectro resultante da reconstrução dos coeficientes de aproximação (filtragem passa-baixas) e dos coeficientes de detalhes (filtragem passa-altas), respectivamente. Vale notar a diferença de escalas no eixo das ordenadas.

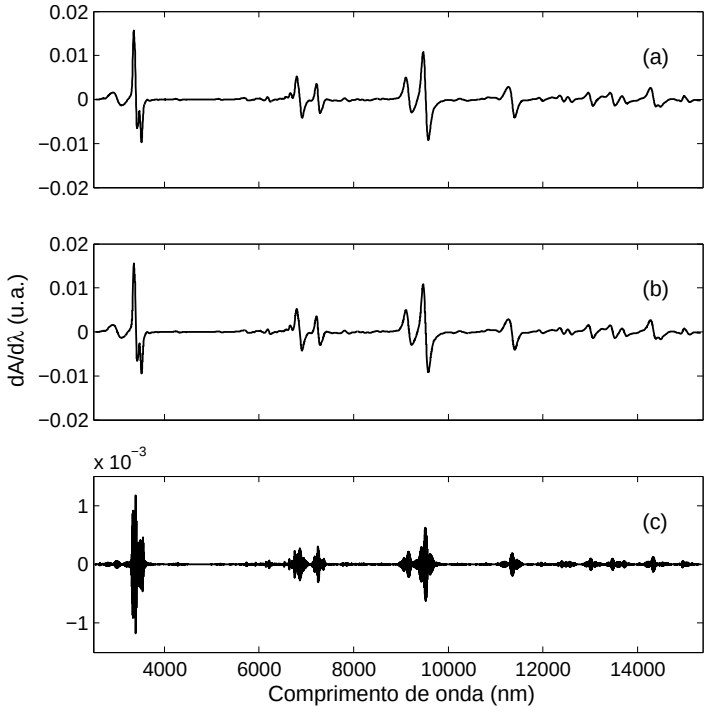


FIGURA 5.5 – Espectro (a) original, (b) resultante da filtragem passa-baixas e (c) resultante da filtragem passa-altas de uma amostra arbitrária.

A Figura 5.6 apresenta o resultado do WHDS em termos do RMSEP obtido no es-

cravo ($\text{RMSEP}^{\text{S}}_{\text{P-WHDS}}$) em função de *Ntrans*. Neste caso, também é possível notar uma tendência de decréscimo do $\text{RMSEP}^{\text{S}}_{\text{P-WHDS}}$ com o aumento de *Ntrans*, sendo observada uma estabilização com, aproximadamente, 7 amostras de transferência. A partir de 2 amostras de transferência o $\text{RMSEP}^{\text{S}}_{\text{P-WHDS}}$ se mostrou sempre menor que o $\text{RMSEP}^{\text{S}}_{\text{P}}$.

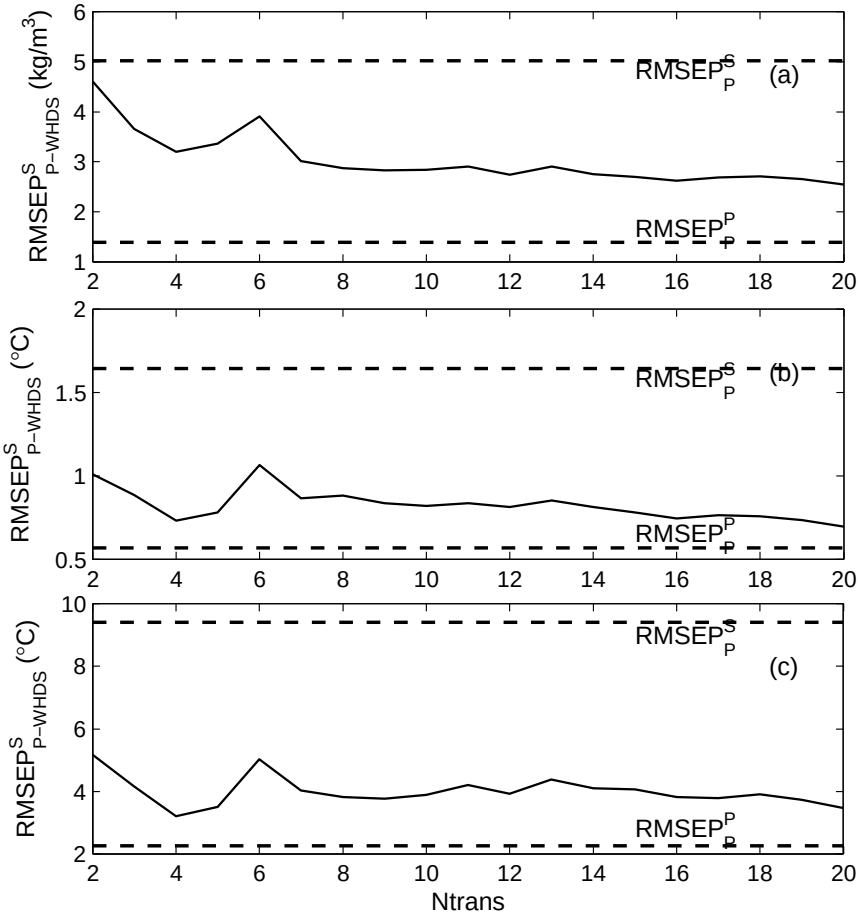


FIGURA 5.6 – RMSEP em função de *Ntrans* obtido com WHDS. (a) ME, (b) T10 e (c) T90.

5.2.1.4 Comparação Entre os Três Métodos de Padronização

Aplicando os algoritmos DS, PDS e WHDS (ver Seção 2.3.1) para a devida transferência de calibração do Primário para o Secundário, pode-se notar que o $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ diminui com o aumento de amostras de transferência. Porém, essa diminuição se torna inexpressiva após um certo número de amostras de transferência. Outra observação interessante

é que o $\text{RMSEP}_{\text{P}}^{\text{S}}$ sempre pode ser melhorado com o uso desses métodos, desde que o número de amostras de transferência seja escolhido adequadamente.

Para fins de comparação, a Tabela 5.3 mostra os valores de $\text{RMSEP}_{\text{P-T}}^{\text{S}}$ obtidos para cada um dos métodos usando-se 9 amostras de transferência, com uma janela de tamanho 3 para o PDS. Esse número de *Ntrans* foi escolhido como um compromisso entre a diminuição do RMSEP para todos as propriedades e baseado no DS, pois foi o método que necessitou do maior números de amostras para atingir um nivelamento na curva do erro. Também é apresentado o valor original obtido sem transferência, bem como a maior redução percentual (Red) do $\text{RMSEP}_{\text{P-T}}^{\text{S}}$ em relação ao $\text{RMSEP}_{\text{P}}^{\text{S}}$ levando em conta os resultados dos três metódos. Tal redução foi calculada como indicado na Equação 5.1 abaixo

$$\text{Red } (\%) = \frac{\text{RMSEP}_{\text{P}}^{\text{S}} - \text{RMSEP}_{\text{P-T}}^{\text{S}}}{\text{RMSEP}_{\text{P}}^{\text{S}}} \times 100\% \tag{5.1}$$

TABELA 5.3 – $\text{RMSEP}_{\text{P-T}}^{\text{S}}$ para 9 amostras de transferência. (PDS com janela de 3 pontos)

Prop	$\text{RMSEP}_{\text{P}}^{\text{S}}$	DS	PDS	WHDS	Red (%)
ME	5.0	3.0	2.8	2.8	44
T10	1.6	0.9	0.8	0.8	50
T90	9.4	4.6	3.7	3.8	61

Como se vê na Tabela 5.3, para $Ntrans = 9$, o uso das técnicas apresentadas melhorou o desempenho no instrumento Secundário. Podem-se observar reduções de 44% a 61% no RMSEP. O PDS e o WHDS apresentaram resultados melhores que o DS. A grande semelhança entre os resultados do PDS e WHDS se deve ao fato de o WHDS usar o PDS no espectro resultante dos coeficientes de aproximação e o DS no espectro resultante dos coeficientes de detalhe. Com efeito, como mostra a Figura 5.5, esse último apresenta uma amplitude muito menor que a do espectro gerado pelos coeficientes de aproximação. Desse modo, em sua maior parte o efeito do WHDS é devido à padronização PDS do espectro

de aproximação, que é similar ao espectro antes da decomposição wavelet. Portanto, o uso do PDS é preferível por ser mais simples que o WHDS. Inspeções visuais que apóiam essa idéia podem ser feitas de comparações entre a Figura 5.6 e figuras da Seção 5.2.2 (p. ex., Figura 5.8).

O uso de técnicas de padronização se mostrou útil, no caso em questão, para a transferência do modelo PLS. Em particular, PDS e WHDS apresentaram melhores resultados que o DS. De certo modo, esse resultado já era esperado dadas as motivações para o uso do PDS (ver Seção 2.3.1.2). As melhoras observadas no RMSEP após a transferência dependem do número de amostras de transferência disponíveis. No estudo em questão, em geral, 9 amostras se mostraram suficientes para a devida transferência.

5.2.2 Métodos de Seleção de Variáveis - Gasolina

Nesta seção serão avaliados os resultados obtidos com as técnicas APSO/W[2] (ver Seção 5.2.2.1) e APSO/W-BAG (ver Seção 5.2.2.2). Essas técnicas buscam a transferência do modelo através do reaproveitamento das medições já realizadas no Primário em conjunto com outras poucas medições realizadas no Secundário. Vale ressaltar que o APSO é uma técnica já proposta na literatura (HONORATO *et al.*, 2005). Este trabalho apresenta duas contribuições principais: o APSW e o uso do *subagging* para transferência dos modelos APSO/W.

Como método clássico de comparação será usado o PDS-PLS com janela de 3 pontos. Tal escolha é baseada nos resultados da seção anterior, em que o PDS mostrou melhores resultados que o DS e resultados similares ao WHDS.

[2]A sigla APSO/W denota as duas técnicas em separado, o APSO e o APSW. Não deve ser confundido com alguma técnica proposta que mescle aspectos do APSO e do APSW.

O modelo APSW foi construído usando-se três níveis de decomposição (*ndec*) na DWT. Tal decisão foi baseada no melhor desempenho médio da transferência para as propriedades da gasolina em estudo. Apesar da escolha de *ndec* ser um fator importante para a construção de modelos APSW, para o APSW-BAG o valor adotado para *ndec* deixa de ser crítico. A Figura 5.7 mostra as curvas de $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ obtidas com o APSW e o APSW-BAG para vários valores de *ndec* para a propriedade T10. Nota-se que as curvas do APSW são sensíveis à escolha de *ndec*, enquanto que o APSW-BAG é menos afetado por tal parâmetro. Apesar de estarem sendo mostradas apenas as curvas para T10, o mesmo comportamento foi observado para as demais propriedades.

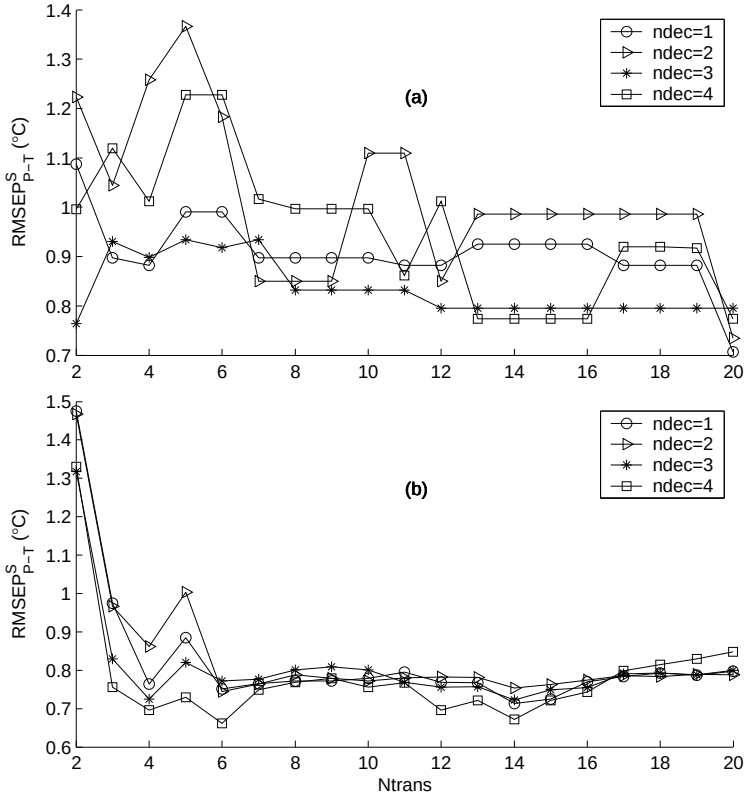


FIGURA 5.7 – Influência do número de decomposições (*ndec*) sobre o $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ de (a) APSW e (b) APSW-BAG em função de *Ntrans* para a propriedade T10.

Primeiramente, serão avaliados os resultados obtidos sem o *subagging*, ou seja APSO/W. Na seção subseqüente será feita a avaliação de APSO/W-BAG.

5.2.2.1 APSO/W

Na Tabela 5.4 são mostrados os valores de RMSEP sem o uso de técnicas de transferência de calibração no instrumento Primário (P) e Secundário (S) para os métodos PLS, APSO e APSW, lembrando que o APSW fez uso de 3 níveis de decomposição. Em termos do resultado observado no Primário, o APSW apresentou desempenho ligeiramente superior às demais técnicas em duas (ME e T90) das três propriedades em análise. Tal fato pode indicar que os modelos de regressão wavelet podem ser de valia para a calibração do instrumento Primário. Considerando o resultado no Secundário, nota-se que APSO/W obtiveram $\text{RMSEP}_{\text{P}}^{\text{S}}$ menor que o PLS e ainda menor variação em relação ao $\text{RMSEP}_{\text{P}}^{\text{S}}$, podendo indicar uma maior robustez de modelos de RLM-APS a variações instrumentais em relação ao PLS. Contudo, à exceção de T10 para o APSO, os resultados no Secundário são piores que os obtidos no Primário, o que justifica o uso de alguma técnica de transferência de calibração.

TABELA 5.4 – $\text{RMSEP}_{\text{P}}^{\text{P}}$ (P) e $\text{RMSEP}_{\text{P}}^{\text{S}}$ (S) para os modelos PLS, APSO e APSW.

Técnica	ME	T10	T90
PLS-P	1.4	0.6	2.3
PLS-S	5.0	1.6	9.4
APSO-P	1.5	0.7	2.5
APSO-S	3.2	0.7	5.3
APSW-P	1.3	0.7	2.1
APSW-S	4.0	1.6	5.1

Vale lembrar que a adaptação do APSO/W como técnica de transferência consiste na modificação do conjunto de validação do APS (ver Seção 5.2.2.1), incluindo as amostras de transferência no mesmo. A Figura 5.8 exibe as curvas de $\text{RMSEP}_{\text{P-T}}^{\text{S}}$ em função de $Ntrans$ para os três métodos de transferência de calibração em estudo, nesta seção. Para o PDS há uma tendência de diminuição do $\text{RMSEP}_{\text{P-T}}^{\text{S}}$ com o aumento de $Ntrans$, havendo um nivelamento em torno de $Ntrans = 7$. Para o APSO e APSW não há uma tendência tão

clara de decrescimento do $RMSEP^{S}_{P\text{-}T}$ com o aumento de *Ntrans*. Entretanto, pode-se notar uma certo nivelamento do $RMSEP^{S}_{P\text{-}T}$ do APSO e APSW com cerca de 6 amostras de transferência. A diferença mais notável pode ser apontada na propriedade T90, em que os desempenhos de APSW e APSO são melhores que o do PDS, com destaque para o APSO.

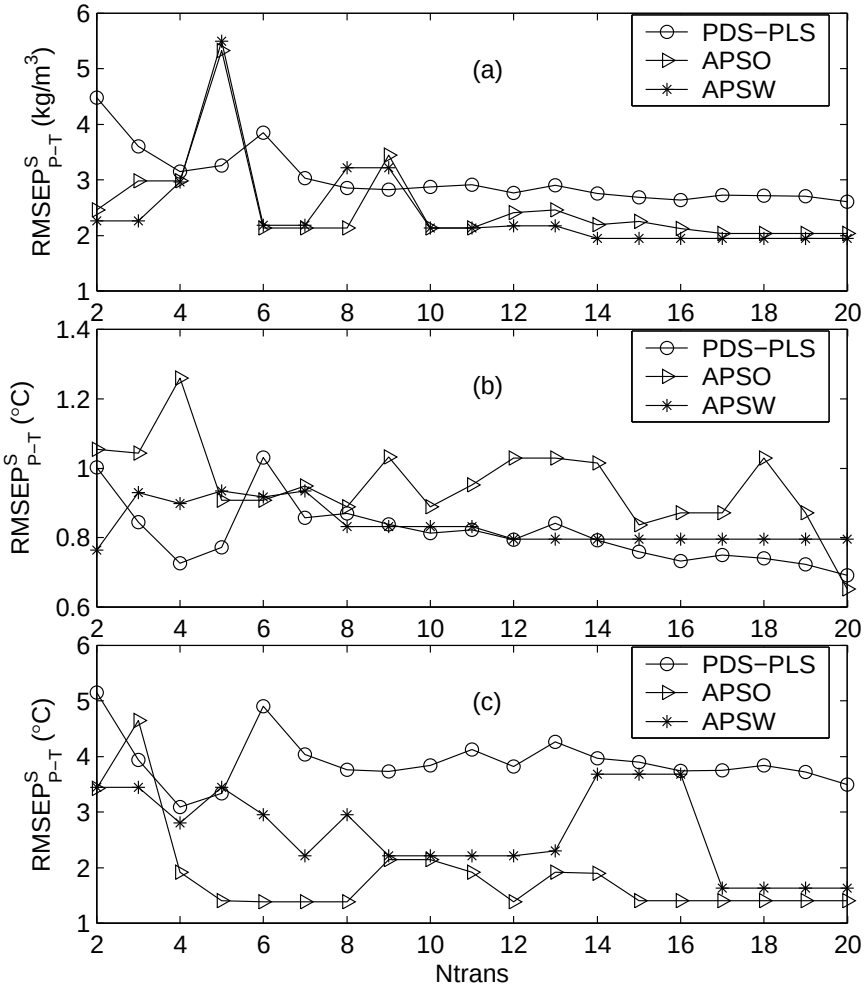


FIGURA 5.8 – $RMSEP^{S}_{P\text{-}T}$ em função de *Ntrans* para PDS-PLS, APSO e APSW. (a) ME, (b) T10 e (c) T90.

Na Tabela 5.5 pode-se ver a comparação do $RMSEP^{S}_{P\text{-}T}$ para PDS-PLS, APSO e APSW fixando-se 7 amostras de transferência[3]. O APSO e o APSW apresentaram melhores resultados que o PDS-PLS em duas (ME e T90) das três propriedades em análise.

[3]Número escolhido com base na análise para o PDS feita na Seção 5.2.1.2

TABELA 5.5 – $RMSEP^{S}_{P\text{-}T}$ para 7 amostras de transferência.

Prop	PDS-PLS	APSO	APSW
ME	3.0	2.1	2.2
T10	0.9	0.9	0.9
T90	4.0	1.4	2.2

Em uma comparação entre a Figura 5.8 e a Tabela 5.4, pode-se notar que as técnicas de transferência fizeram com que o modelo no Secundário melhorasse. Adicionalmente, PDS-PLS e APSW apresentaram o comportamento esperado após a transferência de calibração: melhora do desempenho no Secundário. Por outro lado, pode-se notar um comportamento peculiar no APSO, pois na propriedade T10 a transferência de calibração piorou ligeiramente o resultado obtido no Secundário, enquanto que em T90, a transferência levou a resultados melhores que os do próprio Primário. Esse tipo de comportamento não é muito desejável, uma vez que o resultado da transferência torna-se imprevisível.

5.2.2.2 APSO/W-BAG

A Figura 5.9 ilustra o procedimento de escolha das variáveis para o APSO e APSO-BAG (ambos sem transferência) para a propriedade T90. A Figura 5.9a mostra a seleção de variáveis conduzida pelo APSO (em destaque) sobreposta ao espectro de uma amostra arbitrária. Já a Figura 5.9b apresenta um histograma contendo o número de vezes em que uma determinada variável espectral foi selecionada no APSO-BAG, vale ressaltar que o número máximo possível de vezes é 30, que é o número de iterações do *subagging*. Nota-se que o procedimento de *subagging*, de fato, permite a atribuição de pesos não-nulos a um maior número de variáveis (como argumentado na Seção 5.2.2.2). A saber, o modelo APSO para T90 foi construído usando-se 30 variáveis espectrais, enquanto que o APSO-BAG usou 105 variáveis. É interessante notar que as variáveis escolhidas no APSO foram também usadas no APSO-BAG em alguma das iterações do *subagging*.

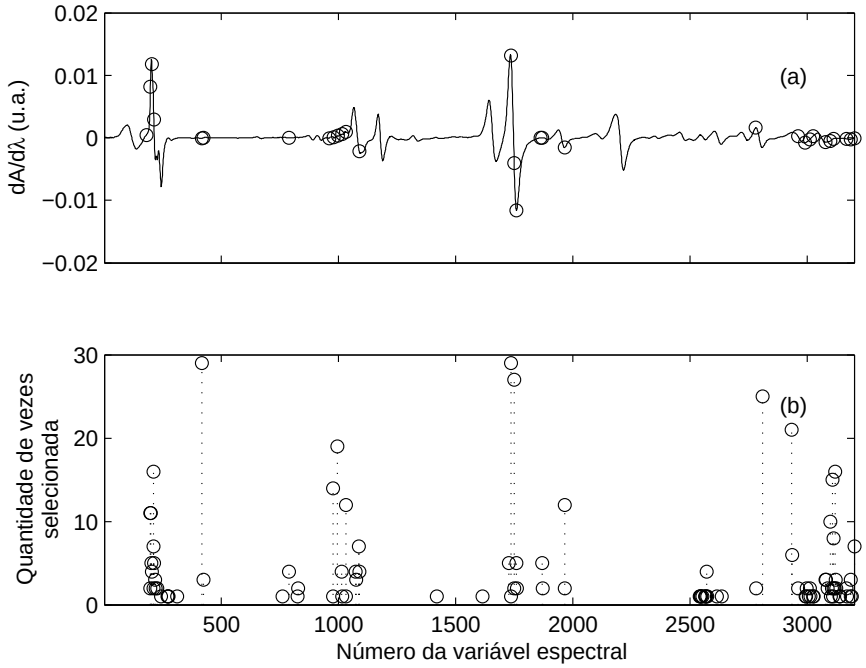


FIGURA 5.9 – Escolha das variáveis para a construção do modelo RLM para T90: (a) APSO e (b) histograma do APSO-BAG.

Na Tabela 5.6 são mostrados os valores de RMSEP sem o uso de técnicas de transferência de calibração no instrumento Primário (P) e Secundário (S) para os métodos PLS, APSO-BAG e APSW-BAG, lembrando que no APSW-BAG uso-se três níveis de decomposição ($ndec = 3$). Em termos do resultado observado no Primário, pode-se notar que os desempenhos das técnicas com *subagging* se mostram iguais ou superiores ao PLS. Tal superioridade também foi apontada para calibração de modelos APSO-BAG em amostras de diesel em (GALVÃO *et al.*, 2006). Considerando o resultado no Secundário, nota-se que em todas as propriedades APSO/W-BAG obtiveram $\text{RMSEP}_{\text{P}}^{\text{S}}$ menor que o PLS e ainda menor variação em relação ao $\text{RMSEP}_{\text{P}}^{\text{S}}$, podendo indicar uma maior robustez a variações instrumentais de modelos de RLM-APS com *subagging* em relação ao PLS. Contudo os resultados no Secundário são invariavelmente piores, o que justifica o uso de alguma técnica de transferência de calibração.

A Figura 5.10 mostra as curvas de $\text{RMSEP}_{\text{P-T}}^{\text{S}}$ em função do número de amostras

TABELA 5.6 – $RMSEP_P^P$ (P) e $RMSEP_P^S$ (S) para os modelos PLS, APSO-BAG e APSW-BAG.

Técnica	ME	T10	T90
PLS-P	1.4	0.6	2.3
PLS-S	5.0	1.6	9.4
APSO-BAG-P	1.0	0.6	1.5
APSO-BAG-S	1.7	1.1	3.2
APSW-BAG-P	1.1	0.6	1.9
APSW-BAG-S	1.8	1.4	2.4

de transferência (*Ntrans*) para as três técnicas de transferência de calibração em estudo nesta seção. Pode-se notar que as técnicas APSO-BAG e APSW-BAG apresentam resultados bastante parecidos. Em comparação com o PDS-PLS, APSO/W-BAG apresentaram resultados significativamente melhores em ME e T90, enquanto que em T10 seus desempenhos foram comparáveis. Outro aspecto positivo apresentado por APSO/W-BAG é a tendência de diminuição sistemática do $RMSEP_{P\text{-}T}^S$ sem as grandes variações com o aumento de *Ntrans* observadas com APSO/W (ver Seção 5.2.2.1).

Na Tabela 5.7 pode-se ver a comparação do $RMSEP_{P\text{-}T}^S$ para PDS-PLS, APSO-BAG e APSW-BAG fixando-se 7 amostras de transferência[4]. O APSO-BAG e o APSW-BAG apresentaram melhores resultados que o PDS-PLS em todas três propriedades em análise.

TABELA 5.7 – $RMSEP_{P\text{-}T}^S$ para 7 amostras de transferência.

Prop	PDS-PLS	APSO-BAG	APSW-BAG
ME	3.0	2.1	2.0
T10	0.9	0.8	0.8
T90	4.0	1.9	2.0

Em uma comparação entre a Figura 5.10 e a Tabela 5.6, pode-se notar que as técnicas de transferência fizeram com que o modelo no Secundário melhorasse. Vale ressaltar que o APSW-BAG já apresentava valores de $RMSEP_P^S$ bem menores que o PLS para as propriedades ME e T90. Não obstante, o uso de amostras de transferência ($Ntrans \geq 10$ para

[4]Número escolhido com base na análise para o PDS feita na Seção 5.2.1.2

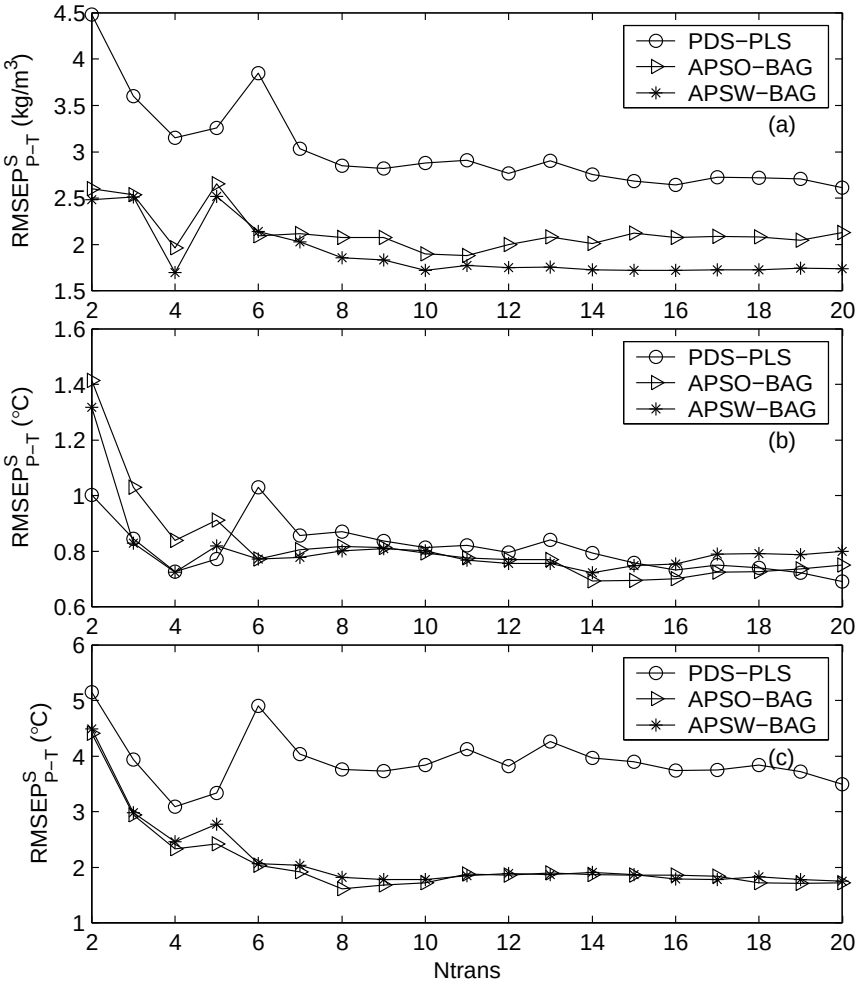


FIGURA 5.10 – RMSEP$^S_{P-T}$ em função de $Ntrans$ para PDS-PLS, APSO-BAG e APSW-BAG. (a) ME, (b) T10 e (c) T90.

ME e $Ntrans \geq 6$ para T90) permitiu uma melhoria na acurácia das previsões realizadas no Secundário. Por outro lado, a transferência não foi bem sucedida no APSO-BAG para a propriedade ME, ou seja, o resultado após a transferência piorou o desempenho obtido no Secundário, independentemente de $Ntrans$. Mas, ainda assim, continuou com resultados melhores que o PDS-PLS.

5.2.2.3 Descarte de Coeficientes Wavelet

Durante a etapa de análise dos resultados, notou-se que poderia ser possível fazer um descarte *a priori* de coeficientes wavelets. Tal fato foi constatado durante a análise

das cadeias de variáveis selecionadas pelo APS. Para a formação das cadeias do APS, foi escolhido, sistematicamente, apenas os coeficientes de aproximação da DWT. Quando a variável inicial no APS foi definida como um coeficiente de detalhe, este foi o único coeficiente de detalhe presente na cadeia de variáveis final. Portanto, o descarte (poda) de coeficientes de detalhe pode ser um grande atrativo computacional para o uso do APSW-BAG em relação ao APSO-BAG.

A Figura 5.11 mostra as curvas de $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ do PDS-PLS, do APSW-BAG e do APSW-BAG-PODA. Este último foi implementado fazendo-se um descarte prévio dos coeficientes de detalhe. Nota-se que o APSW-BAG e o APSW-BAG-PODA apresentam resultados semelhantes. Tal fato indica que o descarte dos coeficientes de detalhe pode ser feito sem detrimento do desempenho da transferência de calibração.

A Tabela 5.8 apresenta o tempo total de processamento envolvido no APSW-BAG, APSW-BAG-PODA e PDS-PLS para o estudo da transferência de apenas uma propriedade da gasolina[5] com sete amostras de transferência. Tal tempo engloba as seguintes etapas no APSW-BAG(-PODA): (1) determinação das cadeias de variáveis, (2) seleção das variáveis mais apropriadas para a construção do modelo e (3) repetição de (2) para todas as 30 iterações do *subagging*. Já o PDS-PLS envolve: (1) padronização dos espectros e (2) calibração do modelo PLS. Como se vê, o APSW-BAG-PODA é cerca de 10 vezes mais rápido que o APSW-BAG e ligeiramente mais rápido que o PDS-PLS. É interessante notar que tal ganho computacional foi obtido, aparentemente, sem perda de acurácia na previsão após a transferência (ver Figura 5.11). Vale ressaltar que APSW-BAG-PODA opera em número oito vezes menor de variáveis que o APSW-BAG, dado que $ndec = 3$ (ver Seção 3.1).

[5]Os dados de gasolina compreendem $m = 3221$ variáveis espectrais e $M = 83$ amostras de modelagem.

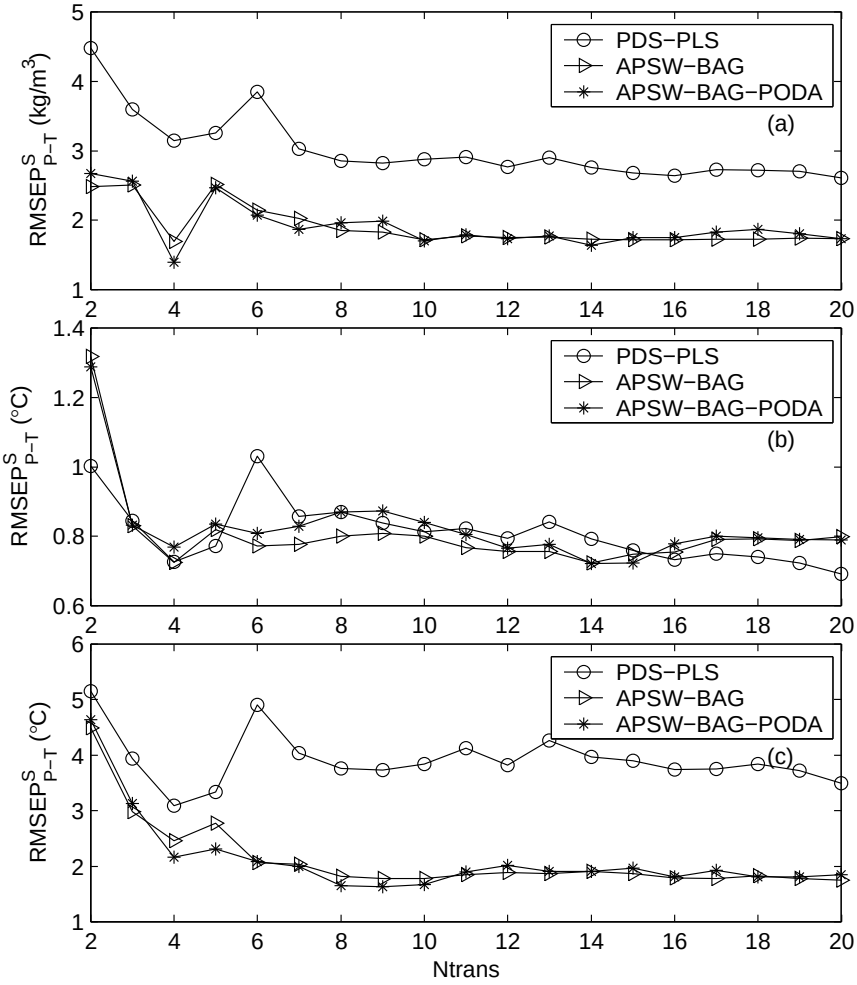


FIGURA 5.11 – $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ em função de $Ntrans$ para PDS-PLS, APSW-BAG e APSW-BAG-PODA. (a) ME, (b) T10 e (c) T90.

TABELA 5.8 – Tempo de processamento do APSW-BAG e APSW-BAG-PODA de uma propriedade para os dados de gasolina.

Técnica	Tempo (min)
APSW-BAG	30.5
APSW-BAG-PODA	3.1
PDS-PLS	4.2

5.2.2.4 Considerações

O uso das técnicas APSO/W e APSO/W-BAG se mostrou útil para transferência de modelos entre distintos instrumentos. Em particular, APSO/W-BAG apresentaram melhores resultados, no geral, tanto no Primário quanto no Secundário.

No APSW, *ndec* é uma escolha importante a ser feita, uma vez que pode levar a resultados bem diferentes (ver Figura 5.7a). Em princípio, o critério aqui adotado para escolha adequada de *ndec* não pode ser usado em aplicações práticas, dado que foi escolhido a partir de uma análise *a posteriori* em que os valores a serem preditos eram conhecidos. No entanto, o estudo realizado é de valia para se ganhar compreensão a respeito dessa técnica de transferência. Tal problema do APSW pode ser contornado usando-se o *subagging*, dado que o APSW-BAG apresenta pouca variação de desempenho entre diferentes valores de *ndec* (ver Figura 5.7b).

O APSO-BAG apresenta desempenho comparável ao APSW-BAG, porém o último pode ser modificado de modo a se reduzir a complexidade computacional através de um descarte preliminar de variáveis sem perda de desempenho na transferência. A preservação dos resultados reside no fato de que a amplitude dos coeficientes de detalhe é muito menor que a amplitude dos coeficientes de aproximação e o APS está sendo implementado sem escalonamento das variáveis. Após o descarte de coeficientes o tempo de processamento pode ser reduzido em cerca de 5.5 vezes, a depender das condições do experimento.

5.3 Dados de Milho

Os dados de milho foram detalhados na Seção 4.2. Vale lembrar que a propriedade em análise é a umidade. A seguir, são apresentados os resultados obtidos na calibração e na sua transferência.

5.3.1 Calibração

A Tabela 5.9 mostra o resumo da etapa de calibração dos modelos para os dados de milho. Nota-se que, em termos de $\text{RMSEP}_{\text{P}}^{\text{P}}$, o PLS e o APSO obtiveram resultado melhor que as demais técnicas. No Secundário, os resultados foram, em geral, parecidos, com pequeno destaque para os métodos com *subagging*. Vale ressaltar que ocorreu uma piora expressiva em todos os métodos quando o modelo do Primário foi aplicado diretamente ao Secundário.

TABELA 5.9 – Resumo da etapa de calibração. Os valores da média, do mínimo e do máximo da umidade nas amostras de predição são, respectivamente, 10.282, 9.430 e 10.882.

Técnica	$\text{RMSEP}_{\text{P}}^{\text{P}}$	$\text{RMSEP}_{\text{P}}^{\text{S}}$
PLS	0.0129 (15)	1.5247
APSO	0.0130	1.4939
APSW	0.0182	1.5733
APSO-BAG	0.0619	1.4690
APSW-BAG	0.0598	1.4730

Em parêntese o número de LV's usadas no PLS.

5.3.2 Transferência de Calibração

O modelo APSW foi construído usando-se quatro níveis de decomposição (*ndec*) na DWT. Tal decisão foi baseada no melhor desempenho médio da transferência do APSW-BAG para a propriedade de milho em estudo (umidade). A Figura 5.12 mostra as curvas

de $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ obtidas com o APSW e o APSW-BAG para vários valores de *ndec*. Como no caso da gasolina, as curvas do APSW-BAG tendem a seguir uma mesma tendência e se aproximarem de um determinado valor, indicando uma robustez a variações em *ndec*.

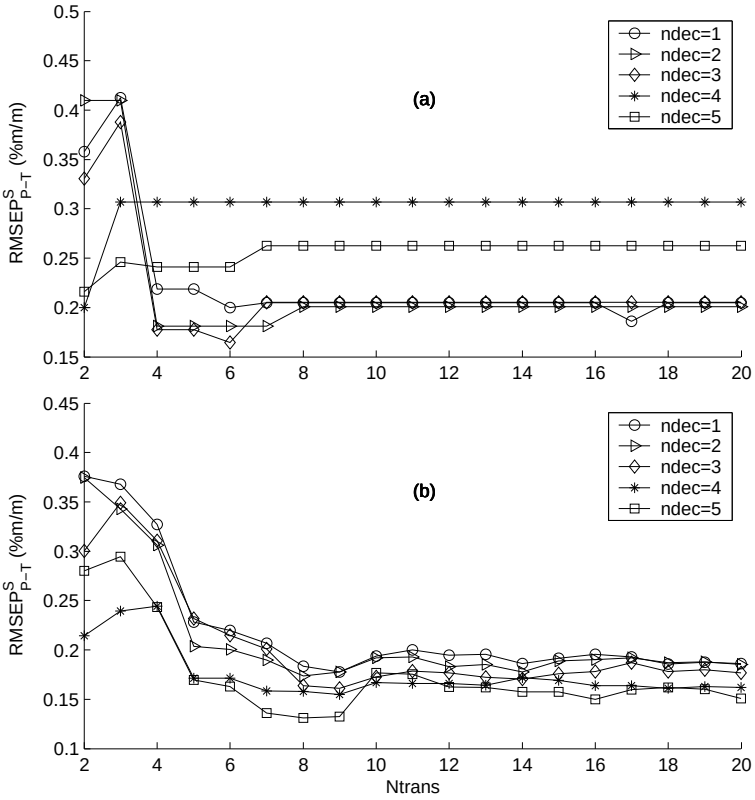


FIGURA 5.12 – Influência do número de decomposições (*ndec*) sobre o $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ de (a) APSW e (b) APSW-BAG em função de *Ntrans* para a umidade do milho.

A Figura 5.13 mostra as curvas de $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ para os métodos em análise nesta seção. Nota-se que, para todas as técnicas em comparação, a transferência de calibração melhorou o resultado obtido no Secundário. Em geral, as técnicas analisadas obtiveram $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ menor que o PDS-PLS. Destaque pode ser dado para o APSW-BAG, que obteve o melhor desempenho para todos os valores de *Ntrans*, à exceção de $Ntrans = 2$ em que o APSW foi um pouco melhor.

Vale ressaltar que, também para os dados de milho, nota-se um comportamento de diminuição sistemática do $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ com o aumento de *Ntrans* em APSO/W-BAG. Outra observação interessante é a semelhança entre as curvas de APSO-BAG e do APSW-

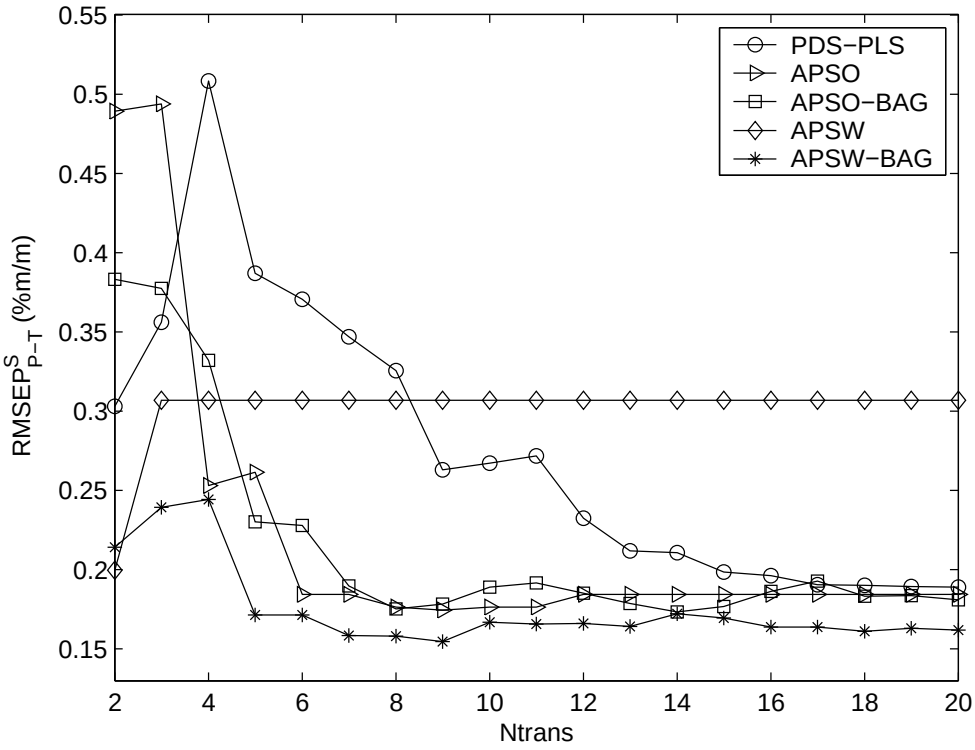


FIGURA 5.13 – Curvas de $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ para PDS-PLS, APSO, APSW, APSO-BAG e APSW-BAG para a umidade do milho.

BAG para $ndec = 1$ (ver Figuras 5.12 e 5.13).

5.3.2.1 Descarte de Coeficientes Wavelet

Nos dados de milho, notou-se que as cadeias de variáveis no APS eram formadas com coeficientes de aproximação até $ndec = 3$, porém quando $ndec = 4$ os coeficientes de detalhe do último nível de decomposição passaram a fazer parte de tais cadeias. Portanto, foram preservados os coeficientes de aproximação e os de detalhe do último nível na implementação do APSW-BAG-PODA. Ao final, o número de variáveis foi reduzido em oito vezes.

A Figura 5.14 mostra as curvas de $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ do PDS-PLS, do APSW-BAG e do APSW-BAG-PODA. Este último foi implementado fazendo-se um descarte prévio de co-

eficientes de detalhe como descrito acima. Nota-se que o APSW-BAG e o APSW-BAG-PODA apresentam resultados semelhantes. Tal fato indica que o descarte de coeficientes de detalhe pode ser feito sem detrimento do desempenho da transferência de calibração.

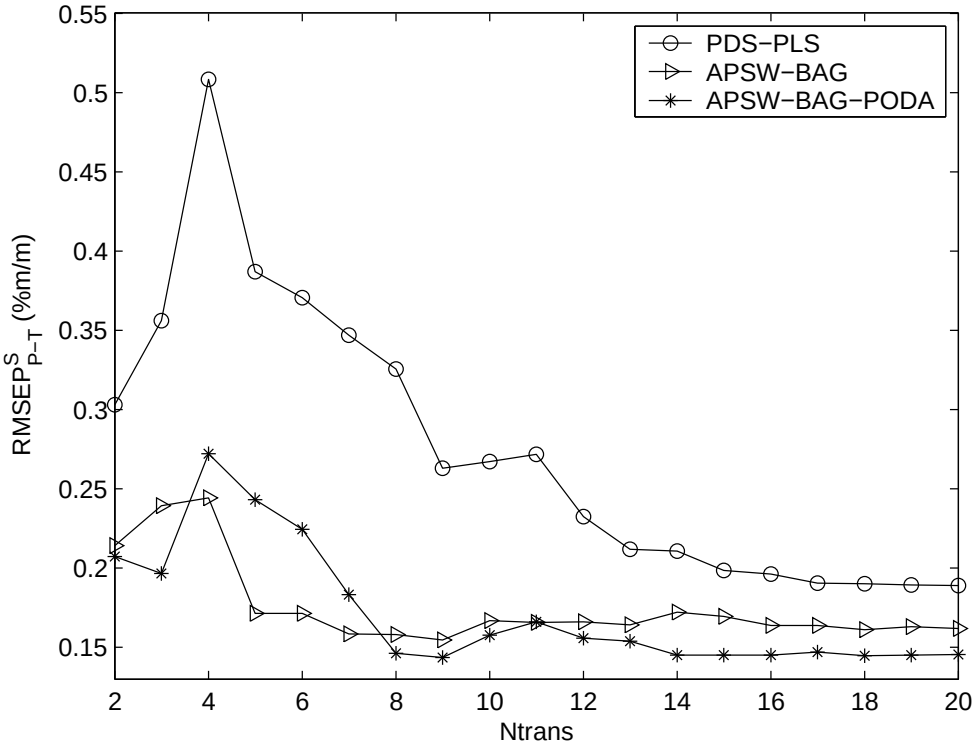


FIGURA 5.14 – $\text{RMSEP}^{\text{S}}_{\text{P-T}}$ em função de $Ntrans$ para PDS-PLS, APSW-BAG e APSW-BAG-PODA.

A Tabela 5.10 apresenta o tempo total de processamento envolvido no APSW-BAG, APSW-BAG-PODA e do PDS-PLS (para fins de comparação) para o estudo da transferência de apenas uma propriedade do milho[6]. Tal tempo engloba as mesmas etapas descritas para os dados de gasolina (ver Seção 5.2.2.3). Como se vê, o APSW-BAG-PODA é cerca de 2.5 vezes mais rápido que o APSW-BAG, contudo cerca de 20 vezes mais lento que o PDS-PLS. Interessante notar que tal ganho computacional foi obtido com uma pequena perda de acurácia na previsão após a transferência para $Ntrans = 5$ e 6 e sem perda para os demais valores de $Ntrans$ (ver Figura 5.14).

[6]Os dados de gasolina compreendem $m = 700$ variáveis espectrais e $M = 60$ amostras de modelagem.

TABELA 5.10 – Tempo de processamento do APSW-BAG e APSW-BAG-PODA para os dados de milho.

Técnica	Tempo (seg)
APSW-BAG	188.3
APSW-BAG-PODA	63.7
PDS-PLS	2.8

5.3.3 Considerações

O uso dos dados de milho teve como objetivo validar as técnicas propostas (APSO/W-BAG). Por essa razão os resultados foram apresentados de maneira concisa. Como pôde-se notar, o uso do APSO/W-BAG como técnica de transferência também se mostrou de valia para este conjunto de dados. O descarte preliminar de variáveis no APSW-BAG também se mostrou útil.

6 Conclusão

Neste trabalho foi estudado o problema de transferência de calibração entre dois instrumentos para análise da qualidade de gasolina por espectroscopia no infravermelho médio. Adicionalmente, para fins de validação das técnicas propostas, foi estudada a transferência de modelos de espectroscopia no infravermelho próximo de amostras de milho.

Não foi objetivo deste trabalho desenvolver um aparelho que possa ser usado diretamente nos postos de gasolina, nem estudar a validade das técnicas propostas como novos padrões para aferição da qualidade da gasolina. Contudo, espera-se que as técnicas aqui apresentadas possam servir para o amadurecimento da área da espectrometria conhecida como transferência de calibração que traz consigo grande apelo econômico e ambiental.

A seguir, são resumidas as contribuições deste trabalho, apresentando-se ainda as conclusões gerais, auto-crítica e sugestões para trabalhos futuros.

6.1 Contribuições

- Estudo comparativo entre técnicas de padronização da resposta espectral (DS, PDS e WHDS) para transferência de modelos PLS. Em particular, a técnica PDS apresentou melhores resultados.

- Estudo das possíveis vantagens advindas da seleção de variáveis no domínio wavelet para a construção de modelos de regressão e transferência de calibração. Notou-se que o descarte de coeficientes de detalhes da transformada wavelet aumenta a velocidade de processamento do APSW-BAG em cerca de 2.5 a 10 vezes (a depender do número de variáveis) e o desempenho da transferência não se deteriora.

- Por fim, a principal contribuição deste trabalho foi a proposição e estudos iniciais do uso de uma técnica de reamostragem[1] e combinação de modelos conhecida como *subagging* para fins de transferência de calibração entre instrumentos distintos. Tal técnica apresentou melhores resultados que a técnica clássica PDS-PLS nos dois conjuntos de dados estudados.

6.2 Conclusões Gerais e Auto-crítica

Técnicas de Padronização

As técnicas de padronização DS, PDS e WHDS foram usadas para a transferência de modelos PLS entre dois instrumentos distintos (Primário e Secundário) para análise de propriedades da gasolina. Técnicas de padronização fazem, obrigatoriamente, uso de amostras de transferência, que devem ser medidas em ambos os instrumentos. Em geral, o erro observado no Secundário tende a diminuir com o aumento do número de amostras de transferência ($Ntrans$). Porém, tal diminuição tende a cessar após um determinado valor de $Ntrans$. Em média, o valor observado para essa estabilização do erro no Secundário se deu 9, 7 e 7 amostras de transferência para os métodos DS, PDS e WHDS, respectivamente.

[1] A rigor é realizada uma subamostragem do conjunto de dados disponível.

Este estudo de caso revelou que tais técnicas podem ser de valia no monitoramento de parâmetros de qualidade da gasolina por espectroscopia no Infravermelho. Com efeito, o estudo mostra que um modelo desenvolvido para uso em um dado espectrômetro pode conduzir a erros até quatro vezes maiores quando aplicado a outro instrumento de mesmo modelo e fabricante. Métodos de padronização permitiram reduzir o erro no segundo instrumento por até 60%.

Os resultados obtidos com PDS e WHDS foram superiores aos obtidos com DS. PDS e WHDS tiveram desempenho bastante parecidos, porém com uma pequena vantagem para o PDS. A grande semelhança entre os resultados do PDS e WHDS se deve ao fato de o WHDS usar o PDS no espectro resultante dos coeficientes de aproximação e o DS no espectro resultante dos coeficientes de detalhe. Com efeito, como mostra a Figura 5.5, esse último apresenta uma amplitude muito menor que a do espectro gerado pelos coeficientes de aproximação. Desse modo, em sua maior parte o efeito do WHDS é devido à padronização PDS do espectro de aproximação, que é similar ao espectro antes da decomposição wavelet.

Uma dificuldade no PDS é se estabelecer o tamanho apropriado na janela usada durante a padronização. No entanto, no caso em estudo, notou-se que a partir de $Ntrans = 7$ esse escolha deixa de ser crítica. Notou-se também que o uso de janela de três pontos apresenta resultados satisfatórios na maioria dos casos.

Outra desvantagem do PDS em relação ao DS é seu maior esforço computacional. Entretanto, ainda assim, o uso do PDS em lugar do DS se justifica uma vez que a transferência de calibração é um procedimento realizado esporadicamente em um mesmo instrumento.

Com relação ao WHDS, o uso dessa técnica não apresentou vantagens com relação ao uso do PDS. Adicionalmente, o WHDS ainda possui maior complexidade computacional

que o PDS, dado que os espectros ainda são processados com a transformada wavelet. Por outro lado, o algoritmo do WHDS usado adotou os parâmetros de projeto sugeridos em (TAN; BROWN, 2001), mas é possível que diferentes parâmetros da DWT (níveis de decomposição e filtro wavelet usado) levassem a resultados diferentes, melhores ou piores. O ajuste de parâmetros do WHDS não foi estudado neste trabalho.

Técnicas de Seleção de Variáveis

A técnica de seleção de variáveis para transferência de calibração proposta em (HONORATO *et al.*, 2005), aqui denominada de APSO foi comparada com o uso do PDS em conjunto com o PLS. Adicionalmente, outras três técnicas aqui propostas foram acrescentadas à comparação. A primeira, o APSW opera de modo semelhante ao APSO, porém as variáveis são selecionadas no domínio wavelet, enquanto o APSO seleciona no domínio original de comprimento de onda. Com o intuito de se melhorar o desempenho obtido no Secundário, foi usado o *subagging* em conjunto com o APSO e APSW (APSO/W). Essas técnicas, propostas aqui, são designadas como APSO-BAG e APSW-BAG (APSO/W-BAG).

O Algoritmo das Projeções Sucessivas (APS) (ARAÚJO *et al.*, 2001) foi usado para que as variáveis selecionadas apresentem simultaneamente boa capacidade preditiva, baixa multicolinearidade e robustez a variações instrumentais. Com efeito, as quatro técnicas (APSO/W e APSO/W-BAG) apresentaram menor erro e menor variação relativa no Secundário que o PLS quando nenhuma amostra de transferência foi usada, possivelmente indicando uma certa robustez desses métodos a variações instrumentais/ambientais.

O uso das cinco técnicas para transferência do modelo se mostrou bem sucedido, no

sentido de ter diminuído o erro observado no Secundário. Comparando APSO/W com PDS-PLS nota-se que na maioria dos casos (propriedades e $Ntrans$) APSO/W apresentaram desempenho superior no Secundário. Entretanto, para APSO/W não há uma tendência muito clara de diminuição do erro no Secundário com o aumento de $Ntrans$. Adicionalmente, observa-se um comportamento instável das curvas de erro em função de $Ntrans$, provavelmente decorrente da decisão "dura" de incluir uma variável ou não durante a etapa de validação no APS. Em particular, APSO e APSW apresentaram desempenhos comparáveis para as propriedades estudadas, não havendo vantagens claras de uma abordagem sobre outra.

Verificou-se que o uso do esquema de reamostragem e combinação de modelos conhecido como *subagging* propicia uma melhora expressiva na capacidade preditiva dos modelos obtidos por seleção de variáveis. APSO/W-BAG se mostraram superiores ao PDS-PLS e APSO/W no estudo realizado. Adicionalmente, o uso do *subagging* em APSO/W tornou o comportamento da transferência mais previsível e "bem comportado", no sentido de que pode-se observar uma tendência de diminuição sistemática do erro no Secundário e sem grandes variações entre valores consecutivos de $Ntrans$. APSO/W-BAG também se mostraram úteis para calibração do modelo no Primário, obtendo menores erros que as demais técnicas (PLS e APSO/W), nos dados de gasolina. Em particular, APSO-BAG e APSW-BAG apresentaram desempenhos comparáveis para as propriedades estudadas, não havendo vantagens claras de uma abordagem sobre outra.

O uso do *subagging* também se mostrou de valia para reduzir a sensibilidade dos resultados com relação à escolha do número de níveis de decomposição wavelet. Adicionalmente, pode-se argumentar que as técnicas APSW e APSW-BAG são casos gerais de APSO e APSO-BAG, estes últimos seriam obtidos quando $ndec = 0$. Tal argumento é

ilustrado nas curvas de erro do APSW-BAG para diferentes níveis de decomposição (Figuras 5.7 e 5.12) em que o APSO-BAG tem comportamento semelhante ao APSW-BAG com $ndec = 1$.

Algumas dificuldades das técnicas que utilizam o *subagging* residem na escolha dos parâmetros utilizados nas etapas de subamostragem e combinação, a saber, a proporção da divisão entre conjunto de calibração e de validação dos dados disponíveis para modelagem e o número total de iterações do algoritmo do *subagging*. Outro ponto frágil do método é sua grande carga computacional em comparação com as demais técnicas aqui estudadas. Entretanto, a carga computacional pode ser substancialmente diminuída se for utilizado um esquema de descarte de coeficientes wavelet de detalhe no APSW-BAG. Tal diminuição foi observada sem degeneração do desempenho na transferência de calibração.

Com o intuito de validar as técnicas propostas em outro conjunto de dados, usou-se ainda espectros de milho adquiridos na região do infravermelho próximo. No que se refere à etapa de transferência de calibração, os comentários feitos para os dados de gasolina se aplicam, também, aos dados de milho.

Vale ressaltar que a abordagem de seleção de variáveis adotada não requer que as amostras de transferência sejam medidas nos dois instrumentos, ao contrário da técnica clássica PDS-PLS. Tal característica pode ser de valia quando os instrumentos primário e secundário não estão instalados no mesmo local.

6.3 Trabalhos Futuros

- Um fato que pode ser analisado é a influência dos pesos na função custo do APSO e APSW, neste trabalho definiu-se pesos iguais aos erros no conjunto de validação

e de transferência.

- Seria de valia estudar o efeito do uso do APS em pacotes wavelets (VETTERLI; KOVACEVIC, 1995). Características robustas poderiam ser melhor identificadas em um particionamento mais completo do plano tempo-escala da transformada wavelet.

- Um outro aspecto é a escolha do número de decomposições adequado para a devida transferência do modelo. O efeito do número de decomposições é perceptível quando poucas amostras de transferência estão disponíveis. O uso de alguma técnica de otimização seria de valia para situações como essa. Vale ressaltar que o APSO e APSO-BAG poderiam ser incluídos no procedimento de otimização considerando-os como o caso em que o número de decomposições é zero.

- Futuros trabalhos poderiam estudar metodologias para a escolha da melhor wavelet a ser empregada em função das características dos espectros em questão. Tal abordagem apresentou bons resultados no estudo de calibração de modelos espectrométricos em (COELHO, 2002) e em um estudo sobre compressão de sinais eletrocardiográficos em (SANTOS; MARTINS; GALVÃO, 2006). Vale ressaltar que o descarte de coeficientes proposto nesse trabalho pode ser visto como uma forma de compressão de dados.

- Outro aspecto interessante a ser estudado seria a influência do particionamento dos conjuntos de calibração e validação no *subagging*. Como mostrado em (BÜHLMANN; YU, 2002), a curva de erro em função da proporção da divisão tende a apresentar um mínimo. Porém, a localização de tal mínimo dependente da natureza dos dados. Neste trabalho estudou-se apenas uma proporção.

- Poderia-se estudar, ainda, o uso de outras técnicas de reamostragem e combinação de

modelos fazendo-se as devidas adaptações para uso em transferência de calibração. Entre essas técnicas pode-se citar o *bagging* (BREIMAN, 1996a), o *iterated bagging* (BREIMAN, 2001) e o *adaboosting* (FREUND; SHAPIRE, 1996).

Referências Bibliográficas

ALSBERG, B. K.; WOODWARD, A. M.; KELL, D. B. An introduction to wavelet transforms for chemometricians: A time-frequency approach. **Chemometrics and Intelligent Laboratory Systems**, v. 37, p. 215–239, 1997.

ANDREW, A.; FEARN, T. Transfer by orthogonal projection: Making near-infrared calibrations robust to between-instrument variation. **Chemometrics and Intelligent Laboratory Systems**, v. 72, p. 51–56, 2004.

ANP. **Regulamento Técnico Nº 2**. [S.l.], 2005.

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.

BAUER, E.; KOHAVI, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. **Machine Learning**, v. 36, p. 105–139, 1999.

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

BLANK, T. B.; SUM, S. T.; BROWN, S. D. Transfer of near-infrared multivariate calibrations without standards. **Analytical Chemistry**, v. 68, p. 2987–2995, 1996.

BOHACS, G.; OVADI, Z.; SALGO, A. Prediction of gasoline properties with near infrared spectroscopy. **Journal of Near Infrared Spectroscopy**, v. 6, p. 341–348, 1998.

BOUVERESSE, E.; HARTMANN, C.; MASSART, D. L.; IR, I. R. L.; PREBBLE, K. A. Standardization of near-infrared spectrometric instruments. **Analytical Chemistry**, v. 68, p. 982–990, 1996.

BOUVERESSE, E.; MASSART, D. L. Standardisation of near-infrared spectrometric instruments: A review. **Vibrational Spectroscopy**, v. 11, p. 3–15, 1996.

BOUVERESSE, E.; MASSART, D. L.; DARDENNE, P. Modified algorithm for standardization of near-infrared spectrometric instruments. **Analytical Chemistry**, v. 67, p. 1381–1389, 1995.

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.

BREIMAN, L. Using iterated bagging to debias regressions. **Machine Learning**, v. 45, p. 261–277, 2001.

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

BÜRCK, J.; WIEGAND, G.; ROTH, H. M. S.; KRÄMER, K. Monitoring of technical oils in supercritical CO2 under continuous flow conditions by NIR spectroscopy and multivariate calibration. **Talanta**, v. 68, p. 1497–1504, 2006.

COELHO, C. J. **Calibração Multivariada Empregando Transformada Wavelet Adaptativa**. Tese (Doutorado) — Instituto Tecnológico de Aeronáutica, 2002.

DANTAS FILHO, H. A. **Um Método Para Determinação Simultânea de Parâmetros de Controle de Qualidade de Óleo Diesel Usando Espectroscopia NIR, Seleção de Variáveis e Calibração Multivariada**. Dissertação (Mestrado) — Universidade Federal da Paraíba, 2003.

DAUBECHIES, I. **Ten Lectures on Wavelets**. Philadelphia: SIAM, 1992.

DAVIES, R. J.; RUSZNAK, C.; DEVALIA, J. L. Why is allergy increasing? environmental factors. **Clinical and Experimental Allergy**, v. 28, n. 6, p. 8–14, 1998.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap Methods and their Application**. Cambridge, UK: Cambridge University Press, 1997.

DESPAGNE, F.; MASSART, D. L.; JANSEN, M.; DAALEN, H. V. Intersite transfer of industrial calibration models. **Analytica Chimica Acta**, v. 406, p. 233–245, 2000.

DOCKERY, D. W.; SPEIZER, F. E.; STRAM, D. O.; WARE, J. H.; SPENGLER, J. D. Effects of inhalable particles on respiratory health of children. **American Review of Respiratory Disease**, v. 139, p. 587–594, 1989.

DREASSI, E.; CERAMELLI, G.; PERRUCCIO, P. L.; CORTI, P. Transfer of calibration in near-infrared reflectance spectrometry. **Analyst**, v. 123, p. 1259–1264, 1998.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2nd. ed. New York: Wiley, 2001.

EFRON, B. **The Jackknife, the Bootstrap and Other Resampling Plans**. Philadelphia: SIAM, 1982.

FEARN, T. Standardisation and calibration transfer for near infrared instruments: A review. **Journal of Near Infrared Spectroscopy**, v. 9, p. 229–244, 2001.

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, p. 724–731, 1999.

FEUDALE, R. N.; WOODY, N. A.; TAN, H.-W.; MYLES, A. J.; BROWN, S. D.; FERRÉ, J. Transfer of multivariate calibration models: A review. **Chemometrics and Intelligent Laboratory Systems**, v. 64, p. 181–192, 2002.

FREUND, Y.; SHAPIRE, R. Experiments with a new boosting algorithm. **Machine Learning: Proceedings of the XIII International Conference**, p. 148–156, 1996.

GALVÃO, R. K. H.; ARAÚJO, M. C. U.; MARTINS, M. 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**, v. 81, p. 60–67, 2006.

GELADI, P.; MACDOUGALL, D.; MARTENS, H. Linearization and scatter-correction for near-infrared reflectance spectra of meat. **Applied Spectroscopy**, v. 39, p. 491–500, 1985.

GOOD, P. I. **Resampling Methods: A Practical Guide to Data Analysis**. Boston: Birkhauser, 1999.

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, p. 1193–1202, 1988.

HANSEN, L. K.; SALAMON, P. Neural network ensembles. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 12, p. 993–1001, 1990.

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

KALIVAS, J. H.; ROBERTS, N.; SUTTER, J. M. Global optimization by simulated annealing with wavelength selection for ultraviolet visible spectrophotometry. **Analytical Chemistry**, v. 61, p. 2024–2030, 1989.

KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. **Technometrics**, v. 11, p. 137–148, 1969.

KRZANOWSKI, W. J. Cross-validation in principal component analysis. **Biometrics**, v. 43, p. 575–584, 1987.

LI, B.; WANG, D.; LV, J.; ZHANG, Z. Flow-injection chemiluminescence simultaneous determination of cobalt(II) and copper(II) using partial least squares calibration. **Talanta**, v. 69, p. 160–165, 2006.

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

LIU, Y.; BROWN, S. D. Wavelet multiscale regression from the perspective of data fusion: New conceptual approaches. **Analytical and Bioanalytical Chemistry**, v. 380, p. 445–452, 2004.

MARK, H.; JR., J. W. A new approach to generating transferable calibrations for quantitative near-infrared spectroscopy. **Spectroscopy**, v. 3, n. 11, p. 28–36, 1988.

MARTENS, H.; NAES, T. **Multivariate Calibration**. New York: John Wiley, 1989.

MARTINS, M. N.; GALVÃO, R. K. H.; PIMENTEL, M. F. Transferência de calibração de modelos de regressão wavelet empregando subagging para análise da qualidade de gasolina por espectroscopia no infravermelho. **(Submetido ao XVI Congresso Brasileiro de Automática)**, 2006.

MARTINS, M. N.; HONORATO, F. A.; GALVÃO, R. K. H.; PIMENTEL, M. F. Avaliação de técnicas de transferência de calibração para monitoração de parâmetros de qualidade de gasolina empregando espectroscopia no infravermelho. **(Submetido ao XVI Congresso Brasileiro de Engenharia Química)**, 2006.

MARTINS, M. N.; HONORATO, F. A.; GALVÃO, R. K. H.; PIMENTEL, M. F. Seleção de variáveis no domínio wavelet para determinação espectroscópica de parâmetros de qualidade de gasolina com robustez a variações entre instrumentos. **(Submetido ao XVI Congresso Brasileiro de Engenharia Química)**, 2006.

MOROS, J.; INON, F. A.; GARRIGUES, S.; de la Guardia, M. Determination of the energetic value of fruit and milk-based beverages through partial-least-squares attenuated total reflectance-fourier transform infrared spectrometry. **Analytica Chimica Acta**, v. 538, p. 181–193, 2005.

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

OPITZ, D. W.; MACLIN, R. J. Popular ensemble methods: An empirical study. **Journal of Artificial Intelligence Research**, v. 11, p. 169–198, 1999.

OZDEMIR, D.; MOSLEY, M.; WILLIAMS, R. Hybrid calibration models: An alternative to calibration transfer. **Applied Spectroscopy**, v. 52, n. 4, p. 599–603, 1998.

PAPOULIS, A. **Probability, Random Variables and Stochastic Processes**. 3rd. ed. New York: McGraw Hill, 1991.

PARK, K.-S.; KO, Y.-H.; H., L.; JUN, C.-H.; CHUNG, H.; KU, M.-S. Near-infrared spectral data transfer using independent standardization samples: A case study on the trans-alkylation process. **Chemometrics and Intelligent Laboratory Systems**, v. 55, p. 53–65, 2001.

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

PIERNA, J. A. F.; MASSART, D. L.; NOORD, O. E. de; RICOUX, P. Direct orthogonalization: Some case studies. **Chemometrics and Intelligent Laboratory Systems**, v. 55, p. 101–108, 2001.

POPE, C. A.; BURNETT, R. T.; THUN, M. J.; CALLE, E. E.; KREWSKI, D.; ITO, K.; THURSTON, G. D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. **Journal of the American Medical Association**, v. 287, p. 1132–1142, 2002.

SANTOS, R. N. F.; MARTINS, M. N.; GALVÃO, R. K. H. Compressão de sinais de ECG usando wavelets adaptativas construídas via programação linear semi-infinita. **(Submetido ao XX Congresso Brasileiro de Engenharia Biomédica)**, 2006.

SASAKI, K.; KAWATA, S.; MINAMI, S. Optimal wavelength selection for quantitative-analysis. **Applied Spectroscopy**, v. 40, p. 185–190, 1986.

SEKULIC, S.; SEASHOLTZ, M. B.; Y., W. Z.; R., K. B.; E., L. S.; HOLT, B. R. Nonlinear multivariate calibration methods in analytical-chemistry. **Analytical Chemistry**, v. 65, p. A835–A845, 1993.

SHENK, J. S.; WESTERHAUS, M. O.; TEMPLETON, W. C. Calibration transfer between near-infrared reflectance spectrophotometers. **Crop Science**, v. 25, n. 1, p. 159–161, 1985.

SKOOG, D. A.; HOLLER, F. J.; NIEMAN, T. A. **Principles of Instrumental Analysis**. 5th. ed. South Melbourne: Thomson Learning, 1998.

SKURICHINA, M.; DUIN, R. P. W. Bagging for linear classifiers. **Pattern Recognition**, v. 31, p. 909–930, 1998.

SMITH, B. M.; GEMPERLINE, P. J. Bootstrap methods for assessing the performance of near-infrared pattern classification. **Journal of Chemometrics**, v. 16, p. 241–246, 2002.

STRANG, G.; NGUYEN, T. **Wavelets and Filter Banks**. Wellesley: Wellesley-Cambridge, 1996.

SVENSSON, O.; KOURTI, T.; MACGREGOR, J. F. An investigation of orthogonal signal correction algorithms and their characteristics. **Journal of Chemometrics**, v. 16, p. 176–188, 2002.

SWIERENGA, H.; GROOT, P. J. de; WEIJER, A. P. de; DERKSEN, M. W. J.; BUYDENS, L. M. C. Improvement of PLS model transferability by robust wavelength selection. **Chemometrics and Intelligent Laboratory Systems**, v. 41, p. 237–248, 1998.

SWIERENGA, H.; HAANSTRA, W. G.; WEIJER, A. P. de; BUYDENS, L. M. C. Comparison of two different approaches toward model transferability in NIR spectroscopy. **Applied Spectroscopy**, v. 52, n. 1, p. 7–16, 1998.

TAN, H.-W.; BROWN, S. D. Wavelet hybrid direct standardization of near-infrared multivariate calibrations. **Journal of Chemometrics**, v. 15, p. 647–663, 2001.

TAN, H.-W.; BROWN, S. D. Wavelet analysis applied to removing non-constant, varying spectroscopic background in multivariate calibration. **Journal of Chemometrics**, v. 16, n. 5, p. 228–240, 2002.

TANIGUCHI, M.; TRESP, V. Averaging regularized estimators. **Neural Computation**, v. 9, p. 1163–1178, 1997.

THOMAS, E. V.; HAALAND, D. M. Comparison of multivariate calibration methods for quantitative spectral analysis. **Analytical Chemistry**, v. 62, p. 1091–1099, 1990.

VALVERDE, R. S.; GARCIA, M. D. G.; GALERA, M. M.; GOICOECHEA, H. C. Determination of tetracyclines in surface water by partial least squares using multivariate calibration transfer to correct the effect of solid phase preconcentration in photochemically induced fluorescence signals. **Analytica Chimica Acta**, v. 562, p. 85–93, 2006.

VETTERLI, M.; KOVACEVIC, J. **Wavelets and Subband Coding**. New Jersey: Prentice Hall, 1995.

WALCZAK, B.; BOUVERESSE, E.; MASSART, D. L. Standardization of near-infrared spectra in the wavelet domain. **Chemometrics and Intelligent Laboratory Systems**, v. 36, p. 41–51, 1997.

WANG, Y.; LYSAGHT, M. J.; KOWALSKI, B. R. Improvement of multivariate calibration through instrument standardization. **Analytical Chemistry**, v. 64, p. 562–564, 1992.

WANG, Y.; VELTKAMP, D. J.; KOWALSKI, B. R. Multivariate instrument standardization. **Analytical Chemistry**, v. 63, p. 2750–2756, 1991.

WANG, Z. M.; HUA, W. Y.; WEI, Z. K.; ZHANG, H. H.; WU, H. Z. Forecasting octane numbers of gasoline by NIR spectroscopy and the manufacture of octane number analyzer. **Spectroscopy and Spectral Analysis**, v. 19, n. 5, p. 684–686, 1999.

WANG, Z. Y.; DEAN, T.; KOWALSKI, B. R. Additive background correction in multivariate instrument standardization. **Analytical Chemistry**, v. 67, p. 2379–2385, 1995.

WOLD, S. Cross-validation estimation of the number of components in factor and principal component analysis. **Technometrics**, v. 24, p. 397–405, 1978.

WOLD, S.; ANTTI, H.; LINDGREN, F.; ÖHMAN, J. Orthogonal signal correction of near-infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 44, p. 175–185, 1998.

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

YOON, J.; LEE, B.; HAN, C. Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. **Chemometrics and Intelligent Laboratory Systems**, v. 64, p. 1–14, 2002.

ZAREI, K.; ATABATI, M.; MALEKSHABANI, Z. Simultaneous spectrophotometric determination of iron, nickel and cobalt in micellar media by using direct orthogonal signal correction-partial least squares method. **Analytica Chimica Acta**, v. 556, p. 247–254, 2006.

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. DOCUMENTO N°	4. N° DE PÁGINAS
TM	06 de junho de 2006	CTA/ITA-IEE/TM-009/2006	113

5. TÍTULO E SUBTÍTULO:

Transferência de Calibração de Instrumentos Para Análise Espectrométrica Empregando Seleção de Variáveis, Reamostragem e Combinação de Modelos

6. AUTOR(ES):

Marcelo do Nascimento Martins

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica. Divisão de Engenharia Eletrônica – ITA/IEE

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Qualidade de Combustíveis; Gasolina; Espectroscopia no Infravermelho; Transferência de Calibração;Transformada Wavelet.

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Controle de qualidade; Combustíveis; Espectroscopia no infravermelho; Calibração; Gasolina; Algoritmo das projeções sucessivas; Análise de ondas localizadas; Análise estatística multivariada; Matemática; Controle

10. APRESENTAÇÃO: **X Nacional** **Internacional**

ITA, São José dos Campos, 2006, 113 páginas.

11. RESUMO:

A espectroscopia no infravermelho tem se mostrado uma ferramenta de valia para o monitoramento da qualidade de combustíveis. Contudo, tal técnica requer a calibração de modelos empíricos para relacionar medidas espectrais com parâmetros físico-químicos de interesse. Neste trabalho propõe-se um método que permite explorar um conjunto de dados já adquirido por um espectrômetro (instrumento Primário) na construção de um modelo para um segundo instrumento (Secundário). Tal método evita a duplicação de custo e esforço experimental no processo de calibração do modelo. Para isso, emprega-se o Algoritmo das Projeções Sucessivas para selecionar variáveis que sejam minimamente redundantes e portem informação relevante nos dois instrumentos. Adicionalmente, é empregado um método de reamostragem e combinação de modelos conhecido como *subagging*. Para validação do método proposto, apresenta-se um estudo de caso envolvendo a determinação de densidade e temperaturas para 10% e 90% de evaporados em amostras de gasolina, assim como um referente a determinação do teor de umidade em amostras de milho. É apresentado, também, um estudo comparativo com técnicas de transferência baseadas em padronização. Os resultados da técnica proposta se mostraram superiores aos obtidos através da técnica clássica de Mínimos-Quadrados Parciais empregando Padronização Direta por Partes. Em particular, verificou-se que o *subagging* propicia uma melhora expressiva na capacidade preditiva dos modelos obtidos por regressão linear múltipla. As técnicas propostas se dividem em abordagens no domínio original de comprimento de onda e no domínio da transformada wavelet. O desenvolvimento no domínio wavelet proporcionou uma redução no esforço computacional.

12. GRAU DE SIGILO:

(X) OSTENSIVO **( ) RESERVADO** **( ) CONFIDENCIAL** **( ) SECRETO**

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