



**UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA**

Dissertação de Mestrado

**Um novo critério para seleção de
variáveis usando o Algoritmo das
Projeções Sucessivas**

Sófacles Figueredo Carreiro Soares

João Pessoa – PB - Brasil

Setembro / 2010



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Dissertação de Mestrado

Um novo critério para seleção de variáveis usando o Algoritmo das Projeções Sucessivas

Sófacles Figueredo Carreiro Soares*

Dissertação apresentada ao programa de Pós-Graduação em Química, da Universidade Federal da Paraíba, como parte dos requisitos para obtenção do título de Mestre em Química.

Orientador: Prof. Dr. Mário César Ugulino de Araújo

Co-orientador: Prof. Dr. Roberto Kawakami Harrop Galvão

***Bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior**

João Pessoa – PB - Brasil

Setembro/2010

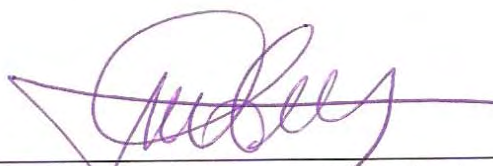
S676u Soares, Sófacles Figueredo Carreiro.
Um novo critério para seleção de variáveis usando o
Algoritmo das Projeções Sucessivas / Sófacles Figueredo
Carreiro Soares.- João Pessoa, 2010.
107f. : il.
Orientador: Mário César Ugulino de Araújo
Co-orientador: Roberto Kawakami Harrop Galvão
Dissertação (Mestrado) – UFPB/CCEN
1. Química. 2. Regressão Linear Múltipla. 3. Seleção de
variáveis. 4. Algoritmo das Projeções Sucessivas. 5. Calibração
multivariada.

UFPB/BC

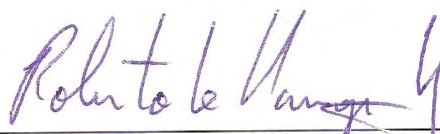
CDU: 54(043)

Um Novo Critério para Seleção de Variáveis Usando o Algoritmo das Projeções Sucessivas.

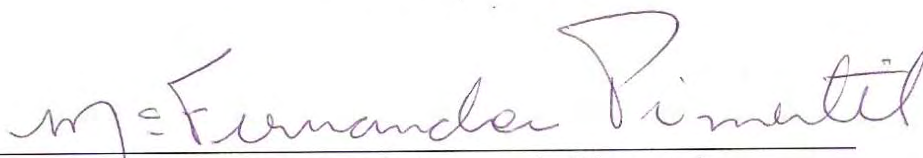
Aprovada pela banca examinadora:



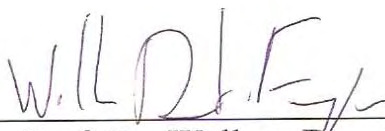
Prof. Dr. Mário César Ugulino de Araújo
Orientador/Presidente



Prof. Dr. Roberto Kawakami Harrop Galvão
2º. Orientador



Prof. Dra. Maria Fernanda Pimentel
Examinadora



Prof. Dr. Wallace Duarte Fragoso
Examinador

*Aos meus pais, Pedro e Neide e meus
irmãos, Pietros e Perla, por todo
amor, amizade e confiança,
com muito carinho eu dedico.*

Agradecimentos

- À Deus;
- À Coordenação de Aperfeiçoamento de Nível Superior, CAPES, pela bolsa concedida;
- Ao professor Mário Cesar Ugulino de Araújo, pela orientação e confiança;
- Ao professor Roberto Kawakami pela orientação e pela atenciosa recepção prestada durante o período de missão realizada no Instituto Tecnológico de Aeronáutica;
- Aos amigos Gledson Emídio e Márcio Coelho, pelo companheirismo e o conhecimento transmitido ao longo de boas conversas;
- Aos amigos Stéfani Yuri e Flaviano Leite pelas medidas dos espectros UV-VIS das amostras de corantes;
- À professora Claudete Fernandes Pereira pela cessão dos espectros de amostras gasolinas;
- Ao programa Nacional de Cooperação Acadêmica (PROCAD), PROCAD 0081/05-1, pelo auxílio financeiro;
- Ao Professor Edvan Cirino da Silva, por ter me dado a oportunidade de iniciar cientificamente;
- Ao Professor Wallace Duarte Fragoso pelas discussões e boas dicas sobre Quimiometria;
- Aos amigos Cleilson, Anabel, Hebertty, Marcelo, Renato, Monte, Adamastor e a todos do LAQA, que de forma indireta me propiciaram um bom ambiente para o desenvolvimento desta dissertação;

- À Arnayra por toda paciência e compreensão;
- À todos que contribuíram para o desenvolvimento deste trabalho.

Sumário

Lista de Figuras.....	ix
Lista de Tabelas	xii
Lista de Siglas e Abreviaturas.....	xiv
Resumo	xv
Abstract	xvi
CAPÍTULO 1.....	1
1. INTRODUÇÃO	2
1.1. Estado-da-Arte do SPA	4
1.2. Objetivos	20
1.2.1. Objetivos específicos	20
CAPÍTULO 2.....	21
2. CALIBRAÇÃO MULTIVARIADA.....	22
2.1. Classificação dos Métodos de Calibração	25
2.2. Regressão Linear Múltipla	26
2.3. Regressão em Componentes Principais	27
2.4. Regressão por Mínimos Quadrados Parciais	28
2.5. Técnicas de Seleção de Variáveis	30
2.5.1. Busca Exaustiva.....	30
2.5.2. Stepwise Regression.....	31
2.5.3. Método de eliminação de variáveis não-informativas	32
2.5.4. <i>Interval</i> PLS	34
2.5.5. Análise de Componentes Independentes.....	34
2.5.6. Algoritmo Genético.....	35
2.5.7. Algoritmo das Projeções Sucessivas	37
2.5.7.1 O SPA para seleção de amostras.....	42
2.5.7.2 O SPA para classificação	42

2.5.7.3 O SPA para previsão na presença de interferentes.....	43
CAPÍTULO 3.....	48
3. Metodologia	49
3.1. Algoritmos usados	49
3.2. Aplicações	51
3.2.1. Dados Simulados	51
3.2.2. Corantes Alimentícios	52
3.2.3. Determinação de Álcool em gasolina	53
CAPÍTULO 4.....	54
4. Resultados e discussões.....	55
4.1. Aplicações aos dados simulados.....	56
4.1.1. Previsão sem interferente.....	57
4.1.2. Previsão na presença de interferente	59
4.2. Determinações de corantes.....	64
4.2.1. Determinações de corantes sem interferente.....	65
4.2.2. Determinações de corantes com interferente.....	68
4.3. Determinação de álcool em gasolina.....	73
CAPÍTULO 5.....	80
5. Conclusões.....	81
5.1. Propostas futuras	81
CAPÍTULO 6.....	83
6. Referencias Bibliográficas.....	84

Lista de Figuras

Figura 2.1	- Principais matrizes envolvida no processo de calibração.....	22
Figura 2.2	- Exemplo do processo de eliminação de variáveis.....	32
Figura 2.3	- Codificação binária usada na seleção de variáveis.	35
Figura 2.4	- Esquema de cruzamento e de mutação no GA.	37
Figura 2.5	- Esquema geral de obtenção de um modelo usando o SPA-SPE-MLR. As variáveis usadas nesse exemplo foram x_1 , x_3 e x_5	46
Figura 2.6	- Representação de quando S é menor que 1.....	47
Figura 3.1	- Aba referente à etapa de calibração para o SPA usando o novo critério.....	49
Figura 3.2	- Aba referente à etapa de previsão para o SPA usando o novo critério.....	50
Figura 4.1	- (a) Espectros puros simulados para os analitos A, B, C e o interferente I. Espectros de misturas para: (b) calibração, (c) validação e (d) previsão.	56
Figura 4.2	- Valores de referência para o analito A <i>versus</i> valores previstos sem interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.	58
Figura 4.3	- Valores de referência para o analito B <i>versus</i> valores previstos sem interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.	58
Figura 4.4	- Valores de referência para o analito C <i>versus</i> valores previstos sem interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.	59
Figura 4.5	- Valores de referência para o analito A <i>versus</i> valores previstos na presença do interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.	60
Figura 4.6	- Valores de referência para o analito B <i>versus</i> valores previstos na presença do interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.	61
Figura 4.7	- Valores de referência para o analito C <i>versus</i> valores previstos na presença do interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.	61

Figura 4.8	- Variáveis usadas para determinação do analito A pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.....	63
Figura 4.9	- Variáveis usadas para determinação do analito B pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR.	63
Figura 4.10	- Variáveis usadas para determinação do analito C pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR.	63
Figura 4.11	- a) Espectros dos corantes tartrazina, vermelho 40, amarelo crepúsculo e o interferente eritrosina puros e os espectros de mistura das amostras de: b) calibração, c) validação e d) previsão.....	65
Figura 4.12	- Valores de referência do corante amarelo crepúsculo <i>versus</i> valores previstos sem a presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.	66
Figura 4.13	- Valores de referência do corante tartrazina <i>versus</i> valores previstos sem a presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.....	67
Figura 4.14	- Valores de referência do corante vermelho 40 <i>versus</i> valores previstos sem a presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.....	67
Figura 4.15	- Espectros de eritrosina variando a concentração entre 1,0 e 10,0 mg L ⁻¹	68
Figura 4.16	- Valores de referência do corante amarelo crepúsculo <i>versus</i> valores previstos na presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.	69
Figura 4.17	- Valores de referência do corante tartrazina <i>versus</i> valores previstos na presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.	70
Figura 4.18	- Valores de referência do corante vermelho 40 <i>versus</i> valores previstos na presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.	70
Figura 4.19	- Variáveis selecionadas para determinação do corante amarelo crepúsculo usando o conjunto de previsão com interferente pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.	72
Figura 4.20	- Variáveis selecionadas para determinação do corante tartrazina usando o o conjunto de previsão com interferente pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.....	72

- Figura 4.21** - Variáveis selecionadas para determinação do corante vermelho 40 usando o conjunto de previsão com interferente pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR..... 72
- Figura 4.22** - Espectros NIR de 36 amostras de gasolina com a concentração de etanol variando entre 10 e 38% (v/v). 73
- Figura 4.23** - Valores de referência de etanol *versus* valores previstos sem interferentes pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS. 75
- Figura 4.24** - Valores de referência de etanol *versus* valores previstos, na presença do interferente tolueno, pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS. 75
- Figura 4.25** - Valores de referência de etanol *versus* valores previstos, na presença do interferente hexano, pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS. 76
- Figura 4.26** - Valores de referência de etanol *versus* valores previstos, na presença do interferente iso-octano, pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS. 76
- Figura 4.27** - Espectros de absorção de gasolina tipo C, tolueno e etanol. Regiões de absorção dos grupos: (a) C-H de aromático; (b) e (c) metila e metileno, (d) metila/OH e metileno/OH, (e) O-H e (f) OH/C-H de aromático (fonte: Pereira et al.^[76]). 78
- Figura 4.28** - Variáveis selecionadas para construção de modelos MLR para determinação de etanol usando: ● SPA-SPE-MLR sem interferentes e interferidas por hexano, ● SPA-SPE-MLR interferido por tolueno e iso-octano e ● SPA-MLR sem interferentes e interferidas por hexano, iso-octano e tolueno. ... 78

Lista de Tabelas

Tabela 2.1	- Classificação do uso dos modelos a partir dos valores de RPD...	25
Tabela 3.1	- Níveis de concentração (mg L^{-1}) dos corantes nas misturas de calibração.	52
Tabela 3.2	- Níveis de concentração (mg L^{-1}) dos corantes nas misturas de validação.	53
Tabela 4.1	- RMSEPs obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS para o conjunto de previsão sem interferente. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.	57
Tabela 4.2	- RMSEPs obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS para o conjunto de previsão com interferente. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.	60
Tabela 4.3	- Valores de t_{cal} e t_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação dos analitos simulados. O número de graus de liberdade encontra-se entre parêntesis.	62
Tabela 4.4	- Valores de F_{cal} e F_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação dos analitos simulados. Os números de graus de liberdade encontram-se entre parêntesis.	62
Tabela 4.5	- RMSEPs obtidos para determinação dos corantes sem a presença de eritrosina (mg L^{-1}) pelos modelos SPA-SPE-MLR, SPA-MLR e PLS. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.	66
Tabela 4.6	- RMSEPs obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS na presença do interferente eritrosina (mg L^{-1}). O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.	69
Tabela 4.7	- Valores de t_{cal} e t_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação dos corantes. O número de graus de liberdade encontra-se entre parêntesis.	71

- Tabela 4.8** - Valores de F_{cal} e F_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação de corantes. Os números de graus de liberdade encontram-se entre parêntesis..... 71
- Tabela 4.9** - RMSEPs do teor de etanol (%v/v) em amostras de gasolinas obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis. 74
- Tabela 4.10** - Valores de t_{cal} e t_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS). Os números de graus de liberdade encontram-se entre parêntesis. 77
- Tabela 4.11** - Valores de F_{cal} e F_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS). Os números de graus de liberdade encontram-se entre parêntesis. 77

Lista de Siglas e Abreviaturas

BIAS	Medida estatística de erro sistemático
CWSPA	Algoritmo das projeções sucessivas com correlação ponderada
F_{cal}	Valor calculado para o teste F
F_{custo}	Função usada como critério de escolha
F_{crit}	Valor tabelado para o teste F
GA	Algoritmo genético
iPLS	Regressão por intervalos em mínimos quadrados parciais
LIBS	Espectrometria de emissão em plasma induzido por laser
MLR	Regressão linear múltipla
NIR	Infravermelho próximo
PCR	Regressão em componentes principais
PLS	Regressão em mínimos quadrados parciais
RMSECV	Raiz quadrada do erro médio quadrático de validação cruzada
RMSEP	Raiz quadrada do erro médio quadrático de previsão
RMSEV	Raiz quadrada do erro médio quadrático de validação
RSECV%	Erro relativo padrão de validação cruzada
RSEP%	Erro relativo padrão de previsão
S	Razão do erro de previsão estatística para o conjunto de previsão e calibração
SEP	Erro padrão de previsão
SPA	Algoritmo das projeções sucessivas
SPA-MLR	Modelo de calibração obtido com as variáveis selecionadas pelo SPA
SPA-SPE-MLR	Modelo de calibração obtido com as variáveis selecionadas pelo SPA-SPE
SPE_i	Erro de previsão estatística para a i-ésima amostra
$\overline{SPE}$	Erro de previsão estatística médio
t_{cal}	Valor calculado para o teste t
t_{crit}	Valor tabelado adotado para o teste t
UVE	Eliminação de variáveis não-informativas
UV-VIS	Ultravioleta visível
X_{cal}	Matriz de medidas usadas no conjunto de calibração

Resumo

Este trabalho propõe uma modificação no Algoritmo das Projeções Sucessivas (*Sucessive Projection Algorithm* - SPA), com objetivo de aumentar a robustez a interferentes nos modelos de Regressão Linear Múltipla (*Multiple Linear Regression* - MLR) construídos. Na formulação original do SPA, subconjuntos de variáveis são comparados entre si com base na raiz do erro quadrático médio obtido em um conjunto de validação. De acordo com o critério aqui proposto, a comparação é feita também levando em conta o erro estatístico de previsão (*Statistical Prediction Error* - SPE) obtido para o conjunto de calibração dividido pelo erro estatístico de previsão obtido para o conjunto de previsão. Tal métrica leva em conta a leverage associada a cada amostra. Três estudos de caso envolvendo a determinação de analitos simulados, corantes alimentícios por espectrometria UV-VIS e álcool em gasolinas por espectrometria NIR são discutidos. Os resultados são avaliados em termos da raiz do erro quadrático médio em um conjunto de previsão independente (*Root Mean Square Error of Prediction* - RMSEP), dos gráficos das variáveis selecionadas e através do testes estatísticos *t* e *F*. Os modelos MLR obtidos a partir da seleção usando a nova função custo foram chamados aqui de SPA-SPE-MLR. Estes modelos foram comparados com o SPA-MLR e PLS. Os desempenhos de previsão do SPA-SPE-MLR apresentados foram melhores em quase todos os modelos construídos quando algum interferente estava presente nos espectros de previsão. Estes modelos quando comparados ao SPA-MLR, revelou que a mudança promoveu melhorias em todos os casos fornecendo RMSEPs e números de variáveis menores. O SPA-SPE-MLR só não foi melhor que alguns modelos PLS. As variáveis selecionadas pelo SPA-SPE-MLR quando observadas nos espectros se mostraram em regiões onde a ação do interferente foi à menor possível revelando o grande potencial que tal mudança provocou. Desta forma a modificação aqui apresentada pode ser considerada como uma ferramenta útil para a formulação básica do SPA.

Palavras chave: Regressão linear múltipla, seleção de variáveis, algoritmo das projeções sucessivas e calibração multivariada.

Abstract

This study proposes a modification in the Successive Projections Algorithm (SPA), that makes models of Multiple Linear Regression (MLR) more robust in terms of interference. In SPA, subsets of variables are compared based on their root mean square errors for the validation set. By taking into account the statistical prediction error obtained for the calibration set, and dividing by the statistical prediction error obtained for the prediction set, SPA can be improved. Also taken into account is the leverage associated with each sample. Three case studies involving; simulated analytic determinations, food colorants (UV-VIS spectrometry), and ethanol in gasoline (NIR spectrometry) are discussed. The results were evaluated using the root mean square error for an independent prediction set (Root Mean Square Error of Prediction - RMSEP), graphs of the variables, and the statistical tests *t* and *F*. The MLR models obtained by the selection using the new function were called SPE-SPA-MLR. When an interferent was present in the prediction spectra, almost all of the models performed better than both SPA-MLR and PLS. The models when compared to SPA-MLR showed that the change promoted better models in all cases giving smaller RMSEPs and variable numbers. The SPE-SPA-MLR was not better in some cases, than PLS models. The variables selected by SPA-SPE-MLR when observed in the spectra were detected in regions where interference was the at its smallest, revealing great potential. The modifications presented here make a useful tool for the basic formulation of the SPA.

Keywords: Multiple linear regression, variable selection, successive projections algorithm and multivariate calibration.

CAPÍTULO 1

1. INTRODUÇÃO

Durante muitos anos a química analítica fez uso de métodos quantitativos que se baseavam em medidas titulométricas e gravimétricas^[1]. Essas medidas, em sua maioria, sempre estiveram associadas ao uso de grandes quantidades de reagentes químicos e demanda de tempo para a execução das análises.

O desenvolvimento da ciência básica e da tecnologia permitiu o aparecimento de muitos equipamentos nos laboratórios de química analítica. Esses instrumentos, com seus princípios de funcionamentos fundamentados em fenômenos físicos ou físico-químicos, fizeram nascer o que ficou conhecido como métodos instrumentais^[1]. Tais métodos permitiram que os químicos analíticos passassem a viver uma nova maneira de fazer a química analítica, registrando suas diversas medidas com uma relativa facilidade e realizando suas determinações de forma mais rápida e sensível.

Os métodos instrumentais alavancaram a aquisição de muitas informações por amostras, como exemplos hoje se tem a espectroscopia de emissão em plasma induzido por laser (*Laser Induced Breakdown Spectroscopy - LIBS*)^[2], e a espectroscopia no infravermelho próximo (*Near Infrared - NIR*)^[3] capazes de gerar rapidamente um considerável número de medidas por cada amostra.

A quimiometria que é definida pela sociedade internacional de quimiometria^[4] como “a ciência de relacionar medidas feitas em um sistema ou processo químico com propriedades do sistema ou processo via aplicação de métodos estatísticos ou matemáticos” surgiu com a finalidade de auxiliar no processamento dessas informações de forma rápida e eficiente.

Com seu uso contínuo a quimiometria passou a ser usada para melhorar de diversas formas os estudos químicos, entre esses estudos destaca-se: o planejamento experimental, processamento de sinais, reconhecimento de padrões e classificação, métodos de inteligência artificial, calibração multivariada, entre outros. Dentre todos os objetos de estudo da quimiometria a calibração multivariada é uma das áreas que atraem o maior interesse ao longo dos anos^[5].

A calibração multivariada tem como objetivo relacionar um parâmetro, com propriedades mais simples de serem obtidas, de forma que exista uma boa relação matemática entre elas, gerando previsões dos parâmetros com rapidez e precisão^[6]. A calibração multivariada tem sido aplicada nas últimas décadas com

sucesso em diversos campos como: química de alimentos, análise farmacêutica, agricultura, meio ambiente, indústria e química clínica^[5].

Um dos maiores problemas para se realizar uma calibração multivariada consiste em fazer com que informações redundantes não eliminem a capacidade de construir modelos satisfatórios^[7]. Em quimiometria quando há muitas variáveis no registro de uma medida química, é possível que essas apresentem alguma relação. Como exemplo simples, pode-se citar medidas de DBO (demanda bioquímica de oxigênio) e OD (oxigênio dissolvido). Com efeito, uma quantidade alta de oxigênio dissolvido em uma amostra está naturalmente associada a um valor baixo para a medida de DBO.

Assim como DBO e OD, as variáveis em medidas espectroscópicas guardam uma grande relação entre elas, possibilitando que mais de duas variáveis possam possuir colinearidade. Quando mais de duas variáveis possuem natureza colinear, o termo colinearidade é substituído por multicolinearidade^[8].

Para resolver os problemas causados pela multicolinearidade na calibração multivariada duas maneiras estão mais difundidas na literatura: os métodos que usam as variáveis originais e os que transformam as variáveis originais em variáveis latentes para a regressão^[6-7,9]. Dentre os métodos que usam combinações lineares das variáveis originais temos a regressão por componentes principais (*Principal Component Regression* – PCR) e a regressão usando mínimos quadrados parciais (*Partial Least Squares* – PLS). A regressão linear múltipla (*Multiple Linear Regression* – MLR) usa a regressão a partir das variáveis originais.

A idéia principal destes dois métodos (PCR e PLS) consiste em estabelecer combinações lineares das variáveis originais e usar essas nas equações de regressão. Essa é uma maneira de descartar as informações irrelevantes e usar somente a parte da variação relevante para a regressão em coordenadas ortogonais. Quando isso é realizado, os modelos PCR e PLS podem não permitir uma interpretação direta do comportamento físico-químico do sistema, pois a regressão é realizada no domínio dos dados transformados^[9,10]. A técnica MLR, ao contrário, não somente realiza a regressão no domínio dos dados originais, mas também possibilita a construção de modelos mais simples e fáceis de interpretar.

Apesar das vantagens da modelagem MLR, a regressão é muito sensível a problemas de multicolinearidade nos dados instrumentais (matriz X). Para

minimizar esse problema, foram desenvolvidas técnicas para seleção de variáveis em calibração multivariada^[3,11-13]. Entre essas estratégias, destaca-se o algoritmo das projeções sucessivas (*Successive Projections Algorithm* - SPA).

Em diversas aplicações, o SPA se mostrou capaz de produzir modelos MLR com uma capacidade de previsão melhor (erros menores nas estimativas dos parâmetros de interesse) que os modelos PCR/PLS construídos usando todas as variáveis. Além disso, o SPA apresenta a vantagem de selecionar, de maneira reproduzível, variáveis que portam informações químicas ou físicas do sistema ao contrário, por exemplo, do algoritmo genético (GA). A seguir, realiza-se uma revisão bibliográfica do SPA mostrando seu desenvolvimento e o sucesso de suas aplicações.

1.1. Estado-da-Arte do SPA

O algoritmo das projeções sucessivas (*Successive Projection Algorithm* – SPA) foi proposto em 2001 por Araújo et al.^[13], com objetivo de selecionar variáveis para a construção de modelos multivariados usando medidas espectrométricas UV-VIS para a determinação simultânea de complexos de Co^{2+} , Cu^{2+} , Mn^{2+} , Ni^{2+} e Zn^{2+} . Entretanto ao longo dos últimos anos o SPA foi bastante usado em calibração multivariada, classificação, seleção de amostras, transferência de calibração, modelagem envolvendo estrutura atividade (QSAR) e seleção de coeficientes no domínio wavelet. A sua primeira aplicação foi comparada com o algoritmo genético, isso resultou em um RMSEV (*Root Mean Square Error of Validation*) de 0,02 mg L⁻¹, enquanto os melhores e piores resultados do algoritmo genético, usado para comparação, foram 0,01 e 0,03 mg L⁻¹. Nesse trabalho foi demonstrado que o SPA, a partir de operações simples poderia permitir o seu uso para a seleção de variáveis minimamente multicolineares, com uma baixa demanda computacional e de maneira determinística.

Embora o SPA tenha sido aplicado com sucesso para análise multicomponente espectrofotométrica UV-VIS, não havia provas da sua capacidade para lidar com um conjunto de variáveis que alternam entre alta e baixa razão sinal/ruído. Esta questão foi abordada em um trabalho de Galvão et al.^[14] em que o SPA foi aplicado a dados simulados com intuito de demonstrar mais claramente o seu funcionamento, sua maior robustez a ruído e redução do

esforço computacional promovido pela seleção de variáveis por SPA quando comparada a busca exaustiva. Nesse trabalho, o SPA foi ainda aplicado com intuito de determinar simultaneamente Mn, Mo, Cr, Ni e Fe usando um espectrômetro de plasma de baixa resolução com sistema de detecção de arranjo de díodos. Os resultados mostraram que os modelos MLR em comprimentos de onda selecionados pelo SPA apresentaram uma melhor capacidade de previsão que os modelos construídos usando PCR e PLS. O algoritmo genético foi utilizado para fins de comparação e produziu resultados semelhantes aos do SPA para Mn, Cr e Fe, e melhores previsões para Mo e Ni. Contudo, em todos os casos, o GA resultou em modelos menos parcimoniosos do que o SPA.

Em 2003, o SPA foi usado por Coelho et al.^[15] em um conjunto de dados processado com o uso de uma transformada wavelet otimizada. A otimização foi formulada com base em uma programação linear semi-infinita, que não apresenta problemas de máximos locais e pode ser resolvido de maneira reproduzível e com baixo esforço computacional. Após a otimização o SPA foi usado na seleção de subconjuntos dos coeficientes wavelet com menor multicolinearidade possível. A seleção permitiu a realização da regressão linear múltipla direta sobre os coeficientes wavelet em uma aplicação ilustrativa envolvendo a determinação simultânea de Mn, Mo, Cr, Ni e Fe em amostras de aço por ICP-OES (*Inductively Coupled Plasma Optical Emission Spectroscopy*). Essa estratégia resultou em previsões mais adequadas do que os modelos PCR, PLS e a regressão wavelet não-otimizada.

Ainda em 2003, Coelho et al.^[16] propuseram a otimização da transformada wavelet para modelos multivariados. Para maximizar o desempenho, foram usados bancos de filtros baseados no ajuste de Quadratura Espelhada (*Quadrature Mirror-Filter* - QMF) para processar os espectros. Após a fase de processamento dos espectros o SPA é usado para selecionar os subconjuntos de coeficientes wavelet. A aplicação a dados ICP-OES de baixa resolução para determinação multicomponente simultânea de Mn, Mo, Cr, Ni e Fe em amostras de aço é intrinsecamente complexa, devido à multicolinearidade forte e severa sobreposição espectral, problemas que são agravados pelo uso da óptica de baixa resolução. Os resultados demonstraram que a otimização wavelet combinada com SPA conduziu a modelos com maior parcimônia e capacidade de previsão.

Após o uso para a seleção de coeficientes no domínio wavelet, o SPA foi usado para seleção de comprimentos de onda em espectros na região do infravermelho próximo para determinação de enxofre em diesel^[17]. Os desempenhos dos métodos quimiométricos de calibração multivariada PCR e PLS foram comparados com a regressão MLR realizada após a seleção de variáveis usando GA e o SPA. Embora a análise de componentes principais identificasse a presença de três grupos, o PLS, PCR e MLR promoveram modelos cujas previsões foram independentes do tipo de diesel. A calibração com PLS e PCR empregando todos os comprimentos de onda produziu RMSEPs de 0,036%/m/m e 0,043%/m/m para o conjunto de validação, respectivamente. O uso de GA e SPA promoveu modelos de calibração com base em 19 e 9 comprimentos de onda, com RMSEP de 0,031%/m/m (GA-PLS), 0,022%/m/m (SPA-MLR) e 0,034%/m/m (GA-MLR). Estes modelos podem vir a substituir os métodos oficiais, pois promoveram RMSEPs equivalentes a reprodutibilidade de 0,05%/m/m. A seleção de variáveis ofereceu modelos de calibração mais robustos, sendo o SPA mais parcimonioso que o GA, nesta aplicação.

O SPA foi usado por Dantas Filho et al.^[18] na seleção de amostras para modelos de calibração multivariada. O SPA selecionou um subconjunto de amostras que são minimamente redundantes, mas ainda representativa do conjunto de dados. Tal procedimento foi capaz de reduzir a carga de trabalho experimental e computacional envolvido na calibração multivariada. A estratégia foi aplicada a análise simultânea multicomponente de complexos de Co^{2+} , Cu^{2+} , Mn^{2+} , Ni^{2+} e Zn^{2+} , por espectrometria UV-VIS e também na determinação de enxofre em diesel por espectrometria NIR. Em ambas as aplicações, o SPA reduziu o número de variáveis e amostras consideravelmente sem nenhuma perda significativa da capacidade de previsão, quando comparado com modelos MLR e PLS construídos com o conjunto completo de amostras de calibração. Esse trabalho mostrou que as amostras selecionadas portam as informações necessárias para a modelagem. Além disso, na aplicação NIR, a seleção de amostras pelo SPA, proporcionou resultados significativamente melhores que o algoritmo Kennard-Stone (KS).

Nenhuma tentativa tinha sido feita na exploração dos dados estatísticos da matriz de parâmetros previstos (matriz Y) no processo de otimização wavelet, Galvão et al.^[19] propuseram uma estratégia para a otimização que visa minimizar diretamente o erro de previsão de um modelo de regressão wavelet

aplicado a um conjunto de validação. O SPA foi usado para guiar a seleção dos coeficientes wavelet. A estratégia foi ilustrada em um exemplo simulado de calibração multivariada envolvendo dois analitos e também em um problema de determinação de enxofre total em amostras de diesel por espectrometria de absorção NIR. Os exemplos, simulado e real, mostraram que o procedimento de otimização melhorou a capacidade de previsão dos modelos de regressão wavelet. Em acréscimo, os modelos wavelet resultaram em um menor erro de previsão do que o método tradicional PLS.

Após o desenvolvimento de uma metodologia que possibilitasse o uso do SPA para selecionar amostras^[18], Honorato et al.^[20] usaram o SPA para selecionar variáveis e amostras para a construção de modelos MLR robustos e transferíveis. As amostras de transferência foram selecionadas pelo algoritmo clássico Kennard-Stone (KS) ou uma versão modificada do SPA que opera nas linhas da matriz de respostas instrumentais, em vez das colunas. Dois conjuntos de dados foram utilizados na avaliação, cada conjunto composto de espectros obtidos por espectrometria no infravermelho em dois instrumentos diferentes. O primeiro conjunto composto de espectros de gasolina, que foram utilizados para prever a temperatura de destilação na qual 90% da amostra tenha evaporado (T90%). O segundo conjunto composto de espectros de milho foi empregado para a determinação de umidade. Os modelos MLR robustos foram comparados a um modelo PLS empregando Padronização Direta por Partes (*Piecewise Direct Standardization* - PDS) para corrigir os espectros registrados no aparelho escravo. Em ambos os conjuntos de dados, a média dos erros de previsão no instrumento escravo para os modelos robustos SPA-MLR e PLS-PDS foram comparáveis, porém um pouco melhor para o SPA-MLR.

Com as variáveis selecionadas pelo SPA foram construídos modelos MLR para determinação espectrofotométrica simultânea dos íons divalentes de cobre, manganês e zinco para análise de medicamento polivitamínico/polimineral^[21]. Os resultados dos modelos PCR, PLS e MLR obtidos a partir das variáveis selecionadas pelo GA foram comparados com os SPA-MLR. Os RMSEPs dos modelos mostraram desempenhos semelhantes ao prever as concentrações dos três analitos no medicamento. Todavia os modelos MLR foram mais simples, pois necessitaram de um número menor de comprimentos de onda.

Pontes et al.^[22] exploraram o papel de minimização de multicolinearidade pelo SPA no contexto dos métodos de classificação. A análise de discriminante

linear (*Linear Discriminant Analysis* - LDA) foi empregada na construção de modelos de classificação com base em um subconjunto de variáveis espectrais selecionadas pelo SPA e o GA. Os modelos foram usados na classificação de óleos vegetais com relação ao tipo (milho, soja, canola e girassol) usando a espectrometria UV-VIS e para discriminar amostras de diesel no que diz respeito à concentração de enxofre usando espectrometria NIR. Nos dois exemplos, SPA-LDA é comparado com o método de classificação SIMCA (*Soft Independent Modeling of Class Analogy* - SIMCA), bem como com o GA-LDA. Os resultados mostraram que o SPA-LDA é superior ao SIMCA e comparável ao GA-LDA no que diz respeito à exatidão da classificação em um conjunto de previsão independente, entretanto o SPA-LDA foi menos sensível ao ruído instrumental que o GA-LDA.

Galvão et al.^[23] empregaram o SPA para selecionar variáveis em modelos multivariados MLR usando uma técnica conhecida como subamostragem de forma a produzir a variabilidade da composição de amostras reais. Os modelos assim obtidos foram aplicados a problemas envolvendo a determinação por espectrometria NIR em diesel de quatro parâmetros de qualidade (massa específica, teor de enxofre, e as temperaturas de destilação T10% e T90%). A utilização de 30 iterações de subamostragem proporcionou melhorias de 16%, 33% e 35% na precisão da previsão dos modelos PLS, SPA-MLR e GA-MLR, respectivamente, com relação aos resultados esperados para cada modelo individual.

Como novo desafio, o SPA, foi aplicado para selecionar descritores moleculares na construção de um modelo não-linear de relações quantitativas estrutura-atividade (*Quantitative Structure-Activity Relationship* - QSAR)^[24], para a previsão da atividade anti-HIV-1 de uma série de derivados da molécula 1-[(2-hidroxietoxi)metil]-6-(tiofenil)timina conhecidos como HEPT. Utilizando o SPA e o *Stepwise Backward Elimination* no descarte de variáveis não-informativas, um subconjunto de 11 descritores foi selecionado. Três redes neurais com funções de base radial (*Radial Basis Function Neural Networks* - RBFNs) foram utilizadas para construir os modelos não-lineares QSAR em todas as fases do estudo. O percentual de erro padrão relativo das previsões da atividade anti-HIV para o conjunto de treinamento através da aplicação de validação cruzada (RSECV%) foi 9,94%, e para o conjunto de previsão (RSEP%) foi de 9,99%. O modelo obtido supera vários modelos encontrados na literatura,

tanto na fase de montagem como na previsão. A determinação resultou em atividades previstas em excelente concordância com os valores experimentais.

Caneca et al.^[25] propuseram o uso do SPA-MLR para a previsão da viscosidade de óleos lubrificantes a partir de medidas de refletância total atenuada na região do infravermelho médio. Os modelos por SPA se mostraram superiores quando comparados aos modelos PLS, PCR e GA-MLR, fornecendo um RMSEP de 3,8 cSt e um erro relativo médio de 3,2%.

O uso de um conjunto de validação separado e de validação cruzada amostra-a-amostra para guiar a seleção de variáveis SPA para calibração multivariada foi comparado^[26]. Embora em Araújo et al.^[13] uma formulação teórica em que o RMSECV foi apresentado para guiar a escolha da seleção de variáveis pelo SPA, nenhum estudo havia sido realizado com o uso de tal abordagem. Para isso foram realizadas análises de diesel e milho por espectrometria NIR^[26], e os resultados mostraram que tanto o uso da validação por série de teste como a validação cruzada pode ser de grande valia usando o SPA.

No ano de 2007 o SPA com correlação ponderada (*Correlation Weighted Successive Projections Algorithm* - CWSPA) foi proposto^[27]. Esse algoritmo é uma versão modificada do SPA. O CWSPA foi usado para a seleção de descritores, para obter modelos QSAR, no estudo de uma série de derivados das moléculas 1-[(2-hidroxietoxi)metil]-6-(tiofenil)timina (HEPT). RBFNs foram utilizadas para construir os modelos QSAR não-lineares. O desempenho do CWSPA foi medido no conjunto de treinamento através da aplicação de validação cruzada. O valor de RSECV% foi 9,77%, e para o conjunto de previsão, o RSEP% foi 8,61% quando o número de descritores selecionados foram 20. O modelo obtido supera os indicados na literatura, tanto na construção do modelo como na previsão.

Di Nezio et al.^[28] propuseram um método analítico para determinação direta e simultaneamente de cinco compostos fenólicos (4-nitrofenol, 2-nitrofenol, fenol, 2,4,6-triclorofenol e 4-clorofenol) em água do mar. O SPA foi usado para selecionar variáveis em medidas espectrométricas na região do ultravioleta e visível (UV-VIS). O uso de modelos quimiométricos de calibração multivariada permitiu maior rapidez e economia de reagentes (etapas de separação e reagentes de derivatização são evitadas). A comparação estatística

entre o PLS, e SPA-MLR, mostrou que o modelo SPA-MLR apresentou um melhor desempenho analítico.

Uma revisão geral das técnicas de transferência de calibração foram apresentadas por Honorato et al.^[29]. Conceitos básicos foram revisados, bem como as principais vantagens e desvantagens de cada técnica. Um estudo de caso baseado em um conjunto de 80 espectros NIR de amostras de milho registradas em dois diferentes instrumentos foi utilizado para ilustrar as principais técnicas de transferência de calibração (Padronização Direta, Padronização Direta por Partes, Correção de Sinal Ortogonal e Seleção de Variáveis Robustas). Nesse trabalho o SPA foi mostrado como uma alternativa viável as técnicas tradicionais, produzindo resultados mais satisfatórios para três das quatro propriedades no estudo de caso apresentado.

Uma simples modificação na formulação SPA aumentou a parcimônia do modelo MLR resultante^[30]. A utilidade da modificação foi ilustrada em um estudo simulado, e em dois exemplos que envolvem a análise de amostras de diesel e milho por espectrometria NIR. Os resultados demonstraram que o número de variáveis selecionadas pelo SPA pôde ser reduzida sem comprometer significativamente o desempenho de previsão. Além disso, o SPA apresentou um melhor desempenho quando comparado com *stepwise regression* clássico e o método PLS aplicado aos espectros completo.

Para acompanhar a síntese enzimática de penicilinas semi-sintéticas, produzidas em um reator semi-contínuo com cristalização simultânea dos antibióticos, Ribeiro et al.^[31] propuseram o uso da calibração multivariada com objetivo de reduzir os tempos de análise em sistemas multicomponentes. Três métodos de calibração multivariada (SPA-MLR, PCR e PLS) são comparados em termo da eficiência da previsão das concentrações dos componentes presentes na síntese enzimática de ampicilina. Para a construção dos modelos, espectros no UV foram medidos e suas derivadas de primeira ordem foram utilizadas. Os resultados mostraram que todos os métodos (SPA-MLR, PCR e PLS) apresentaram desempenhos semelhantes. A validade foi testada com base em ensaios de bancada da síntese de ampicilina usando Penicilina G acilase imobilizada como catalisador. A comparação dos resultados com análises HPLC (*High Performance Liquid Chromatograph* - HPLC) mostrou que o método proposto para controlar as concentrações de ampicilina (produto desejado) e ácido 6-aminopenicilânico (6-APA, substrato mais caro) pode ser obtida com precisão

similar. As previsões dos compostos mais sensíveis (éster metílico de fenilglicina, PGME e fenilglicina, PG) não foram satisfatórias, porém informações prévias do estado inicial do reator podem ser usadas para melhorar a estimativa desses componentes.

Um sistema enzimático em fluxo batelada com detecção espectrométrica foi desenvolvido para determinação simultânea de levodopa [ácido (S)-2-amino-3-(3,4-dihidroxifenil) propanóico] e carbidopa [ácido (S)-3-(3,4-diidroxifenil)-2-hidrazino-2-metil propanóico] em preparações de produtos farmacêuticos^[32]. A enzima polifenol oxidase (*Poly Phenol Oxidase* - PPO) foi utilizada para oxidar os dois analitos as suas respectivas dopaquinonas, que apresentaram uma forte absorção espectrofotométrica entre 295 e 540 nm. A partir dessas medidas, modelos foram obtidos pelos métodos de calibração: univariada, PLS e SPA-MLR. Um estudo comparativo entre a análise univariada, PLS em diferentes faixas, e SPA-MLR, foi realizada através da aplicação da região de confiança elíptica comum. Os resultados foram satisfatórios para o SPA-MLR e o PLS na região espectral entre 295 e 540 nm. Comprimidos de amostras comerciais foram determinados e os resultados obtidos estiveram em concordância com os métodos que a farmacopéia recomenda.

Pereira et al.^[33] combinaram a seleção de faixa espectral por mínimos quadrados parciais em intervalos (*interval Partial Least Squares* - iPLS) com a seleção de variáveis pelo SPA para obtenção de modelos MLR, objetivando determinar simultaneamente índice de acidez, índice de refração e viscosidade por espectrometria NIR, em quatro tipos de óleos vegetais comestíveis (milho, soja, canola e girassol). Um conjunto de amostras independentes foi empregado para avaliar a capacidade de previsão dos modelos resultantes iPLS-SPA-MLR. Como resultado, valores de correlação de 0,94, 0,98 e 0,96 foram obtidos entre os valores previstos e valores de referência para a determinação do índice de acidez, índice de refração e viscosidade, respectivamente. Os resultados mostraram que uma única calibração pode ser realizada com sucesso para cada parâmetro, sem a necessidade de desenvolver um modelo separado para cada tipo de óleo vegetal.

O SPA combinado com a eliminação de variáveis não-informativas foi proposto como uma nova abordagem de seleção de variáveis para calibração multivariada^[34]. O método de eliminação de variáveis não informativas (*uninformative variable elimination* - UVE) foi usado e o SPA é seguido para

selecionar o menor número de variáveis com um mínimo de informações redundantes. O método proposto foi aplicado a dados de transmissão no infravermelho próximo (NIR) para a análise de nicotina em lâmina de tabaco e dados de reflexão NIR para determinação de ingrediente farmacêutico ativo (IFA) em comprimidos. Para análise NIR de nicotina em lâmina de tabaco, o número de variáveis selecionadas pelo SPA foi 48 enquanto que o UVE-SPA selecionou 35, os RMSEPs dos modelos MLR correspondentes diminuíram de 0,174 %mg/mg para 0,160 %mg/mg. Na análise NIR de IFA, o número de variáveis selecionadas a partir de 650 variáveis espectrais foram reduzidas de 46 pelo SPA, a 17 pelo UVE-SPA, e os RMSEPs dos modelos MLR diminuíram de 0,842 %mg/mg para 0,473 %mg/mg. Os modelos UVE-SPA-MLR produziram previsões melhores do que as realizadas pelo PLS usando o espectro completo.

A fim de extrair as informações de baixa intensidade, relacionadas a concentração de glicose, o método de eliminação de variáveis não informativas modificado (*modified Uninformative Variable Elimination* – mUVE) combinado com o SPA, nomeado como mUVE-SPA, foi usado^[35]. A interferência de fundo, derivada da absorção óptica de componentes da matriz; baixa seletividade e sensibilidade espectral são os principais fatores interferentes para a determinação não-invasiva de glicose no sangue humano usando medidas na região do infravermelho próximo (NIR). O mUVE é usado para eliminar a matriz de fundo e o ruído de alta frequência pela tecnologia multi-resolução wavelet e o SPA é usado para selecionar as variáveis com menor multicolinearidade possível. O mUVE-SPA e o SPA foram aplicados em dois conjuntos de dados espectrais NIR para a determinação de glicose a partir das variáveis selecionadas para a construção de modelos PLS. O primeiro conjunto é composto por medidas em amostras de plasma humano in vitro e o segundo por medidas não invasivas in vivo. Os resultados indicaram que o método híbrido proposto pode dar um caminho alternativo para extrair informações de glicose em sangue humano de forma não invasiva. Os modelos mUVE-SPA-PLS apresentaram uma maior parcimônia e precisão nas previsões quando comparado com os modelos SPA-PLS.

A espectrometria na região do visível e infravermelho próximo (VIS/NIR) foi usada para determinar o ácido acético, tartárico e láctico em vinagre de ameixa com base nas variáveis selecionadas pelo SPA^[36], combinada com a regressão pelo método dos mínimos quadrados usando máquinas de vetores de

suporte (*Least Squares-Support Vector Machine* - LS-SVM)^[37] que é uma técnica capaz de lidar com análise multivariada linear ou não-linear e resolver esses problemas de uma forma relativamente rápida. Uma descrição detalhada do método LS-SVM pode ser visto em Suykens et al.^[38]. Os modelos obtidos a partir das variáveis selecionadas pelo SPA foram comparados com os obtidos por meio do método de seleção de variáveis baseado na análise dos coeficientes de regressão (*Regression Coefficients Analysis* - RCA) PLS. Modelos MLR e PLS foram desenvolvidos para comparação com o LS-SVM. Os resultados indicaram que SPA-LS-SVM alcançou o melhor desempenho para os três ácidos em comparação com PLS usando o espectro completo, SPA-MLR, SPA-PLS, RCA-PLS e RCA-LS-SVM. Os RMSEPs foram 0,3581, 0,0714 e 0,0201 g L⁻¹ para determinação dos ácidos acético, tartárico e láctico, respectivamente para os modelos usando SPA-LS-SVM. Os resultados globais mostraram que a espectroscopia VIS/NIR em conjunto com técnicas como o SPA-LS-SVM pode ser útil para determinação de ácidos orgânicos de vinagres de ameixa de forma rápida e precisa.

O SPA foi usado para selecionar variáveis em medidas espectrométricas VIS/NIR na construção de modelos para determinação do teor sólidos solúveis em cerveja. Para efeito de comparação mais dois métodos de seleção de variáveis foram utilizados, a análise de coeficientes de regressão PLS e a análise de componentes independentes (*Independent Component Analysis* - ICA)^[39]. As variáveis foram selecionadas por SPA, RCA e ICA, e com elas foram construídos modelos de regressão, lineares (PLS e MLR) e não-lineares usando LS-SVM. Dez variáveis foram selecionadas pelo SPA e o modelo que alcançou o melhor desempenho linear foi o SPA-MLR, em relação ao SPA-PLS, RCA-MLR, RCA-PLS, ICA-MLR e PLS. O coeficiente de correlação e o RMSEP do SPA-MLR foi 0,9762 e 0,1808 °Brix respectivamente. Entre os modelos não-lineares, o SPA-LS-SVM obteve desempenho semelhante aos modelos RCA-LS-SVM e ICA-LS-SVM, o valor do coeficiente de correlação e o RMSEP foram 0,9818 e 0,1628 °Brix, respectivamente. Os resultados globais mostraram que o SPA foi bem sucedido para a seleção de variáveis na determinação de sólidos solúveis em cerveja.

Um método foi introduzido para a determinação quantitativa do teor de proteínas em amostras de iogurte com base na absorbância característica na região espectral de 1800-1500 cm⁻¹ por espectrometria no infravermelho médio com transformada de Fourier MID-FTIR. Nesse caso, o SPA selecionou variáveis para regressão usando redes neurais artificiais treinadas por *back propagation*

(*Back Propagation Artificial Neural Network* - BP-ANN)^[40]. A vantagem de usar esse método está no fato de que as relações com a concentração não são sempre lineares^[40]. Desse modo as previsões realizadas pelo método BP-ANN foram comparadas com o mesmo método levando em conta apenas as variáveis selecionadas pelo SPA. Os erros relativos de previsão (REP) dos métodos BP-ANN e SPA-BP-ANN para o conjunto de calibração foram de 7,25 % e 3,70 %, e em um conjunto de previsão independente 9,32 % e 6,98 %, respectivamente. Deste modo os modelos SPA-BP-ANN, podem ser usados de forma rápida e simples, para a determinação de proteínas.

No contexto de classificação, Pontes et al.^[2] propuseram uma nova metodologia analítica para classificação de solos usando a espectroscopia de emissão em plasma induzido por laser e técnicas quimiométricas. Subconjuntos de variáveis espectrais selecionadas por três técnicas diferentes, SPA, GA, e Stepwise (*Stepwise* - SW) foram usadas para a construção de modelos usando LDA, para efeito de comparação a modelagem independente e flexível por analogia de classe (*Soft Independent Modeling of Class Analogy* - SIMCA) também foi empregada. Com objetivo de reduzir a carga computacional um processo de compressão de dados no domínio wavelet foi proposto. A metodologia foi validada em um estudo de caso envolvendo a classificação de 149 amostras de solos brasileiro em três diferentes ordens (Argissolo, Latossolo e Nitossolo). A melhor discriminação dos tipos de solos foi atingida pelo SPA-LDA, que alcançou uma taxa de classificação média de 90% em um conjunto de validação separado e 72% usando validação cruzada. A proposta do processo de compressão wavelet proporcionou uma redução de 100 vezes na carga computacional, sem comprometer significativamente a exatidão da classificação dos modelos resultantes.

Medidas eletroanalíticas realizadas usando a voltametria de onda quadrada (*Square Wave Voltammetry*-SWV), que são medidas simples e não-dispendiosas, foram usadas para a classificação de óleos vegetais comestíveis com relação ao tipo (canola, girassol, milho e soja) e estado de conservação (vencido e não vencido)^[41]. O SIMCA e o LDA com seleção de variáveis pelo SPA foram comparados para a classificação dos voltamogramas resultantes. Os resultados foram avaliados em termos de erros em um conjunto de amostras não incluídas no processo de modelagem. Os melhores resultados foram obtidos com o método

SPA-LDA, que classificou corretamente todas as amostras utilizando apenas 10 variáveis.

A espectrometria NIR foi utilizada para determinar o teor de proteína de *Auricularia aurícula* usando o PLS, e o SPA aplicado para selecionar as variáveis eficazes^[42] para a regressão MLR e LS-SVM. As combinações de vários pré-tratamentos e métodos de calibração foram comparados com base no desempenho de previsão, e o melhor modelo PLS usando os espectros completos foi alcançado através dos espectros brutos, enquanto os modelos ideais usando SPA-MLR, SPA-PLS e SPA-LS-SVM foram obtidos com o tratamento MSC (*Multiplicative Scatter Correction*). O melhor desempenho de previsão foi realizado pelo modelo SPA-LS-SVM, com coeficientes de correlação, 0,9839 e um RMSEP igual a 0,16%. Os resultados mostraram que a espectrometria NIR combinada com SPA-LS-SVM foi capaz de determinar adequadamente o teor de proteínas na *Auricularia aurícula*.

As determinações da quantidade de ferro e zinco no conteúdo de leite em pó^[43] foram realizadas a partir de modelos LS-SVM baseados nas variáveis selecionadas pelo método UVE combinado com SPA. Os modelos foram construídos através de medidas espectrométricas NIR e MIR, sendo que os modelos que usaram as variáveis selecionadas no MIR foram melhores que os modelos construídos com as variáveis selecionadas no NIR. Os resultados de previsão do conteúdo de ferro com o modelo UVE-SPA-LS-SVM com 18 variáveis MIR produziram coeficiente de correlação de 0,920 e o desvio de previsão residual (*Residual Predictive Deviation* - RPD) de 3,321, e RMSEP 1,444 mg/100g. Para a previsão do teor de zinco 12 variáveis MIR produziram coeficiente de correlação 0,946, e RPD igual a 4,361, o RMSEP foi 0,321 mg/100g. O bom desempenho mostrou que UVE-SPA é uma poderosa ferramenta de seleção de variáveis, e a espectroscopia MIR incorporada ao UVE-SPA-LS-SVM pode ser aplicada como uma alternativa rápida e precisa para determinar o teor mineral de ferro e zinco em leite em pó.

Descritores físico-químicos foram calculados para 152 moléculas por meio do software Dragon, com objetivo de construir modelos para prever a atividade biológica de uma série de inibidores de glicogênio sintase quinase-3 β (*Glycogen Synthase Kinase* - GSK-3 β). Três métodos diferentes de seleção de variáveis foram utilizados, GA, SPA e um método conhecido como otimização por colônia de formigas (*Ant Colony Optimization* - ACO) que ainda não tinha sido aplicado a

dados químicos. Os modelos construídos com base no conjunto de descritores escolhido por meio dos algoritmos de seleção de variáveis foram relacionados com a atividade biológica^[44], através de métodos de regressão lineares (MLR) e não lineares (ANN e SVM). O modelo ACO-SVM forneceu a melhor previsão, menor RMSEP, indicando assim o caráter não-linear desta análise e introduzindo o ACO como um método melhorado de seleção de variáveis em QSAR para a classe de GSK-3 β inibidores. A aplicação do SPA permitiu a previsão satisfatória, porém quando comparado ao ACO o valor do RMSEP foi um pouco maior.

A classificação de cigarros em diferentes marcas de fabricantes foi realizada empregando espectrometria de refletância difusa NIR e modelos SPA-LDA^[45]. Desta maneira, o SPA foi empregado para escolher um subconjunto adequado de números de onda para o LDA. Para efeito comparativo o método SIMCA também foi utilizado. Os modelos foram obtidos a partir de um conjunto de 210 amostras de cigarros de quatro marcas diferentes. O SPA-LDA classificou todas as amostras com sucesso em relação a suas marcas utilizando apenas dois números de onda. Em contraste, os modelos SIMCA não foram capazes de atingir 100% de exatidão na classificação, independentemente do nível de significância adotado para o teste F.

Com objetivo de prever os coeficientes de adsorção de alguns pesticidas (K_{OC}), o estudo da relação quantitativa estrutura-propriedade (*Quantitative Structure-Property Relationship* - QSPR) foi realizado^[46]. A modelagem da relação entre os descritores moleculares selecionados e os coeficientes de adsorção, foram realizadas pelos métodos de regressão, linear (MLR) e não-linear (ANN). Os descritores usados para o desenvolvimento do modelo foram selecionados pelo SPA. Os modelos QSPR foram validados por meio da validação cruzada, bem como a aplicação dos modelos para prever o K_{OC} de compostos de referência externa, que não contribuem para modelar as etapas de desenvolvimento. Ambos os métodos lineares e não lineares promoveram previsões condizentes com os resultados esperados, apesar de se obter melhores resultados por meio de modelos SPA-ANN. Os erros quadráticos médios do conjunto de previsão obtidos pelos modelos SPA-MLR e SPA-ANN foram 0,3705 e 0,2888, respectivamente.

O SPA foi utilizado para selecionar variáveis em espectros de Ressonância Magnética Nuclear de Hidrogênio (*Nuclear Magnetic Resonance* - NMR) para classificação de óleo de canola por marcas^[47]. Os modelos foram avaliados por

dois parâmetros, taxa de resposta correta (*Correct Answer Rate* - CAR) e o coeficiente kappa de Cohen (K). A aplicação direta do SPA no espectro só melhorou um pouco a taxa de resposta correta (CAR) e o coeficiente kappa de Cohen (K) em relação ao modelo aplicado aos espectros completos usando LS-SVM. Dessa maneira dois processos de eliminação de variáveis foram usados, o UVE e o UVE combinado com o algoritmo *Simulated annealing* (Simulated annealing -SA) para estimar o ponto de corte ótimo do UVE. Os resultados mostram que a discriminação UVE-SA funcionou melhor do que UVE convencional. Somente 13 variáveis foram selecionadas pelo UVE-SA-SPA, enquanto o UVE-SPA escolheu 77 variáveis. O modelo UVE-SA-SPA-LS-SVM forneceu os melhores valores de CAR e K (97,5% e 0,967) e mostram que é possível distinguir as diferentes marcas de óleos de canola com espectros ^1H NMR. Dessa maneira uma combinação de SA, UVE e SPA fornecem um método eficaz para a classificação de óleos de canola.

Para determinar o grau de polimerização de papel isolante para transformadores com maior rapidez e menor custo por análise, um método alternativo usando espectrometria NIR^[48], foi desenvolvido. Setenta e cinco amostras de papel kraft, crepe e papelão, em diferentes estágios de degradação foram coletados em transformadores durante um período de três anos. O conjunto de amostras foi analisado de acordo com a norma ABNT NBR 8148, para obter os valores de referência e serem usados para estabelecer uma relação quantitativa com espectros medidos por reflectância difusa. O modelo SPA-MLR produziu resultados aceitáveis para calibração quando combinado com a espectrometria NIR, contudo quando essas medidas foram usadas no PLS seus resultados produziram erros menores em um conjunto de validação externa.

Wu et al.^[49] usou as variáveis selecionadas pelo SPA objetivando a construção de modelos para a determinação não invasiva do teor de sólidos solúveis em suco de uva e pH, a partir de medidas espectrométricas VIS-NIR. A fim de eliminar variáveis não informativas os métodos, UVE, iPLS, e duas modificações do iPLS, o siPLS (*synergy interval Partial Least Squares* - siPLS) e o biPLS (*backward interval Partial Least Squares* - biPLS) foram utilizados. Após a aplicação desses métodos o SPA selecionou variáveis para a regressão por meio do PLS e o MLR. Os modelos PLS usando todo espectro e os obtidos pelas variáveis selecionadas pelo SPA, iPLS, siPLS (usando dois, três e quatro intervalos), UVE e UVE-SPA foram utilizados. Os modelos MLR foram obtidos

usando somente as variáveis selecionadas pelo SPA e UVE-SPA. O modelo que forneceu o melhor resultado foi o UVE-SPA-MLR, em ambas as análises de pH e sólidos solúveis totais. Os coeficientes de correlação para o conjunto de previsão e o desvio de previsão residual (RPD), obtido por UVE-SPA-MLR são 0,979 e 6,971 °Brix para sólidos solúveis totais, e 0,951 e 5,432 para pH. Os resultados globais mostram que é viável determinar de forma não-invasiva teor de açúcar solúvel e pH em suco de uva utilizando a espectroscopia VIS-NIR e UVE-SPA para seleção de variáveis em modelos MLR.

Uma implementação de regressões seqüenciais para o SPA usando regressão MLR foi proposta por Soares et al.^[50]. A melhoria computacional promovida foi ilustrada por um exemplo envolvendo a determinação de proteína em trigo por espectrometria NIR. As previsões do modelo resultante exibiram um coeficiente de correlação de 0.989 e um RMSEP de 0.2% m/m na faixa de 10.2-16.2 % m/m. A implementação proposta foi de grande valia na promoção de ganhos computacionais de até cinco vezes.

Com objetivo de melhorar a eficiência computacional, tanto do SPA-MLR como de outras técnicas quimiométricas sofisticadas (GA-MLR, GA-LDA e PLS), Soares et al.^[51] apresentaram formas de paralelizar a implementação de diversos algoritmos mediante pequenas modificações em códigos quimiométricos já existentes, tornando possível explorar os benefícios do processamento em múltiplos núcleos por processadores Multi-Core. Os códigos modificados foram aplicados na determinação de proteína em trigo por espectrometria de reflectância NIR e classificação de óleos vegetais comestíveis por voltametria de onda quadrada. Empregando a computação paralela em processadores multi-core, ganhos computacionais de até 204% foram obtidos.

O SPA foi combinado com a técnica de sub-amostragem e agregação de modelos (*subagging*) de modo a realizar transferência de calibração^[52]. Nesse caso os modelos MLR devem ser robustos a diferenças instrumentais entre dois equipamentos (primário e secundário). Para guiar a seleção, um pequeno número de amostras de transferência, com espectros adquiridos no instrumento primário e secundário, foram empregados. Demonstrou-se a viabilidade desse trabalho a partir da determinação por espectrometria NIR de massa específica, T10% e T90% em amostras de gasolina e umidade, em amostras de milho por espectrometria NIR de reflectância difusa. O erro de previsão em termos de RMSEP para o equipamento secundário foi usado para avaliar os modelos. Os

modelos MLR gerados pela abordagem SPA-subagging forneceram resultados melhores que os obtidos por PLS-PDS. Em particular, o uso de subagging resultou em uma redução mais sistemática no erro de previsão com a inclusão progressiva de amostras de transferência.

A seleção de variáveis realizada pelo SPA em espectrometria UV-VIS com objetivo de obter um modelo LDA foi usado na classificação de extratos de cafés brasileiros com relação ao tipo (cafeinado e descafeinado) e estado de conservação (vencido ou não-vencido)^[53]. Para efeito de comparação a modelagem SIMCA foi realizada. Os melhores resultados foram obtidos pelo modelo SPA-LDA, que classificou corretamente todas as amostras, enquanto o SIMCA obteve 27 erros. A exatidão da classificação deste modelo manteve-se elevada (96%), mesmo após a introdução do ruído espectral artificial.

O SPA foi usado na seleção de variáveis para a construção de modelos MLR em dados espectrométricos NIR, e comparado com os modelos PLS produzidos pelas variáveis selecionadas por meio de perturbações na concentração de amostras líquidas usando espectroscopia 2D e um sistema em fluxo^[54]. Modelos PLS, a partir de variáveis selecionadas com base na estabilidade dos coeficientes de regressão e usando todas as variáveis também foram usados^[54]. A avaliação foi realizada por meio da determinação de parâmetros de qualidade da gasolina (Teor de etanol, MON (*Motor Octane Number* - MON) e RON (*Research Octane Number* - RON)). Os modelos gerados usando as variáveis selecionadas pelo SPA para determinação de MON e RON produziram resultados semelhantes aos outros modelos. Na determinação de etanol as amostras de previsão foram contaminadas por tolueno, e percebe-se que tanto o SPA-MLR como os outros métodos não produziram uma boa previsão. Esses resultados comprovam que quando interferentes desconhecidos aparecem nas amostras de previsão, os resultados das análises podem ser comprometidos significativamente.

1.2. Objetivos

O presente trabalho teve como objetivo promover a melhoria da robustez por meio da mudança na função custo usada na escolha das variáveis pelo algoritmo das projeções sucessivas, tornando possível a determinação de parâmetros de interesse de forma mais exata, por meio da redução da influência dos interferentes quando presentes.

1.2.1. Objetivos específicos

- Usar o erro de previsão estatística (*Statistical Prediction Error* - SPE)^[9] em conjunto com o RMSE como critério de escolha do melhor modelo usando o SPA;
- Aplicar o SPA com o novo critério na determinação de analitos simulados, corantes alimentícios e álcool em gasolina, e comparar com o SPA sem tal critério e com o PLS usando todas as variáveis.

CAPÍTULO 2

Fundamentação Teórica

2. CALIBRAÇÃO MULTIVARIADA

A medida apenas de uma variável em muitos casos não fornece uma boa relação com a resposta devido a influências que o sinal medido pode sofrer de outras espécies, gerando assim previsões inadequadas^[6]. Dessa forma um modelo de calibração univariada só pode fornecer resultados precisos, se o sinal medido não tiver influência de outras fontes^[55]. O uso de muitas variáveis é capaz de reduzir a quantidade de ruído, melhorando a precisão e fornecendo a capacidade de determinação simultânea de muitas propriedades de uma amostra^[55]. Diante dessas vantagens a calibração multivariada tornou-se uma ferramenta indispensável para a determinação quantitativa em química analítica.

Um modelo de calibração multivariada pode ser obtido basicamente dividindo o processo em duas etapas^[7,10,56]. Na primeira procura-se estabelecer uma relação matemática entre as variáveis medidas através de métodos rápidos (Matriz X_{cal}) com o vetor que possui os parâmetros de interesse determinados pelos métodos de referência (y_{cal}). Essa relação matemática existente entre a matriz X_{cal} e o vetor y_{cal} é conhecida como modelo de calibração. Na **Figura 2.1** é mostrado um esquema das principais informações necessárias à construção de modelos multivariados.

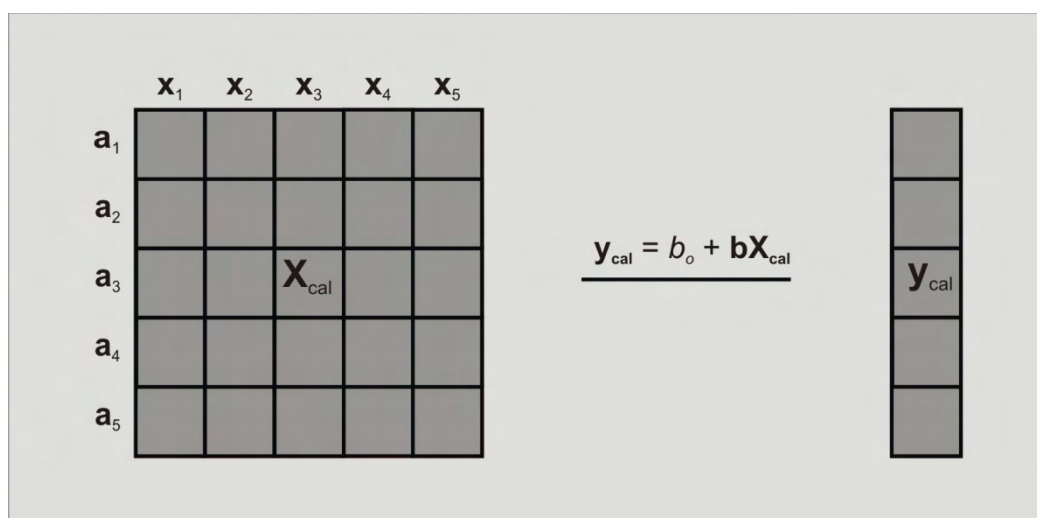


Figura 2.1.- Principais matrizes envolvida no processo de calibração.

Tendo construído o modelo de calibração, em uma segunda etapa é preciso verificar se a relação existente entre a matriz X_{cal} e o vetor y_{cal} é satisfatória para a determinação da propriedade de interesse, essa etapa

conhecemos como validação do modelo. A validação geralmente é realizada de duas formas diferentes:

1. São usados subconjuntos de amostras do conjunto de calibração^[6] para construir o modelo e as amostras restantes formam um subconjunto que é usado para testar a validade do modelo de calibração. Esse tipo de estratégia é conhecida como validação cruzada (*Cross-validation*)^[57];
2. A matriz $\mathbf{X}$ é dividida em amostras de calibração e amostras de validação. Esse procedimento é conhecido como validação externa por serie de teste, e esses novos conjuntos são conhecidos como conjuntos de validação $\mathbf{X}_{\text{val}}$ e $\mathbf{y}_{\text{val}}$.

Para entender como se faz uma validação cruzada, considere as amostras medidas ao longo da coluna $\mathbf{x}_1$ para as amostras $\mathbf{a}_1$ até $\mathbf{a}_5$ relacionando-se com os parâmetros de referência $\mathbf{y}_{\text{cal}}$ (**Figura 2.1**). Um modelo com as amostras de $\mathbf{a}_2$ até $\mathbf{a}_5$ é construído, logo em seguida a amostra $\mathbf{a}_1$ é prevista usando este modelo. Em outra etapa, com as amostras $\mathbf{a}_1$, $\mathbf{a}_3$, $\mathbf{a}_4$ e $\mathbf{a}_5$ um novo modelo é construído, da mesma forma a amostra remanescente $\mathbf{a}_2$ é prevista. Esse procedimento é repetido até que todas as amostras tenham sido previstas. Após isso os valores previstos são comparados com os valores de $\mathbf{y}_{\text{cal}}$. No exemplo foi ilustrado somente uma amostra ficando fora do modelo (*leave-one-out*), porém mais de uma amostra pode ficar fora, de uma forma sistemática ou aleatória.

A comparação entre os valores previstos e os valores de referência requer que sejam usadas algumas métricas que permitam avaliar se os valores previstos a partir das medidas $\mathbf{X}$ são condizentes com os de $\mathbf{y}$. Entre as métricas que dão idéia da dimensão do erro de previsão dos modelos as mais usadas são o PRESS (*Predicted Residual Error Sum of Squares*), RMSE (*Root Mean Squares Error*) e o RSEP (*Relative Standard Error of Prediction*).

O PRESS^[7,9,57] é definido como sendo o somatório do quadrado dos resíduos, que são obtidos subtraindo de $\mathbf{y}$ (valor dado pelo método de referência) os valores de $\hat{\mathbf{y}}$ (valor previsto pelo modelo), como mostrado na **Equação 1**:

$$PRESS = \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (1)$$

O MSE (*Mean Squared Error*), obtido dividindo-se o PRESS pelo número de amostras previstas, também é usado habitualmente. Uma desvantagem de se usar o PRESS ou o MSE é que os seus valores não são obtidos nas unidades originais dos valores de referência^[6]. Para resolver esse problema recorre-se ao uso do RMSE que é obtido a partir da raiz quadrada do MSE.

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

Um bom modelo de calibração deverá produzir um pequeno valor de RMSE^[6]. O RMSE pode ser calculado tanto para as N_c amostras de calibração, RMSEC (*Root Mean Squared Error of Calibration*) quanto para as N amostras de previsão, RMSEP (*Root Mean Squared Error of Prediction*). Entretanto o cálculo do RMSEC difere em relação à **Equação 2**, nesse caso ao invés de N são considerados $N_c - k - 1$, sendo k o número de fatores utilizados. Quando a validação do modelo é feita por validação cruzada o RMSE é conhecido por RMSECV e nesse caso o que muda é o método de obtenção de $\hat{y}$.

O erro padrão de previsão (*Standard Error of Prediction* - SEP)^[6] vem sendo usado para medir o desempenho em diversos trabalhos, e seu cálculo é realizado como mostrado na **Equação 3**.

$$SEP = \sqrt{\sum_{i=1}^N \frac{(y_i - \hat{y}_i - BIAS)^2}{N-1}} \quad (3)$$

e o $BIAS = \sum_{i=1}^N (\hat{y}_i - y_i) / N$ é obtido.

Uma métrica que tem sido utilizada para a avaliação do modelo é o desvio de previsão residual (*Residual Predictive Deviation* - RPD)^[49] que é o desvio padrão do vetor y para as amostras de validação, dividido pelo SEP. De um modo geral essa métrica classifica os modelos para o uso segundo a **Tabela 2.1**^[58].

Tabela 2.1 - Classificação do uso dos modelos a partir dos valores de RPD.

Valor de RPD	Aplicação
0.0-2.3	Não recomendado
2.4-3.0	Triagem Grosseira
3.1-4.9	Triagem
5.0-6.4	Controle de qualidade
6.5-8.0	Controle de Processo
8.1+	Todas as aplicações

Comumente outras ferramentas como os gráficos de y de referência *versus* os valores de $\hat{y}$ previsto e o de resíduos das amostras deixadas pelo modelo de calibração, são utilizados na verificação dos modelos. Uma boa calibração produz um gráfico com as observações aleatórias em volta da bissetriz do gráfico y *versus* $\hat{y}$ (em torno de 45° de cada eixo)^[6,9].

2.1. Classificação dos Métodos de Calibração

A classificação da calibração em univariada ou multivariada depende exclusivamente das dimensões das matrizes envolvidas no processo de calibração, enquanto que a classificação em linear ou não-linear depende da natureza da relação entre a matriz X e a matriz Y . Uma calibração pode ser ainda do tipo direta ou indireta e clássica ou inversa^[59]. Para entender isso será considerado o caso univariado no qual a calibração é feita a partir de uma única variável x e uma propriedade y .

A calibração univariada entre x (sinal medido) e y (propriedade de interesse) pode ser obtida de duas formas:

- (1) Pode-se obter uma relação como sendo $y = b_0 + b_1x + erro$, e a previsão pode ser feita diretamente por $\hat{y} = \hat{b}_0 + \hat{b}_1x$.
- (2) Outra relação entre x e y é $x = a_0 + a_1y + erro$, a previsão nesse caso é realizada invertendo-se a equação, ou seja, $\hat{y} = -(a_0/a_1) + (1/a_1)x$.

Na literatura o caso (1) é conhecido como calibração inversa já o caso (2) é conhecido como calibração clássica^[7]. A razão para isto é histórica, pois o método (2) assume uma relação linear parecida com a que é prevista pela lei de

Beer, que estabelece uma proporcionalidade entre a absorbância e a concentração.

Quando se usa calibração multivariada clássica é necessário obter uma matriz com todos os sinais instrumentais dos analitos puros^[59], tais sinais podem ser medidos ou obtidos indiretamente. Se todos os sinais instrumentais dos analitos puros são medidos a calibração é do tipo direta, caso contrário a calibração é indireta e os sinais dos analitos puros são estimados indiretamente utilizando-se estimativas experimentais da relação entre sinal medido e a propriedade de interesse.

2.2. Regressão Linear Múltipla

A regressão linear múltipla pode ser caracterizada como uma técnica para resolver uma série de equações simultâneas, em sistemas multicomponentes. A análise pode ser descrita por medidas de j variáveis de $\mathbf{X}$ e os parâmetros $\mathbf{Y}$, com o objetivo principal de criar uma relação linear entre elas.

Em calibração usando regressão linear múltipla^[9,10] assume-se que os parâmetros de interesse (matriz $\mathbf{Y}$) relacionam-se linearmente com as respostas instrumentais (matriz $\mathbf{X}$) segundo a **Equação 4**.

$$\mathbf{Y} = \mathbf{XB}_{MLR} + \mathbf{E} \quad (4)$$

Onde,

- $\mathbf{X} \Rightarrow$ matriz dos sinais de m amostras, medidos em j variáveis;
- $\mathbf{Y} \Rightarrow$ matriz dos q parâmetros de m amostras;
- $\mathbf{B}_{MLR} \Rightarrow$ matriz dos coeficientes lineares de regressão.
- $\mathbf{E} \Rightarrow$ matriz dos resíduos que representa a parte não descrita pelo

modelo;

A matriz $\mathbf{B}_{MLR}$ pode ser obtida pelo método dos mínimos quadrados utilizando a **Equação 5**. O método dos mínimos quadrados analisa m diferenças entre cada valor y e o modelo, correspondentes ao respectivo valor $\mathbf{X}$. O modelo selecionado é o que apresenta a menor soma de quadrados de tais diferenças^[60].

$$\mathbf{B}_{MLR} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \quad (5)$$

Todavia, a obtenção da matriz $\mathbf{B}_{MLR}$ pela expressão acima está sujeita aos seguintes problemas:

- A obtenção de um sistema indeterminado quando $j > m$;
- Alta multicolinearidade que ocorre em $\mathbf{X}$. (quando as colunas de $\mathbf{X}$ portam informações redundantes);
- Se $\mathbf{X}$ esta mal condicionada à previsão de $\mathbf{Y}$ está sujeita a propagação de erros^[61].

Para contornar tais problemas no uso do MLR, em medidas que possuem muitas variáveis, recorre-se aos métodos de seleção de variáveis ou ao uso da regressão a partir das variáveis transformadas não colineares (PCR ou PLS).

2.3. Regressão em Componentes Principais

A regressão em componentes principais está baseada no conceito de análise de componentes principais (*Principal Component Analysis* - PCA)^[61]. A PCR estima os escores $\mathbf{T}$ a partir dos dados originais $\mathbf{X}$ e em seguida relaciona linearmente $\mathbf{T}$ com a propriedade a ser determinada.

Para estimar a matriz de escores $\mathbf{T}$ é realizada uma decomposição de $\mathbf{X}$ em componentes principais. Os algoritmos dos mínimos quadrados parciais iterativos não-lineares (*Nonlinear Iterative Partial Least Squares* - NIPALS) e a decomposição por valores singulares (*Singular Value Decomposition* - SVD) têm sido frequentemente utilizados para realizar esses cálculos.

A primeira componente principal, $\mathbf{t}_1$, é calculada a partir da combinação linear das colunas de $\mathbf{X}$ que explique a maior variância possível. O vetor que define a combinação linear $\mathbf{p}_1$ é conhecido como *loading* e é escalonado de modo a ter norma igual a 1. A segunda componente principal $\mathbf{t}_2$ é definida da mesma maneira, sendo que $\mathbf{t}_1$ é perfeitamente não-correlacionada com $\mathbf{t}_2$ e igualmente a direção do segundo vetor é definido por $\mathbf{p}_2$. O processo continua até que o número de componentes principais seja igual a A (número de componentes escolhidos). Geralmente, quando a quantidade de amostras de $\mathbf{X}$ é maior que o número de variáveis esse processo continua até que A seja igual ao número de

variáveis de $\mathbf{X}$. Caso contrário, a quantidade de amostras torna-se o limitante para A .

Em termos matriciais, a matriz que contém todos os escores das componentes de 1 até A é tida como $\mathbf{T}$ e a matriz que contém os *loadings* é conhecida como $\mathbf{P}$. Uma maneira de representar a matriz $\mathbf{X}$ centrada na média é:

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} \quad (6)$$

Sendo $\mathbf{E}$, na **Equação 6**, a parte do resíduo deixado pela modelagem.

Tendo calculado $\mathbf{T}$, agora é possível obter uma equação de regressão entre a propriedade a ser determinada $\mathbf{y}$ e a matriz de escores $\mathbf{T}$. A relação existente entre essas duas propriedades pode ser descrita pela **Equação 7**.

$$\mathbf{y} = \mathbf{T}\mathbf{q} + \mathbf{f} \quad (7)$$

Os coeficientes de regressão $\mathbf{q}$ são estimados a partir do método dos mínimos quadrados. Com o valor de A igual ao número máximo de variáveis da matriz $\mathbf{X}$, o método PCR passa a se comportar de maneira idêntica ao método MLR^[6].

2.4. Regressão por Mínimos Quadrados Parciais

Em PLS os escores de $\mathbf{X}$ são os preditores de $\mathbf{y}$, e as matrizes $\mathbf{y}$ e $\mathbf{X}$ assumem uma relação linear que é modelada pelas variáveis latentes das mesmas, essas são estimadas como combinações lineares das variáveis originais. Para realizar uma previsão significativa um pequeno número de variáveis latentes é necessário^[62].

Os coeficientes das combinações lineares são representados por $\mathbf{W}$ e são conhecidos como *loading weights*^[7]. A relação abaixo mostra em termos matriciais como pode ser representado.

$$\mathbf{T} = \mathbf{XW} \quad (8)$$

Os escores de $\mathbf{X}$ são multiplicados pelos *loadings* $\mathbf{P}$ de modo que os resíduos $\mathbf{E}_x$, na **Equação 9**, sejam pequenos.

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E}_x \quad (9)$$

Da mesma forma, $\mathbf{u}$ os escores de $\mathbf{y}$ são multiplicados por $\mathbf{c}$ (*loadings* de $\mathbf{y}$) de modo que os resíduos $\mathbf{e}_y$, na **Equação 10**, sejam pequenos.

$$\mathbf{y} = \mathbf{uc}^t + \mathbf{e}_y \quad (10)$$

Os escores de $\mathbf{X}$ são usados para a previsão das propriedades de $\mathbf{y}$ a partir da relação dada pela **Equação 11**.

$$\mathbf{y} = \mathbf{Tc}^t + \mathbf{F} \quad (11)$$

Os resíduos, $\mathbf{F}$ expressam os desvios entre as respostas observadas e modeladas. Usando as **Equações 8 e 11**, pode-se obter a seguinte relação para um modelo de regressão multivariado:

$$\mathbf{y} = \mathbf{XWc}^t + \mathbf{F} = \mathbf{Xb} + \mathbf{F} \quad (12)$$

Os coeficientes de regressão PLS, $\mathbf{b}$, podem ser escritos como:

$$\mathbf{b} = \mathbf{Wc}^t \quad (13)$$

Quando se tem mais de um parâmetro $\mathbf{y}$ ($\mathbf{y}_1, \mathbf{y}_2$) todo o procedimento pode ser reescrito considerando os vetores $\mathbf{y}_1$ e $\mathbf{y}_2$ em uma matriz $\mathbf{Y}$, desse modo as equações acima podem ser utilizadas fazendo as devidas considerações de dimensionalidade.

Os pesos $\mathbf{W}$ e $\mathbf{c}$ dão informações a respeito de como as variáveis se combinam para formar a relação quantitativa entre $\mathbf{X}$ e $\mathbf{y}$, proporcionando assim uma interpretação dos escores, $\mathbf{T}$ e $\mathbf{u}$. Assim, esses pesos são indispensáveis para uma compreensão de quais variáveis de $\mathbf{X}$ são numericamente importantes e quais são as variáveis que fornecem informações similares.

Em calibração multivariada usando PLS até a parte dos dados que não são explicados pelo modelo, os resíduos, são de grande interesse para o diagnóstico.

Grandes resíduos em y podem indicar que o modelo não é satisfatório para a determinação. Em PLS os resíduos de X , a parte não utilizada na modelagem de y , são úteis para identificar outliers no espaço de X e processos desviando das operações normais^[62].

2.5. Técnicas de Seleção de Variáveis

As técnicas de seleção de variáveis são baseadas no princípio de escolha de um pequeno número de variáveis selecionadas a partir dos dados originais ou dados transformados, sendo uma etapa muito importante, capaz de remover variáveis não informativas e minimizar a multicolinearidade que atrapalha a regressão. Desse modo tem sido demonstrado que a capacidade preditiva pode ser aumentada e, a complexidade do modelo pode ser reduzida por uma boa pré-seleção de comprimentos de onda^[3]. É agora amplamente aceito que a seleção de variáveis bem executadas pode resultar em modelos que têm uma maior capacidade preditiva^[63].

Existem muitas estratégias disponíveis para a seleção de variáveis, desde abordagens clássicas, métodos seqüenciais, métodos baseados em estratégia de busca e métodos mais sofisticados, como, o algoritmo das projeções sucessivas e os métodos de eliminação de variáveis não informativas^[3]. Desta maneira serão detalhados os principais métodos usados no contexto da química analítica e alguns com objetivo de elucidar as vantagens e desvantagens de alguns métodos.

2.5.1. Busca Exaustiva

A busca exaustiva consiste em avaliar todas as combinações possíveis das variáveis disponíveis para a regressão. Como, se pode imaginar, a aplicabilidade dessa estratégia pode ser comprometida pelo grande esforço computacional requerido. Com efeito, o número de combinações a serem testadas pode ser muito elevado, mesmo para problemas de dimensão relativamente pequena. Galvão et al.^[64] mostram que um modelo para a determinação simultânea de três analitos com sobreposição e 60 variáveis, pode gerar aproximadamente 400 milhões de possibilidades de combinações. Por isso muitos algoritmos foram desenvolvidos como alternativa à busca exaustiva.

2.5.2. Stepwise Regression

O método de seleção de variáveis *stepwise regression* está baseado nos princípios de *forward selection* e *backward elimination*^[64].

No método *forward selection*^[64] o modelo parte de uma variável x_1 (variável que tem maior correlação com a resposta) e vai adicionando progressivamente mais variáveis ao modelo. Após a adição de cada variável, é realizada uma avaliação parcial com base e um teste F. Esse teste usa o cálculo da soma quadrática residual do novo modelo para calcular um valor de F (F_{cal}). A variável que fornecer o maior valor de F_{cal} permanece no modelo. Semelhante ao *forward selection*, o *backward elimination* parte de todas as variáveis disponíveis e vai retirando variáveis. Novamente, o efeito de cada eliminação é avaliado com base em um teste F e, a variável que fornecer o menor valor de F_{cal} é removida do modelo. Em ambos os casos (*forward selection* e *backward elimination*) o processo continua até que não haja variáveis com valores de F_{cal} maiores que um valor crítico (F_{critco}) tabelado ou obtido empiricamente para um determinado nível de confiança e graus de liberdade.

O algoritmo resultante adiciona progressivamente novas variáveis ao modelo, a partir da variável x_i que guarda maior correlação empírica com a variável dependente y , como no método de seleção *forward*. Para evitar que a variável quando incorporada ao modelo possa ser reavaliada em interações subseqüentes, o algoritmo incorpora um mecanismo para a eliminação de variáveis como no método de eliminação *backward*. Cada iteração do processo de seleção inclui uma fase de inclusão seguida por uma fase de exclusão.

Na fase de inclusão, cada uma das variáveis restantes é submetida a um teste F parcial, como na seleção *forward*. Se o maior valor de F obtido desta forma, é maior do que um valor crítico de entrada ($F_{entrada}$), a variável do valor correspondente é inserida ao modelo. Na fase de exclusão, cada uma das variáveis no modelo está sujeito a um teste F parcial, como no método de eliminação *backward*. Se o menor valor de F assim obtido é menor do que um valor crítico de saída ($F_{saída}$), o valor da variável correspondente é eliminado do modelo, e as variáveis ainda estão disponíveis para a seleção. O processo de seleção pára quando não há variáveis que podem ser adicionadas ou removidas do modelo de acordo com os critérios do teste F parcial^[64].

2.5.3. Método de eliminação de variáveis não-informativas

O método de eliminação de variáveis não-informativas (*uninformative variable elimination* – UVE) é fundamentado na estabilidade dos coeficientes de regressão PLS^[65] avaliada por meio da comparação de coeficientes de variáveis medidas e variáveis simuladas. As etapas do UVE podem ser resumidas da seguinte maneira:

- O modelo PLS ideal (número de variáveis latentes - A) em X com o menor RMSEP como critério é determinado;
- Variáveis são simuladas por um gerador aleatório e multiplicadas por uma pequena constante. Isso produz a matriz $R_{n \times p}$ com o número de variáveis p iguais ao número de variáveis em X . A probabilidade a priori para cometer um erro na seleção, ou seja, para eliminar uma variável informativa ou de manter uma não-informativa é a mesma em X e R .
- Junção de $R_{n \times p}$ com $X_{n \times p}$. A matriz resultante é chamada $XR_{n \times 2p}$, as primeiras p colunas sendo de X e as p últimas sendo de R ;

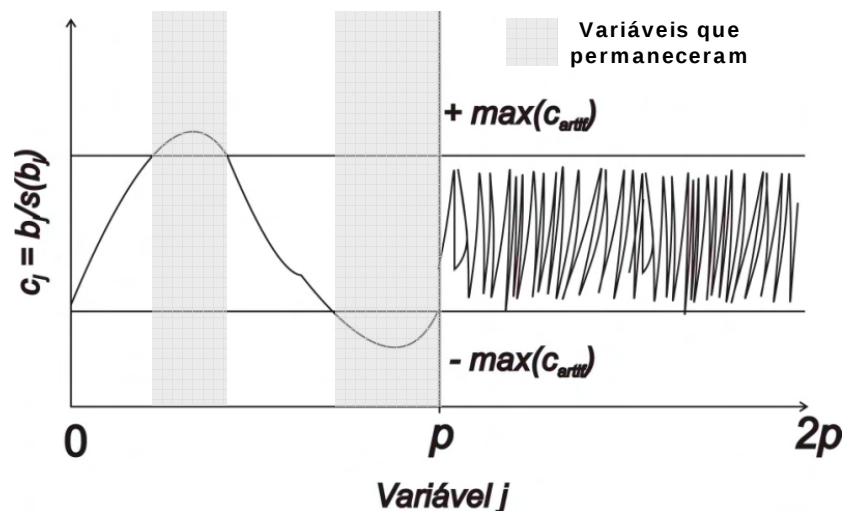


Figura 2.2 - Exemplo do processo de eliminação de variáveis.

- Cálculo dos modelos PLS para a nova matriz XR de acordo com o procedimento *leave-one-out* (têm-se n modelos). O número de fatores (A) é o mesmo que foi determinado para X . Estes n modelos PLS são gerados,

cada um com $2p$ coeficientes de regressão (b), armazenados em uma matriz $\mathbf{B}_{n \times 2p}$;

- Determinação para cada variável j (tanto experimentais como as geradas aleatórias), da média ($\bar{b}_j = \sum_{i=1}^n \frac{b_{ij}}{n}$) e desvio padrão ($s(b_j) = \sqrt{\sum_{i=1}^n \frac{(b_{ij} - \bar{b}_j)^2}{n-1}}$);
- Cálculo para cada variável j do critério $c_j = \bar{b}_j / s(b_j)$;
- Determinação do $\max(\text{abs}(c_{\text{artif}}))$, o maior valor absoluto de c para todas as variáveis artificiais, quando $j > p$;
- Eliminação das variáveis de $\mathbf{X}$ para o qual $\text{abs}(c_j) < \max(\text{abs}(c_{\text{artif}}))$, (para $j = 1, \dots, p$). As demais variáveis constituem a nova matriz $\mathbf{X}$, $\mathbf{X}_{\text{nov0}}$;
- Contruir modelos PLS *leave-one-out* com $\mathbf{X}_{\text{nov0}}$ e A fatores, e com esses modelos, y é previsto;
- Quantificação da capacidade preditiva do novo modelo $\text{RMSEP}_{\text{nov0}}$ de acordo com a **Equação 2**;
- Se o $\text{RMSEP}_{\text{nov0}} > \text{RMSEP}$ conclui-se que a eliminação de variáveis não informativas não melhora a modelagem e o algoritmo é encerrado. Caso contrário, se $\text{RMSEP}_{\text{nov0}} < \text{RMSEP}$, em primeiro lugar procura-se saber se A não é muito grande para evitar sobreajuste do modelo. Para verificar esta possibilidade, o algoritmo começa com uma nova seleção de $\mathbf{X}_R$ a partir do segundo ponto e repete um por um $A = A - 1$ e o RMSEP original passa a ter valor de $\text{RMSEP}_{\text{nov0}}$. Quando a redução de A para $A - 1$ não melhorar a modelagem ($\text{RMSEP}_{\text{nov0}} > \text{RMSEP}$) o algoritmo encerra.

Uma abordagem híbrida entre a UVE e a transformada *wavelet* foi desenvolvida para redução da variação de fundo e do ruído em dados NIR^[66]. O algoritmo consiste em duas partes. Primeiro o algoritmo WP (*wavelet prism*) decompõe o sinal em diferentes componentes de frequência que contêm as mesmas informações do sinal original, e em seguida o critério mUVE é desenvolvido usando a mesma lógica que o UVE. Entretanto, ao invés de selecionar as variáveis para calibração multivariada, o mUVE se propõe a avaliar a importância dos coeficientes *wavelet* com diferentes frequências. O mUVE pode

determinar a escala adequada, em que os coeficientes *wavelet* de aproximação representam a variação de fundo e os coeficientes *wavelet* de detalhe contêm predominantemente a informação do ruído.

2.5.4. Interval PLS

O método iPLS é baseado na divisão dos espectros em intervalos iguais e, em seguida, a construção de modelos de regressão PLS para cada sub-intervalo^[67] com o número de variáveis latentes adequado. O RMSECV para cada sub-intervalo é calculado. O intervalo que fornecer o menor RMSECV é escolhido como o modelo iPLS local ideal para representar um modelo PLS. Porém o método iPLS não testa todas as possibilidades de intervalo, desta maneira outras variantes iPLS, como o siPLS (*synergy interval Partial Least Squares*) e o biPLS (*backward interval Partial Least Squares*) são usados.

No siPLS, calculam-se todas as possíveis combinações de dois, três ou quatro modelos PLS em intervalos. Posteriormente, RMSECV é calculado para cada combinação de intervalos. A combinação de intervalos que produzir o menor RMSECV é o escolhido. O tempo de cálculo é dependente do número de intervalos e o número de combinações. Já no algoritmo biPLS, o modelo PLS é calculado a cada intervalo e em seguida um é deixado fora (por isso a denominação *backward*)^[68]. O intervalo deixado fora é aquele que fornece o modelo com menor desempenho, avaliado pelo RMSECV. Este procedimento continua até que um intervalo permaneça. A combinação ideal é determinada com base no menor RMSECV.

2.5.5. Análise de Componentes Independentes

A ICA^[69] é uma técnica de processamento de sinais que visa separar o sinal observado em componentes independentes (Independent Components – ICs) que são combinações lineares das variáveis de origem. O uso desta técnica pode proporcionar interpretação química e uma maior redução de ruído^[39]. De uma forma geral o modelo ICA é representado por:

$$\mathbf{X} = \mathbf{KS} \quad (14)$$

onde $\mathbf{X}$ é a matriz de sinais, $\mathbf{S}$ são os componentes independente e $\mathbf{K}$ é a matriz de coeficientes. Tendo a matriz de coeficientes $\mathbf{K}$, ela pode ser utilizada para selecionar variáveis. Em^[39] a seleção foi realizada por meio da escolha do comprimento de onda com o maior valor absoluto de $\mathbf{K}$.

2.5.6. Algoritmo Genético

O algoritmo genético (genetic algorithm – GA) é uma técnica que simula matematicamente os mecanismos de seleção natural e teoria da evolução das espécies de Charles R. Darwin^[64,70].

A implementação desse algoritmo para seleção de variáveis é feita através da codificação binária, representada por cada variável como um gene. Genes que tem valor 1 indicam que a variável é incluída no modelo, enquanto genes com valor 0 indicam que as variáveis não entrarão no modelo. O número de variáveis é igual ao número de genes binários e o conjunto desses genes formam os cromossomos. Na **Figura 2.3** é apresentado um esquema de codificação das variáveis.

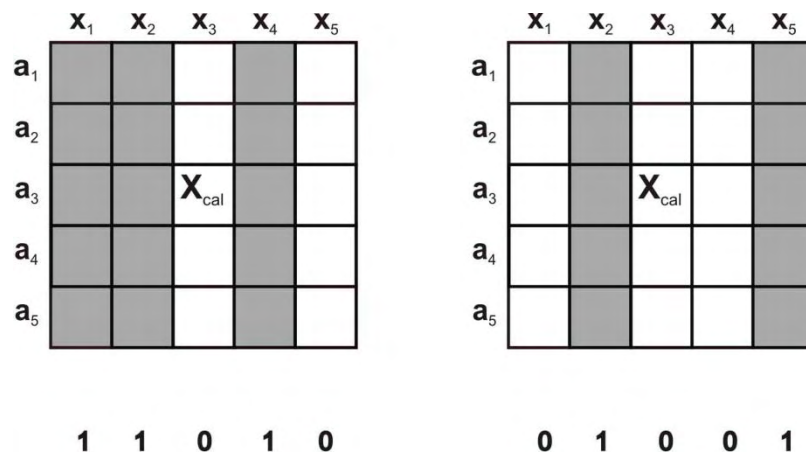


Figura 2.3 – Codificação binária usada na seleção de variáveis.

No início do processo o GA utiliza um gerador randômico para criar uma população inicial de cromossomos, evitando-se a influência tendenciosa na construção dessa população, caso o analista tenha uma previa informação de alguma variável, ele pode incluí-la na população inicial.

A avaliação dos cromossomos é feita com base na aptidão, que é o parâmetro que indicará a habilidade de um indivíduo sobreviver. Matematicamente, quanto maior a aptidão de um indivíduo melhor a resposta

produzida (menor erro). O cálculo da aptidão pode ser realizado construindo modelos MLR ou PLS, baseados nos comprimentos de onda indicados por cada cromossomo. Para isso, o valor da aptidão é calculado como sendo o inverso do PRESS, obtido no conjunto de validação ou para as amostras de calibração se for usado validação cruzada. O subconjunto de variáveis que produzir o menor PRESS (e assim uma maior aptidão) é então adotado como o resultado do GA.

A seleção dos indivíduos que gerarão descendentes em uma nova geração é realizada de forma que os mais aptos tenham maior probabilidade de serem os escolhidos para a reprodução, embora seja necessário dar uma chance aos indivíduos menos aptos, pois eles podem ter características positivas. O procedimento usado para esse fim é conhecido como método da roleta. Neste método a probabilidade de cada indivíduo é dada pela **Equação 15**.

$$P(i) = \frac{Aptidão(i)}{\sum_{i=1}^N Aptidão(i)} \quad (15)$$

Uma nova população é formada cruzando pares de cromossomos aleatoriamente e gerando filhos que possuem material genético dos pais. Esse cruzamento pode ser com ruptura ou por recombinação. Em uma pequena parcela da população é promovido à mutação (que são alterações no código genético) de modo a se obter indivíduos mais aptos. Na representação binária, a mutação é realizada pela troca de 1 por 0 ou vice-versa em um dos genes do cromossomo, como ilustrado na **Figura 2.4**

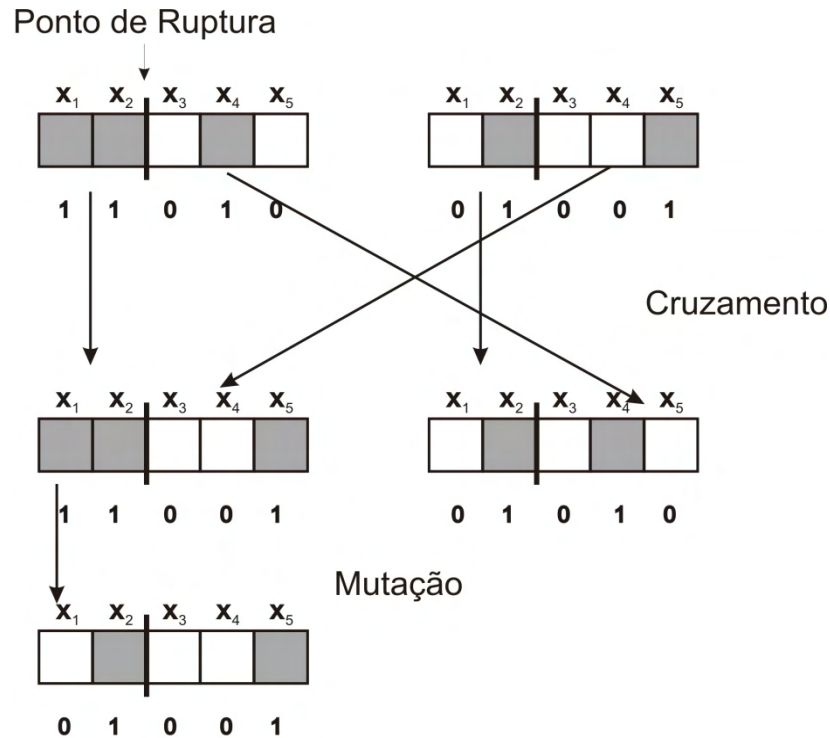


Figura 2.4 - Esquema de cruzamento e de mutação no GA.

Uma das desvantagens desse algoritmo é o fato de ser estocástico, ou seja, ao repetir-se o cálculo dificilmente se obterá o mesmo resultado.

2.5.7. Algoritmo das Projeções Sucessivas

O SPA é uma técnica de seleção *forward* que parte de uma variável x_k e vai incorporando em cada iteração uma nova variável com a menor multicolinearidade possível em relação às variáveis já selecionadas^[13]. Para escolher qual variável a ser incorporada em cada iteração, cálculos de projeções são realizados.

Partindo de cada variável x_k , $k = 1, \dots, K$ a cadeia contendo M variáveis é construída de acordo com as operações^[64]:

Passo 1: (inicialização) faça:

$$z^1 = x_k$$

$$x_j^1 = x_j, j=1, \dots, K$$

$$SEL(1, k) = k$$

$$i = 1$$

Passo 2: Calcular a matriz de projeção P^i no subespaço ortogonal a z^i :

$$\mathbf{P}^i = \mathbf{I} - \frac{\mathbf{z}^i (\mathbf{z}^i)^T}{(\mathbf{z}^i)^T \mathbf{z}^i}$$

onde $\mathbf{I}$ é a matriz identidade de dimensões apropriada.

Passo 3: Calcular os vetores projetados $\mathbf{x}_j^{i+1}$ a partir de:

$$\mathbf{x}_j^{i+1} = \mathbf{P}^i \mathbf{x}_j^i \quad (16)$$

para todos os $j = 1, \dots, K$.

Passo 4: Determinar o índice de j^* do vetor de maior projeção e armazená-lo na matriz **SEL**:

$$j^* = \arg (\max ||\mathbf{x}_j^{i+1}||) \text{ e } \mathbf{SEL}(i + 1, k) = j^*.$$

Passo 5: Fazer $\mathbf{z}^{i+1} = \mathbf{x}_{j^*}^{i+1}$ (vetor que define a próxima operação de projeção)

Passo 6: Fazer $i = i + 1$. Se $i < M$, retorne para o **Passo 2**

A variável de partida $\mathbf{x}_k$ que guarda maior relação com a resposta a ser modelada e o número ideal de variáveis para construir o modelo MLR não é conhecido inicialmente. Assim os subconjuntos de variáveis gerados são avaliados com base no RMSE obtido pela comparação entre os valores previstos e os valores de referência. De uma forma esquemática essa avaliação é realizada da seguinte forma:

Especifique o m_{min} e m_{max} (número máximo e o mínimo de variáveis a selecionar);

De $k = 1$ até K faça

De $m = m_{min}$ até m_{max} faça

- Use as variáveis com índices $SEL(1,k)$, $SEL(2,k)$, ..., $SEL(m,k)$ para construir um modelo MLR. Aplique o modelo para o conjunto de validação e calcule o $RMSE(m,k)$.

Próximo m

Próximo k

A obtenção do RMSE pode ser realizada de duas maneiras:

- Caso seja usado validação por série de teste um conjunto de amostras de validação deve ser definido. Em seguida o melhor subconjunto é determinado pelo menor valor da raiz do erro quadrático médio em um conjunto de validação calculado por meio da **Equação 2**, para todos os subconjuntos de variáveis;
- Se for usada validação cruzada, o melhor subconjunto é determinado pelo menor valor da raiz do erro quadrático médio de validação cruzada no conjunto de calibração, esse é obtido por uma equação semelhante à **Equação 2**.

Os resultados armazenados na matriz RMSE possuem dimensões ($M \times K$). O $RMSE(M,K)$ está associado a cadeia com a variável de partida x_k e um total de m variáveis. A melhor cadeia de variáveis é definida pelo menor valor de RMSE. Outras métricas podem ser usadas para a escolha das variáveis de maneira similar.

As limitações abaixo apresentadas devem ser consideradas na hora da definição do número máximo e mínimo de variáveis a serem selecionadas.

$m_{min} \geq A$, O número de variáveis menor que o número de analitos não é recomendado em calibração multivariada;

$m_{max} \leq M$, O número máximo de variáveis a ser selecionado deve ser menor que o número de amostras de calibração.

O algoritmo das projeções sucessivas com correlação ponderada (*Correlation Weighted Successive Projections Algorithm* - CWSPA)^[27] é uma

versão modificada do SPA, promovido por meio de uma mudança na equação de projeção (**Equação 16**) calculada no **passo 3** do SPA. A mudança leva em conta o fator R_j^L , esse fator é o coeficiente de correlação da j -ésima variável e a resposta. Sua extensão é controlada pelo fator L .

$$\mathbf{x}_j^{i+1} = [\mathbf{P}^i \mathbf{x}_j^i] \times (R_j^L) \quad (17)$$

A extensão da contribuição de R na seleção de variáveis, L , foi otimizado a partir de cálculos com L variando de sete maneiras possíveis ($L = 0, 1, 2, \dots, 6$), com base no menor percentual de erro padrão relativo, r_g^4 - CWSPA foi o estado escolhido ($L = 4$)^[27]. É interessante perceber que quando $L = 0$, a **Equação 17** passa a ser igual à **Equação 16**.

Uma simples modificação na formulação SPA, que visa melhorar a parcimônia do modelo MLR resultante foi proposto por Galvão et al.^[30]. A modificação é um processo de eliminação incorporado ao algoritmo a fim de eliminar variáveis que não contribuem efetivamente para a capacidade de previsão do modelo. Tal modificação consiste na adição de uma nova etapa após a escolha das variáveis para a construção dos modelos. Para isso foi usado o índice de relevância r_j calculado com base na **Equação 18**.

$$r_j = S_{v_j} |b_j|, \quad j = 1, 2, \dots, m \quad (18)$$

Esse índice de relevância é calculado com base no desvio padrão S_{v_j} para a variável v_j obtida no conjunto de calibração. Adicionalmente, b_j é o valor absoluto do coeficiente de regressão da j -ésima variável. É importante ressaltar que esse índice de relevância não pode ser calculado no início da seleção, por que a determinação de b_j , por MLR seria mal condicionada. Dessa forma após o cálculo de r_j as variáveis selecionadas pelo SPA, são colocadas em ordem decrescente de acordo com o índice de relevância. E a seqüência de RMSEV é definida da seguinte maneira

De $j = 1$ até m faça

Construa modelos com $\{v_1 \dots v_j\}$

Aplique esse modelo no conjunto de validação

Calcule o RMSEV(j)

Próximo j

O $RMSEV_{\min}$ é o menor valor da seqüência RMSEV assim obtida. Finalmente, o número mínimo de variáveis j para que o valor $RMSEV(j)$ não seja significativamente maior que $RMSEV_{\min}$ é adotado. Para isso, um teste F é empregado na comparação dos valores de RMSEV quadrático. Este critério é similar ao método de Haaland e Thomas^[71], que é usado para determinar um número adequado de variáveis latentes em PLS. Um nível de significância de 25 % ($\alpha = 0,25$) para o teste F foi o adotado.

Uma importante contribuição referente a ganhos computacionais foi proposto por Soares et al.^[50], no qual o algoritmo das projeções sucessivas foi implementado de forma seqüencial.

A regressão seqüencial parte de um modelo entre duas variáveis $y = \beta_1^{(1)}x_1 + \varepsilon^{y|x_1}$, onde $\beta_1^{(1)}$ é o coeficiente de regressão e $\varepsilon^{y|x_1}$ é o resíduo. O termo sobrescrito (1) e $y|x_1$ indica que a regressão é realizada entre uma variável independente e y . A estimativa de $\beta_1^{(1)}$ pelo método dos mínimos quadrados é dado a partir da **Equação 19**.

$$\hat{\beta}_1^{(1)} = \frac{\sum_{i=1}^N y_i x_{i,1}}{\sum_{i=1}^N (x_{i,1})^2} \quad (19)$$

onde y_i e $x_{i,1}$ representam os valores de y e x_1 para a i -ésima amostra de calibração.

Se a regressão é feita a partir de duas variáveis em x , o modelo pode ser escrito por $y = \beta_1^{(2)}x_1 + \beta_2^{(2)}x_2 + \varepsilon^{y|x_1x_2}$. Dessa maneira para obter $\hat{\beta}_1^{(2)}$ e $\hat{\beta}_2^{(2)}$ é preciso realizar uma regressão entre x_2 e x_1 . Essa regressão é representada pela **Equação 20**.

$$x_2 = \delta_1^{x_2|x_1}x_1 + \varepsilon^{x_2|x_1} \quad (20)$$

A estimativa do coeficiente $\hat{\delta}_1^{x_2|x_1}$ pode ser obtido da seguinte maneira

$$\hat{\delta}_1^{x_2|x_1} = \frac{\sum_{i=1}^N x_{i,2} x_{i,1}}{\sum_{i=1}^N (x_{i,1})^2} \quad (21)$$

Dessa forma $\hat{\beta}_1^{(2)}$ e $\hat{\beta}_2^{(2)}$ pode ser obtido como

$$\hat{\beta}_2^{(2)} = \frac{\sum_{i=1}^N e^{y|x_1} x_{i,2}}{\sum_{i=1}^N e^{x_2|x_1} x_{i,2}}, \hat{\beta}_1^{(2)} = \hat{\beta}_1^{(1)} - \hat{\delta}_1^{x_2|x_1} \hat{\beta}_2^{(2)} \quad (22)$$

onde $e_i^{y|x_1} = y_i - \hat{\beta}_1^{(1)} x_{i,1}$ e $e_i^{x_2|x_1} = x_{i,2} - \hat{\delta}_1^{x_2|x_1} x_{i,1}$. Este procedimento pode ser generalizado para obter modelos com m variáveis.

Dentro desse mesmo contexto Soares et al.^[51] trata da implementação de computação paralela usando o Matlab Parallel Computing Toolbox, que requer somente pequenas modificações em códigos quimiométricos já existentes de modo a explorar os benefícios do processamento em múltiplos núcleos.

2.5.7.1 O SPA para seleção de amostras

Para realizar a seleção de amostras, o SPA é aplicado à matriz $\mathbf{X}_{\text{cal}}^t$ em vez de $\mathbf{X}_{\text{cal}}$ ^[18]. Desta forma, SPA seleciona um subconjunto de amostras que são minimamente redundantes (pouco multicolinearidades entre as colunas selecionadas de $\mathbf{X}_{\text{cal}}^t$) e ainda fornecem informações representativas para calibração do modelo MLR. Porém para evitar problemas de mau condicionamento da regressão, quando o número de variáveis é muito maior do que o número de objetos, é recomendada a realização de seleção de variáveis antes de aplicar SPA para seleção de amostras. Tal procedimento reduz a carga de trabalho experimental e computacional envolvido na calibração multivariada, bem como em modelos de transferência de calibração entre diferentes instrumentos.

2.5.7.2 O SPA para classificação

Uma função custo associado ao risco médio do erro de classificação por análise de discriminante linear (LDA) é usada para orientar a seleção SPA^[22]. Na formulação original o SPA usa a minimização do RMSECV ou RMSEV, para a escolha da variável de partida e do número de variáveis, neste caso esses parâmetros são determinados pela minimização da seguinte função custo:

$$G = \sum_{k=1}^N \frac{r^2(\mathbf{X}_k, \mu_{ik})}{\min_{j \neq i} r^2(\mathbf{X}_k, \mu_{ij})} \quad (23)$$

Na **Equação 23**, o numerador $r^2(\mathbf{X}_k, \mu_{ik})$ é o quadrado da distância de Mahalanobis entre o objeto $\mathbf{X}_k$ e a média de sua classe μ_{ik} e o denominador corresponde ao quadrado da distância de Mahalanobis entre o objeto $\mathbf{X}_k$ e o centro da classe errada mais próxima.

2.5.7.3 O SPA para previsão na presença de interferentes

Um interferente é uma espécie que produz um erro sistemático em uma análise pelo aumento ou atenuação do sinal analítico ou sinal de fundo^[72]. A presença de interferentes (em geral classificadas como físicas, químicas ou espectrais) são comuns e podem causar problemas, invalidando uma análise química^[73].

A regressão por MLR tende a ser menos robusta a uma eventual presença de interferentes que os métodos PCR e PLS usando os espectros completos. De fato, se as variáveis forem selecionadas na mesma região que a ocorrência espectral do interferente a previsão pode ser muito comprometida. Em um modelo usando espectros completos o efeito do interferente é dividido por um número maior de canais obtendo-se assim uma maior robustez. Por outro lado se o interferente não sobrepõe o analito na faixa de trabalho, torna-se possível selecionar informações que não são significativamente afetadas pelo interferente.

Uma alternativa para minimizar os problemas causados pelos interferentes usando MLR seria guiar os modelos a escolher regiões espectrais onde a influência do interferente é diminuída. Uma importante vantagem na utilização de dados multivariados é que os sinais não seletivos podem se tornar seletivos pelo uso da matemática^[55]. Em relação a isso os métodos de seleção de variáveis em vez de eliminar a interferência, modelada como um espectro, selecionam as variáveis específicas relacionadas à resposta^[56]. Assim, interferentes podem ser tratados, desde que o sinal (forma) dos interferentes não sejam completamente idêntico ao sinal a se analisar^[55].

O SPA na formulação original compara subconjuntos de variáveis com base na raiz do erro quadrático médio obtido em um conjunto de validação. Tal métrica não avalia uma possível diferença sistemática existente entre os espectros de previsão e calibração. Desse modo com o objetivo de diminuir a influência dos interferentes em modelos SPA-MLR, é proposto um novo critério para a comparação dos subconjuntos de variáveis gerados pelo SPA. Tal critério usa uma nova função custo que leva em conta também o erro estatístico de previsão junto com o RMSE, para a escolha da variável de partida $\mathbf{x}_k$ e o número de variáveis m .

O erro estatístico de previsão^[9] de uma amostra (usando $k = 1, \dots, m$ variáveis) $\mathbf{x}_{i,k}$ é definido a partir da multiplicação da distância de Mahalanobis^[74] pelo resíduo deixado pelo modelo de calibração. Essa relação é apresentada na **Equação 24**.

$$SPE_{i,k} = s \sqrt{\mathbf{x}_{i,k}^T (\mathbf{X}_{cal,k}^T \mathbf{X}_{cal,k})^{-1} \mathbf{x}_{i,k}} \quad (24)$$

onde s é dado por $s^2 = \frac{1}{N_{cal} - m - 1} \sum e_{cal,k}^2$, sendo e_{cal} o resíduo de calibração e N_{cal} o número de amostras de calibração. As matrizes $\mathbf{X}_{cal}$ e $\mathbf{x}_{i,k}$ são centradas na média do conjunto de amostras de calibração.

Com uso da **Equação 24** um valor de SPE_i para cada amostra é obtido. Como é desejado um valor representativo para o conjunto de amostras o somatório de SPE_i é dividido pelo número de amostras deste conjunto conforme apresentado na **Equação 25**.

$$\overline{SPE} = \frac{1}{N} \sum_{i=1}^N SPE_{i,k} \quad (25)$$

A soma do $\overline{SPE}$ com o RMSE fornece um único valor que poderia ser utilizado para a escolha do melhor subconjunto de variáveis, contudo existem pontos incoerentes nessa consideração.

- O primeiro ponto é o fato de estarmos somando um valor que possui unidade do parâmetro a ser determinado com outro que possui unidade do sinal medido;

- Segundo, a simples soma não nos dá idéia de como varia a o $\overline{SPE}$ do conjunto de previsão em relação ao de calibração, se o valor é realmente grande pela natureza das medidas ou pelo fato da amostra não pertencer ao espaço de calibração.

Para contornar os problemas acima citados, os valores de $\overline{SPE}_{pred}$, obtidos para o conjunto de previsão, e $\overline{SPE}_{cal}$, obtidos para o conjunto de calibração são comparados fazendo a razão entre os dois. Dessa maneira tem-se o seguinte:

$$S = \frac{\overline{SPE}_{pred}}{\overline{SPE}_{cal}} \quad (26)$$

se o valor de S é maior que 1, os subconjuntos de variáveis que fornecem esses valores não são considerados como candidatos para o MLR, em outras palavras significa dizer que as variáveis são extremamente afetadas pelo interferente. A nova função custo que será usada no SPA para a escolha do melhor subconjunto de variáveis que apresente a menor interferência possível é escolhido de acordo com a **Equação 27**.

$$F_{custo} = RMSE \times S \quad (27)$$

Observando as **Equações 24 e 25** percebe-se que os valores dos parâmetros de referência não precisam ser conhecidos. Dessa maneira a estatística das amostras externas que serão previstas podem ser usadas para a escolha do melhor subconjunto de variáveis.

De uma forma esquemática as variáveis são escolhidas como mostrado abaixo.

De $k = 1$ até K faça

De $m = m_{min}$ até m_{max} faça

- Use as variáveis com índices $SEL(1,k)$, $SEL(2,k)$, ..., $SEL(m,k)$ para construir um modelo MLR. Aplique o modelo para o conjunto de validação e calcule o $F_{custo}(m,k)$.

Próximo m

Próximo k

Semelhante a formulação original, os resultados armazenados na matriz F_{custo} possuem dimensões $(M \times K)$. $F_{custo}(M, K)$ está associado a cadeia com a variável de partida x_k e um total de m variáveis. A melhor cadeia de variáveis é definida pelo menor valor de F_{custo} .

Um esquema geral de como ocorre à construção e validação dos modelos SPA-SPE-MLR é apresentado na **Figura 2.5**.

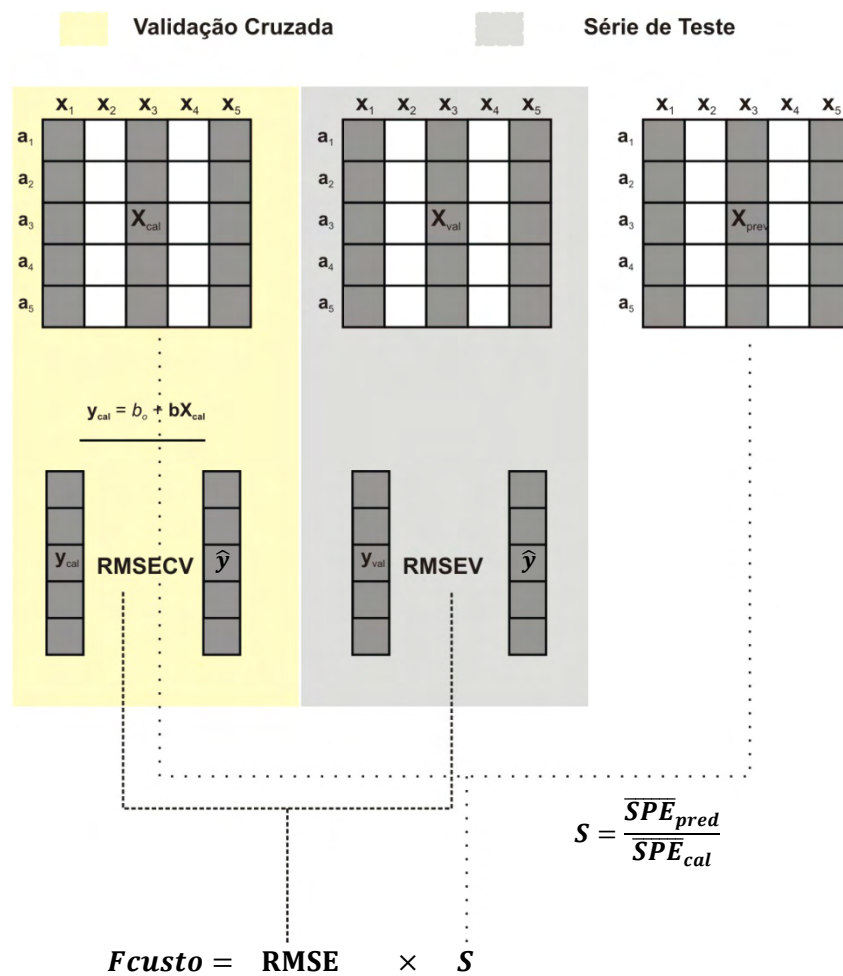


Figura 2.5 – Esquema geral de obtenção de um modelo usando o SPA-SPE-MLR. As variáveis usadas nesse exemplo foram x_1 , x_3 e x_5 .

Após escolhido o melhor subconjunto de variáveis, o método de eliminação de variáveis também foi incorporado ao fim do algoritmo com objetivo de melhorar a parcimônia.

Pode-se perceber que quando S é calculado o fator s da **Equação 24** é anulado, com isso S passa a ser definido prioritariamente pela distância de Mahalanobis^[74] que está intrinsecamente relacionada com o conceito de *leverage* para a detecção de *outlier* e para discriminação entre grupos em trabalhos de classificação envolvendo LDA. Uma interpretação gráfica de quando S é menor que 1 é apresentado na **Figura 2.6**.

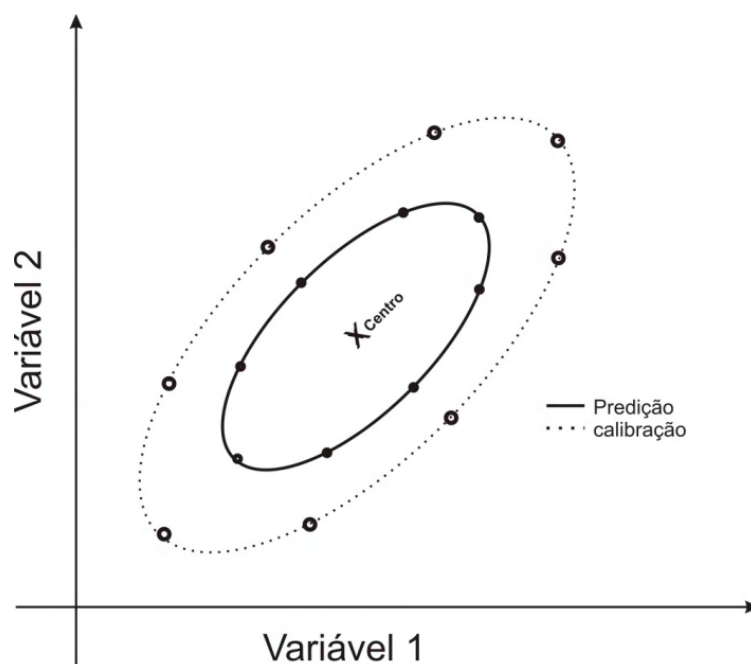


Figura 2.6 – Representação de quando S é menor que 1.

Com esse novo critério espera-se que SPA apresente um melhor desempenho na presença de interferentes.

CAPÍTULO 3

Metodologia

3. Metodologia

3.1. Algoritmos usados

Os modelos MLR foram obtidos por meio das variáveis selecionadas pelo SPA e o SPA-SPE. Esses algoritmos foram desenvolvidos usando o Matlab 6.5[®] e seus códigos encontram-se em um CD anexo à dissertação. Para facilitar o uso destes algoritmos, uma interface gráfica foi desenvolvida, e tanto o SPA como o SPA-SPE foi implementado.

A interface compreende duas etapas, na primeira constrói-se o modelo a partir das matrizes de calibração e validação (quando usada a validação por série de teste). Na etapa de construção são apresentadas as opções ao usuário para montagem do modelo (**Figura 3.1**).



Figura 3.1 – Aba referente à etapa de calibração para o SPA usando o novo critério.

Na segunda etapa, o modelo construído é utilizado na previsão. Caso seja definido o uso do SPE como critério de escolha de novas variáveis para o modelo, esse é recalculado considerando a nova função custo apresentada na **seção 2.5.7.3**. Se a validação por SPE não é escolhida, o SPA-SPE é definido de maneira idêntica ao SPA. As opções ao usuário são apresentadas na **Figura 3.2**.



Figura 3.2 - Aba referente à etapa de previsão para o SPA usando o novo critério.

A modelagem PLS foi realizada usando o software Unscrambler 9.7[®] da (CAMO process AS). Um computador portátil Compaq, equipado com um processador Intel[®] Core[™] 2 Duo de 2.0 MHz, com 2 GB de memória, foi utilizado nos cálculos, em todo o trabalho.

3.2. Aplicações

Para avaliar a os ganhos promovidos pela mudança da função custo para a escolha das variáveis na construção de modelos SPA-SPE-MLR, três aplicações foram realizadas:

- Determinação de três analitos A, B e C em dados simulados;
- Determinações dos corantes amarelo crepúsculo, vermelho 40 e tartrazina utilizando a espectrometria de absorção molecular UV-VIS;
- Determinação de Álcool em gasolina utilizando a espectrometria NIR.

Para efeito comparativo os modelos SPA-MLR e PLS foram obtidos a partir de tais experimentos.

3.2.1. Dados Simulados

Na simulação é assumida uma relação linear entre a matriz de respostas instrumentais $\mathbf{X}$ e a matriz de concentrações $\mathbf{Y}$ para três analitos:

$$\mathbf{X} = \mathbf{Y}\mathbf{W} + \mathbf{N} \quad (28)$$

na **Equação 28**, $\mathbf{N}$ é o ruído artificial adicionado.

O conjunto de dados de calibração composto por 27 amostras foi gerado de acordo com um planejamento fatorial com três níveis (1.0, 5.5, 10.0) para os valores de concentração. De modo similar o conjunto de validação foi obtido de um planejamento fatorial com três níveis (2.0, 5.5, 9.0).

O conjunto de amostras de previsão foi gerado com os valores de concentração variando aleatoriamente dentro da faixa de calibração. Seguindo esse procedimento, 40 amostras foram geradas, sendo que 20 amostras contêm um quarto analito que inexistente no conjunto de calibração e validação.

A previsão foi realizada a partir de dois conjuntos:

- O primeiro compreende 20 amostras que não tem o interferente;

- O segundo consiste de 20 amostras com o sinal de um quarto constituinte contribuindo na matriz de respostas **X**. Os modelos obtidos nesse experimento foram validados por série de teste.

3.2.2. Corantes Alimentícios

Espectros UV-VIS de misturas de 3 corantes sintéticos preparados em solução tampão (fosfato de sódio monobásico anidro/hidróxido de sódio de pH 7,0) foram registrados usando um espectrômetro da HP modelo 8453 a partir de um sistema em fluxo batelada, conforme descrito em Nunes et al.^[75].

O conjunto de calibração foi definido de acordo com um planejamento fatorial completo de 3 níveis e 3 fatores (3^3), totalizando 27 misturas. A validação foi realizada a partir de amostras obtidas de um planejamento fatorial fracionário (3^{3-1}), totalizando 9 misturas. Todos os modelos obtidos a partir desse experimento foram validados por série de teste. Os níveis de concentração considerados para calibração encontram-se na **Tabela 3.1** e os níveis de concentração da validação encontram-se na **Tabela 3.2**.

A Previsão foi realizada em dois conjuntos diferentes:

- No primeiro, para a previsão foram gerados 9 misturas com os níveis de concentração variando aleatoriamente dentro da faixa de calibração;
- Para geração de um conjunto com quatro corantes na previsão (sendo um o interferente eritrosina) foram gerados 30 amostras, variando aleatoriamente dentro da faixa de calibração.

Tabela 3.1 - Níveis de concentração (mg L^{-1}) dos corantes nas misturas de calibração.

Corante	Níveis		
	-1	0	+1
Tartrazina	2,0	6,0	10,0
Vermelho 40	1,5	5,8	10,0
Amarelo Crepúsculo	1,5	5,8	10,0

Tabela 3.2 - Níveis de concentração (mg L^{-1}) dos corantes nas misturas de validação.

Corante	Níveis		
	-1	0	+1
Tartrazina	4,0	6,0	8,0
Vermelho 40	3,6	5,8	7,9
Amarelo Crepúsculo	3,6	5,8	7,9

3.2.3. Determinação de Álcool em gasolina

A partir de uma amostra de gasolina tipo A (que é o tipo de gasolina que não possui etanol em sua composição) prepararam-se 36 amostras com teor de etanol variando entre 10 e 38% (v/v). A partir destas amostras, medidas espectrais NIR foram realizadas utilizando-se uma sonda de transfectância com caminho óptico total de 1,0 mm e uma área circular de $0,5 \text{ cm}^2$. Essa é uma reprodução do experimental de Pereira et al.^[76].

Os espectros de absorção foram obtidos a partir de médias de 100 varreduras na região do infravermelho próximo usando um espectrômetro Brimrose modelo Luminar 2000 que opera na faixa espectral de 850 a 1800 nm.

Das 36 amostras preparadas, 29 foram usadas no conjunto de calibração e as amostras restantes foram previstas pelos modelos. Todos os modelos foram validados por validação cruzada *leave-one-out*.

Um segundo conjunto de 30 amostras, foi obtido a partir de três conjuntos de 10 amostras interferidas por tolueno, hexano e iso-octano, respectivamente, com 5 e 10% (v/v) (cinco amostras com 5% e cinco com 10%). Esses conjuntos só foram usados na etapa de previsão.

Na construção dos modelos multivariados os dados foram pré-processados, por meio da primeira derivada e filtro *Savitzky Golay* com um polinômio de primeira ordem e janela de 7 pontos.

CAPÍTULO 4

Resultados e Discussões

4. Resultados e discussões

Neste trabalho visando superar dificuldades apresentadas pela presença de interferentes nas amostras de previsão. Os modelos MLR e PLS serão obtidos a partir das aplicações descritas no **capítulo 3**, sendo comparados, com intuito de realçar as vantagens da robustez promovida pela mudança da função custo na escolha de variáveis pelo SPA.

Como deve ser entendido, o SPA-SPE-MLR não tem a capacidade de remover o interferente dos espectros. A grande vantagem esperada para essa estratégia é a capacidade de minimizar a ação dos interferentes em amostras externas através da escolha de regiões menos afetadas. Tal vantagem será demonstrada através dos resultados obtidos pelo SPA-MLR e PLS comparados com os obtidos pelo SPA-SPE-MLR.

Para realizar a comparação entre as previsões serão realizados os testes estatísticos, t emparelhado e F , conforme descrito abaixo:

- A diferença entre as previsões realizadas será verificada com base no teste t emparelhado entre os valores previstos pelo SPA-SPE-MLR e o SPA-MLR ou PLS com base na seguinte equação^[77]:

$$t_{cal} = \frac{|bias|\sqrt{N}}{SEP} \quad (29)$$

Caso o valor de t_{cal} seja menor que seu valor crítico (t_{crit}) para $N-1$ graus de liberdade e 95% de confiança, pode-se afirmar que os resultados obtidos pelos dois métodos não apresentam diferença significativa. No cálculo do SEP e do $bias$ apresentado na **Equação 3**, o valor de referência é considerado como sendo o valor previsto pelo SPA-SPE-MLR;

- As precisões dos modelos SPA-SPE-MLR serão comparadas com as obtidas para o SPA-MLR e PLS conforme o teste F descrito abaixo ^[78]:

$$F_{cal} = \frac{(SEP_A)^2}{(SEP_B)^2} \quad (30)$$

O valor de SEP_A deve ser maior que SEP_B e são obtidos de acordo com a **Equação 3**. Um valor crítico (F_{crit}) é obtido com base em 5% de significância e $N-1$ graus de liberdade.

4.1. Aplicações aos dados simulados

Devido à complexidade de espectros de amostras reais, a simulação matemática de dados foi realizada, possibilitando o controle da quantidade de analitos, interferentes, intensidade de ruído associado às medidas, o tipo de ruído e a dimensão da matriz envolvida nos cálculos. O uso de tal banco de dados favorece a interpretação dos resultados devido ao conhecimento do comportamento dos analitos.

Os espectros dos analitos puros e das misturas geradas para as amostras de calibração, validação e previsão são apresentados na **Figura 4.1**.

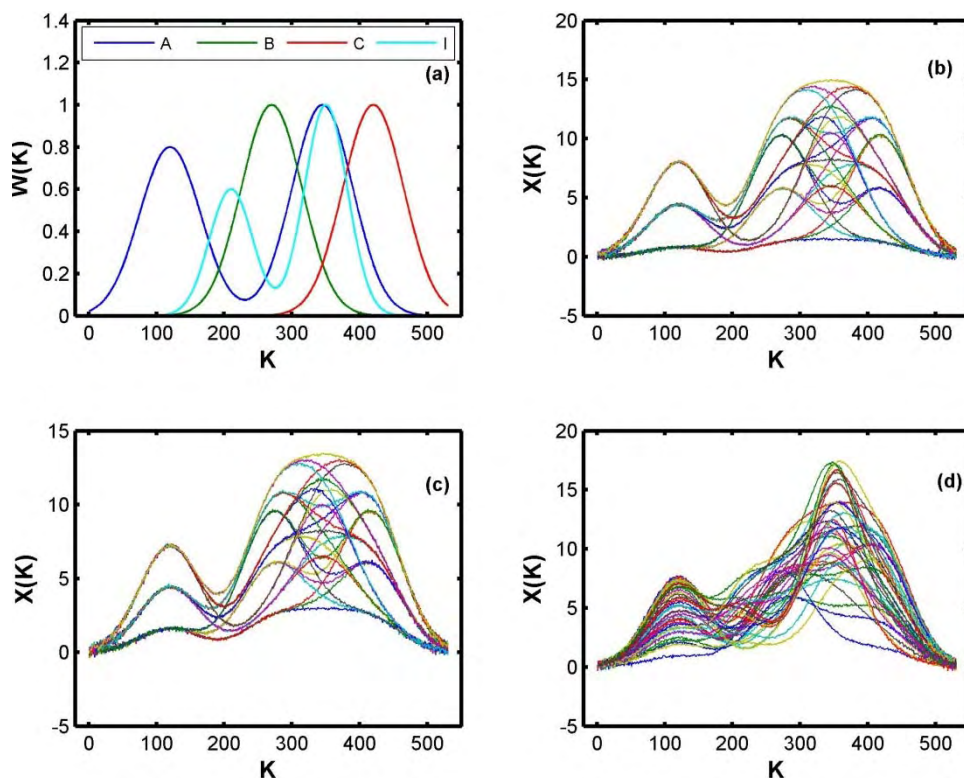


Figura 4.1 – (a) Espectros puros simulados para os analitos A, B, C e o interferente I. Espectros de misturas para: (b) calibração, (c) validação e (d) previsão.

Na **Figura 4.1 (a)** são apresentados os espectros puros dos analitos gerados por meio da superposição de funções gaussianas com médias iguais a 120, 270, 345 e 420 e desvios iguais a 44,72. O espectro do interferente é resultado da soma de duas gaussianas com médias iguais a 210 e 350 e desvios iguais a 31,62. A região que foi menos influenciada pelo interferente foi a região entre 0 e 120.

A partir desses sinais os modelos SPA-MLR, SPA-SPE-MLR e PLS foram obtidos, conforme descrito na **seção 3.2.1**.

4.1.1. Previsão sem interferente

Com objetivo de avaliar a proposta de utilização do SPA com a nova função custo (**Equação 27**), os algoritmos foram aplicados na determinação dos analitos A, B e C sem a presença de nenhum interferente nas amostras de previsão.

A **Tabela 4.1** apresenta os valores de RMSEP para os modelos SPA-MLR, SPA-SPE-MLR e PLS.

Tabela 4.1 – RMSEPs obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS para o conjunto de previsão sem interferente. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.

Analito	SPA-MLR	SPA-SPE-MLR	PLS
A	0,0658 (3)	0,0658 (3)	0,0101 (2)
B	0,0518 (3)	0,0518 (3)	0,0061 (3)
C	0,0650 (4)	0,0564 (3)	0,0084 (3)

Os modelos MLR obtidos para determinação aplicados ao conjunto de previsão sem interferente proporcionaram RMSEPs comparáveis para todos os analitos, mostrando que o novo critério não prejudica a qualidade dos modelos quando comparado com SPA sem tal função custo. Os modelos PLS resultantes foram superiores aos demais. Nesse caso o uso de muitas variáveis foi capaz de amenizar, por efeito de média, a presença do ruído artificial adicionado, fornecendo um menor valor de RMSEP em todos os casos.

As **Figuras 4.2, 4.3 e 4.4** mostram os valores esperados *versus* os valores previstos pelos modelos nesta seção.

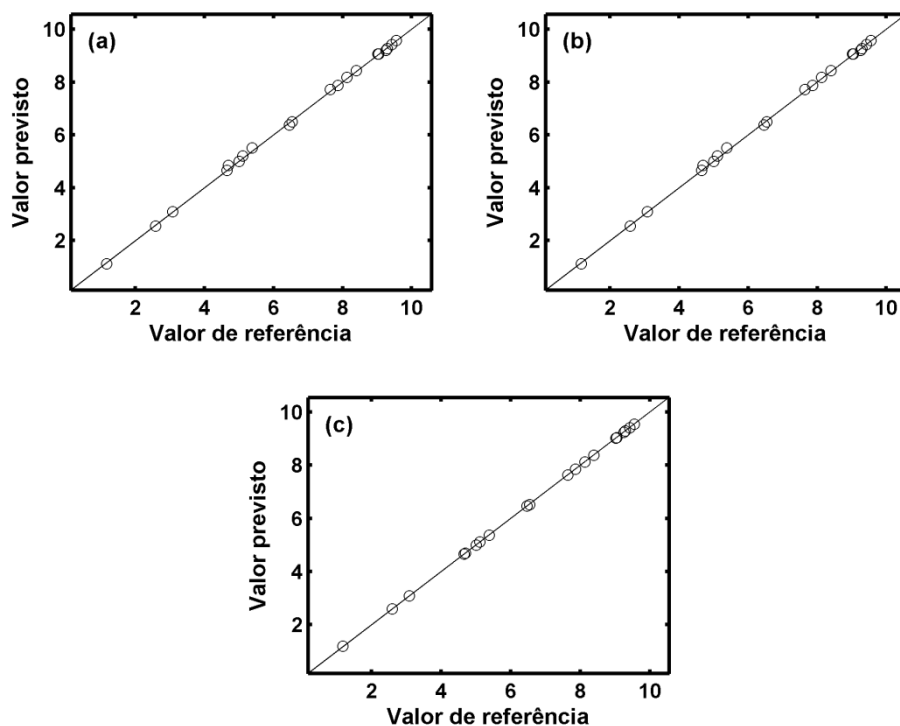


Figura 4.2 – Valores de referência para o analito A *versus* valores previstos sem interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.

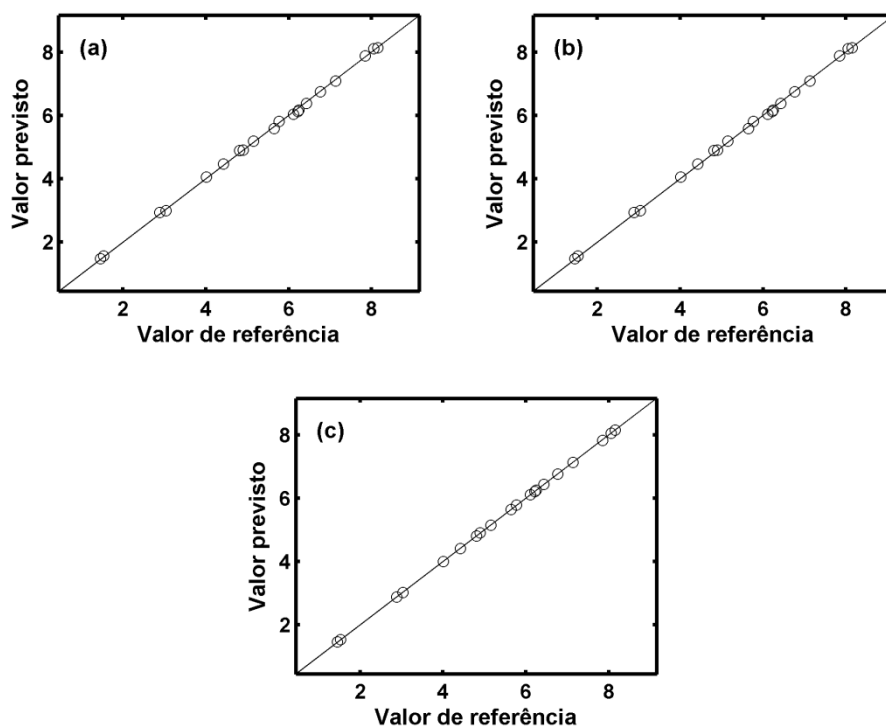


Figura 4.3 – Valores de referência para o analito B *versus* valores previstos sem interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.

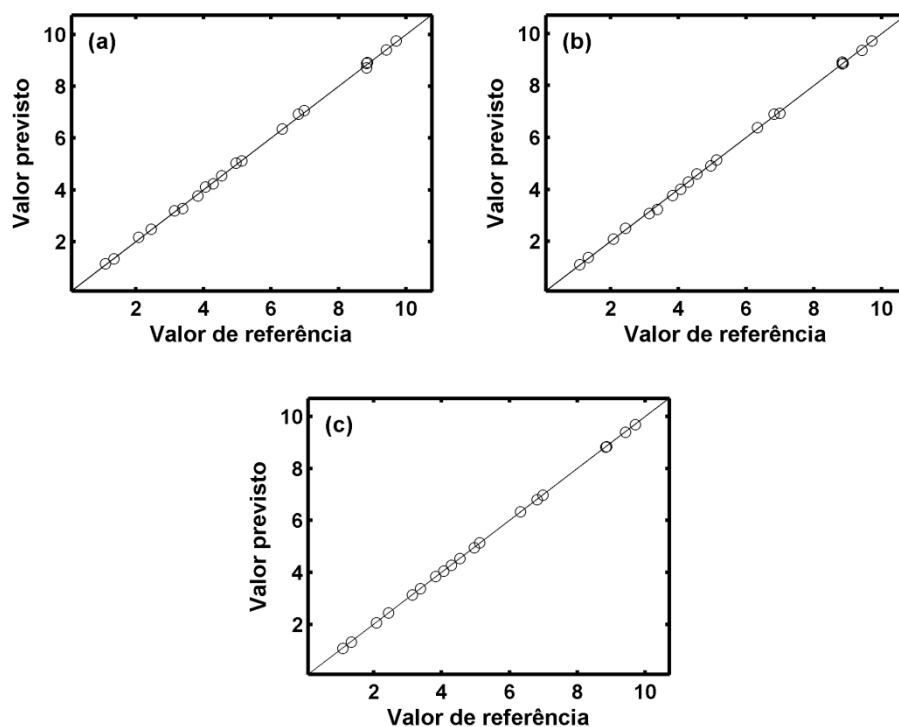


Figura 4.4 – Valores de referência para o analito C *versus* valores previstos sem interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.

As previsões realizadas por todos os modelos (**Figuras 4.2, 4.3 e 4.4**) apresentaram valores que se distribuíram aleatoriamente ao longo da bisetriz, indicando a ausência de erros sistemáticos. Nesse caso o teste *t* emparelhado com 95% de confiança mostrou que não há diferença sistemática entre as previsões realizadas. Quando realizado o teste F, com mesmo nível de confiança, verificou-se que apenas os modelos PLS apresentaram diferenças significativas das previsões realizadas pelo SPA-SPE-MLR.

As variáveis selecionadas em todos os modelos corresponderam a valores muito próximos das médias gaussianas, que são as regiões de maior relação sinal/ruído, indicando que o SPA-SPE é capaz de selecionar variáveis diretamente relacionadas à resposta. Nesta aplicação os modelos SPA-SPE-MLR, quando comparados aos modelos SPA-MLR, se mostraram satisfatórios para determinações dos analitos A, B e C.

4.1.2. Previsão na presença de interferente

Para avaliar a robustez dos modelos na presença de um quarto constituinte presente nas amostras de previsão (interferente I) os algoritmos

SPA-MLR, SPA-SPE-MLR e o PLS foram aplicados para determinação dos analitos A, B e C. Os resultados em termos de RMSEP são apresentados na **Tabela 4.2**.

Tabela 4.2 – RMSEPs obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS para o conjunto de previsão com interferente. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.

Analito	SPA-MLR	SPA-SPE-MLR	PLS
A	5,3687 (3)	0,1002 (1)	2,1797 (2)
B	0,5398 (3)	0,7081 (2)	1,5666 (3)
C	0,8558 (4)	0,5480 (2)	0,6579 (3)

Ao observar-se a **Tabela 4.2** percebe-se que todos os modelos pioraram o seu desempenho quando comparados com os RMSEPs para o conjunto sem interferente (**Tabela 4.1**). O SPA-MLR para a determinação do analito A, foi mais afetado que os outros métodos. Já a seleção de variáveis pelo método SPA-SPE forneceu modelos MLR menos influenciáveis à presença de interferentes em duas das três determinações, enquanto o SPA-MLR forneceu o melhor resultado para determinação do analito B. Esses resultados serão discutidos a partir das variáveis selecionadas.

As **Figuras 4.5, 4.6 e 4.7** contêm os resultados da previsão realizada por meio dos modelos de calibração multivariada SPA-MLR, SPA-SPE-MLR e PLS.

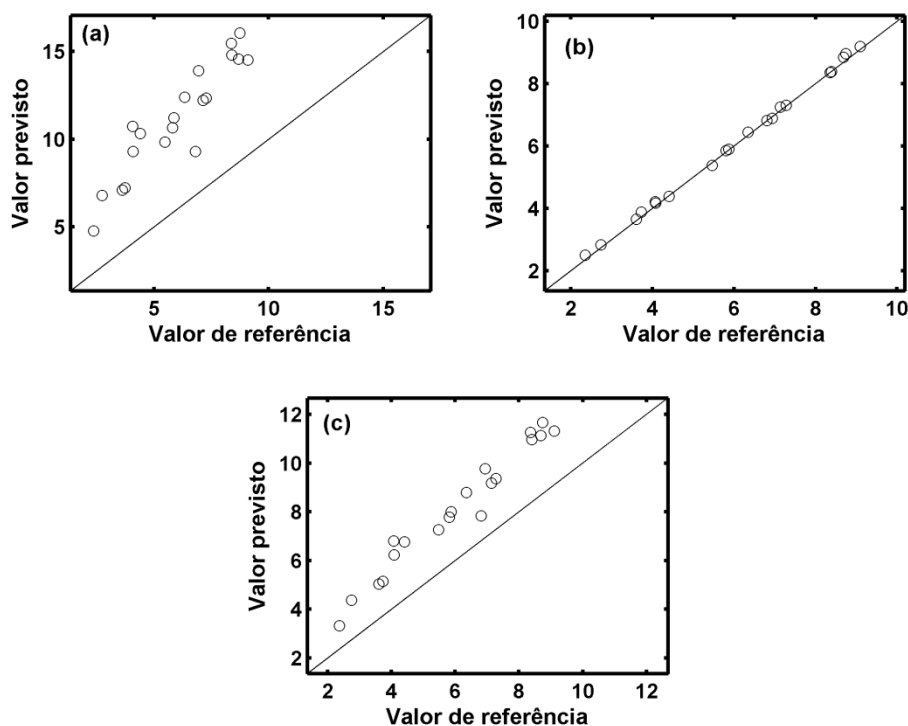


Figura 4.5 - Valores de referência para o analito A *versus* valores previstos na presença do interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.

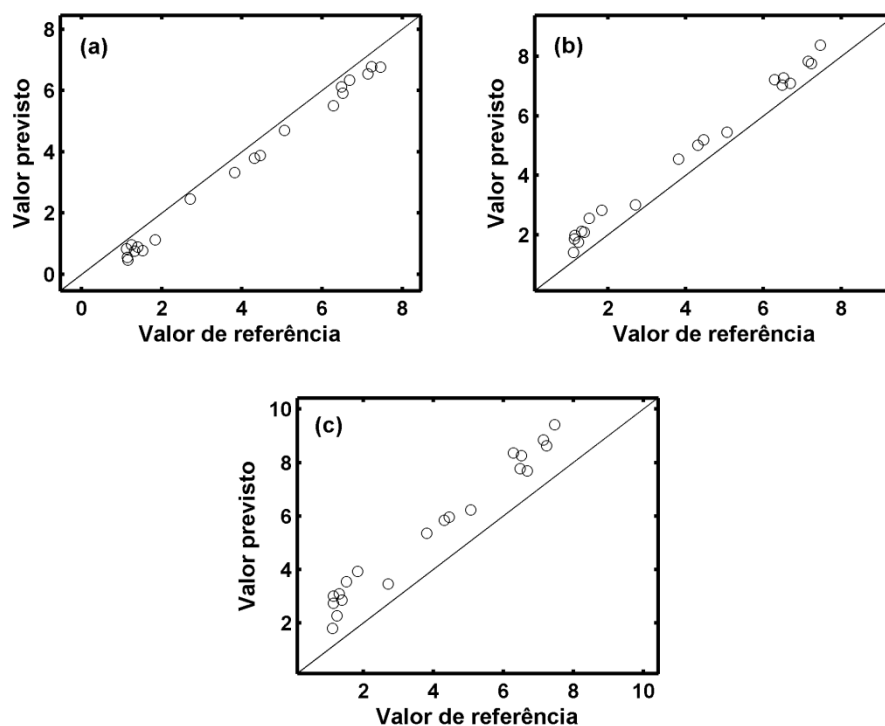


Figura 4.6 - Valores de referência para o analito B *versus* valores previstos na presença do interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.

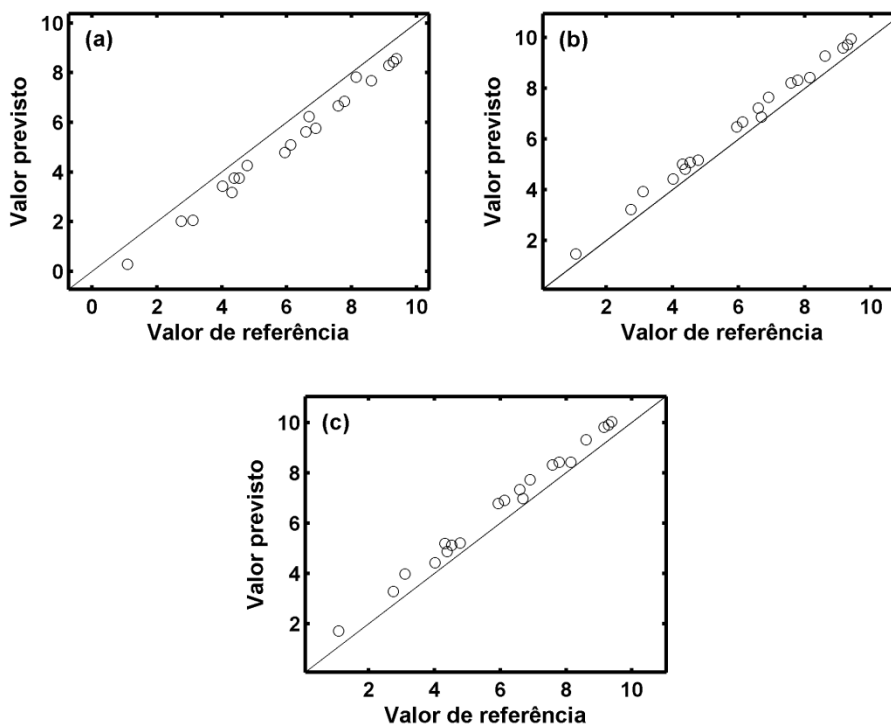


Figura 4.7 - Valores de referência para o analito C *versus* valores previstos na presença do interferente pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR, c) PLS.

A partir das **Figuras 4.5, 4.6 e 4.7** pode ser observado como os interferentes prejudicam os modelos, fazendo que os valores não se distribuam

aleatoriamente ao longo da bissetriz e indicando a existência de erros sistemáticos.

Para avaliar se a previsão realizada pelo modelo SPA-SPE-MLR é significativamente diferente, quando comparado com as obtidas pelo SPA-MLR e o PLS, o teste t emparelhado com 95% de confiança foi realizado. Os valores de t_{cal} e t_{crit} são apresentados na **Tabela 4.3**.

Tabela 4.3 - Valores de t_{cal} e t_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação dos analitos simulados. O número de graus de liberdade encontra-se entre parêntesis.

	SPA-MLR	PLS	t_{crit} (19)
SPA-SPE-MLR	t_{cal}	t_{cal}	
Analito A	15.8141	15.3783	2.0930
Analito B	14.1975	17.3757	
Analito C	16.1354	5.5336	

Observando a **Tabela 4.3** pode-se perceber que os modelos gerados por meio do PLS e SPA-MLR apresentaram valores de t_{cal} maior que t_{crit} quando tiveram seus valores previstos comparados com os previstos pelos modelos SPA-SPE-MLR, ao nível de 95% de confiança. Desse modo, os modelos PLS e SPA-MLR apresentaram exatidões estatisticamente diferentes das previsões realizadas pelo SPA-SPE-MLR.

As precisões dos modelos foram comparadas e os valores de F encontram-se apresentados na **Tabela 4.4**.

Tabela 4.4 - Valores de F_{cal} e F_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação dos analitos simulados. Os números de graus de liberdade encontram-se entre parêntesis.

	SPA-MLR	PLS	F_{crit} (19,19)
SPA-SPE-MLR	F_{cal}	F_{cal}	
Analito A	325.3868	53.7100	2.1683
Analito B	1.8306	3.6756	
Analito C	2.2048	1.3140	

Os valores apresentados na **Tabela 4.4** indicam que não há diferença de precisão entre os modelos SPA-SPE-MLR e SPA-MLR (ao nível de 95% de confiança) para determinação do analito B e entre os modelos SPA-SPE-MLR e PLS para determinação do analito C.

Nas Figuras 4.8, 4.9 e 4.10 são apresentadas as variáveis selecionadas para os modelos MLR que foram usados para as determinações dos analitos A, B e C.

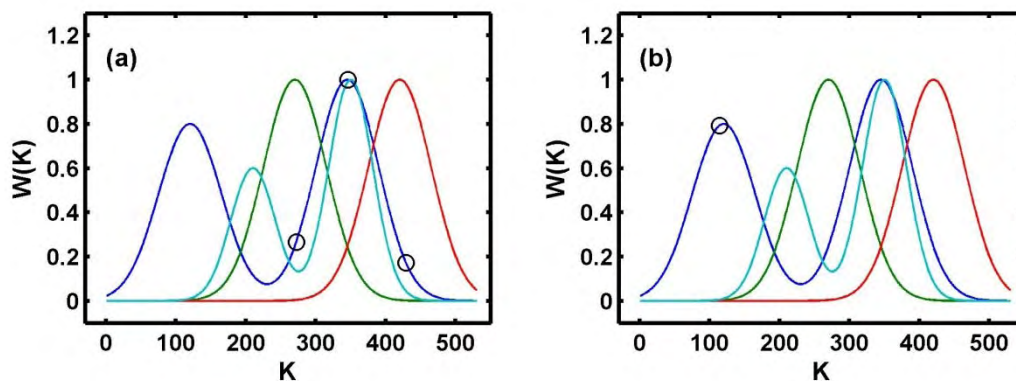


Figura 4.8 - Variáveis usadas para determinação do analito A pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.

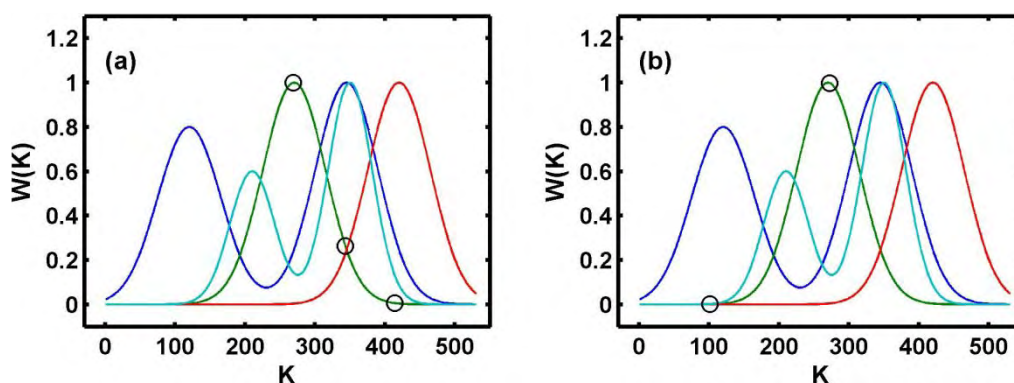


Figura 4.9 - Variáveis usadas para determinação do analito B pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR.

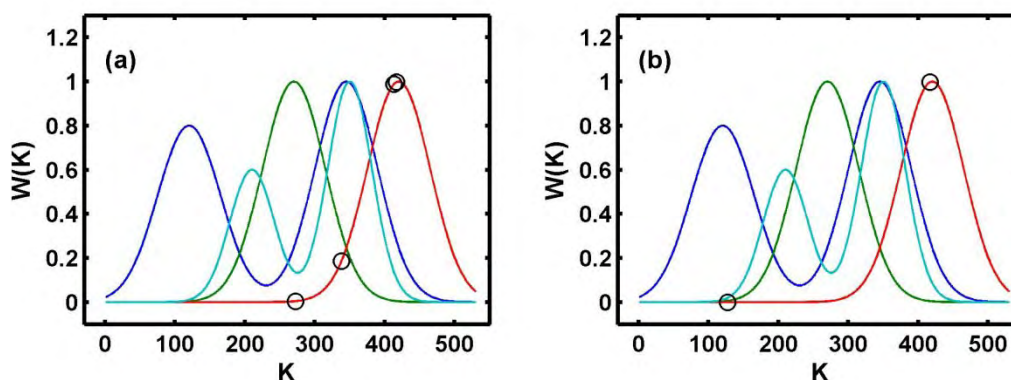


Figura 4.10 - Variáveis usadas para determinação do analito C pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR.

Observando as Figuras 4.8, 4.9 e 4.10, percebe-se que foram usadas pelo SPA-SPE-MLR regiões com menor interferência, enquanto o SPA-MLR em todos os casos selecionou variáveis no qual a intensidade do interferente é mais

pronunciada. O efeito desta seleção foi refletido nos maiores valores de RMSEPs e nos valores se afastando sistematicamente da bissetriz com maior intensidade. As variáveis que foram usadas pelo SPA-SPE-MLR se aproximaram das médias das funções gaussianas de cada analito.

Na determinação do analito A, um indicativo para o melhor resultado pelo SPA-SPE-MLR pode ser atribuído a diminuição da interferência. Os sinais do analito A foram gerados por duas funções gaussianas, com uma mais distante da região do interferente e com menor intensidade (**Figura 4.1 (a)**). Essa liberdade permitiu a mudança da região usada na construção dos modelos SPA-SPE-MLR **Figura 4.8 (a) e 4.8 (b)**, de uma região que tinha maior relação sinal/ruído, para uma região de menor intensidade, entretanto menos afetada pelo interferente. Na determinação dos analitos B e C, percebe-se que as variáveis que foram selecionadas **Figura 4.9 e 4.10** ocorreram em regiões de menor intensidade do interferente.

Dessa forma o SPA-SPE, principalmente para o analito A que foi simulado como tendo uma menor interferência, forneceu menores valores de RMSEPs estatisticamente diferentes dos outros métodos, resultando em um melhor desempenho na presença do interferente.

4.2. Determinações de corantes

Na **Figura 4.11 (a)** são apresentados os espectros dos corantes amarelo crepúsculo, tartrazina, vermelho 40 e o interferente eritrosina, puros. Os espectros de mistura dos corantes para as amostras de calibração, validação e previsão são apresentados na **Figura 4.11 (b), (c) e (d)**.

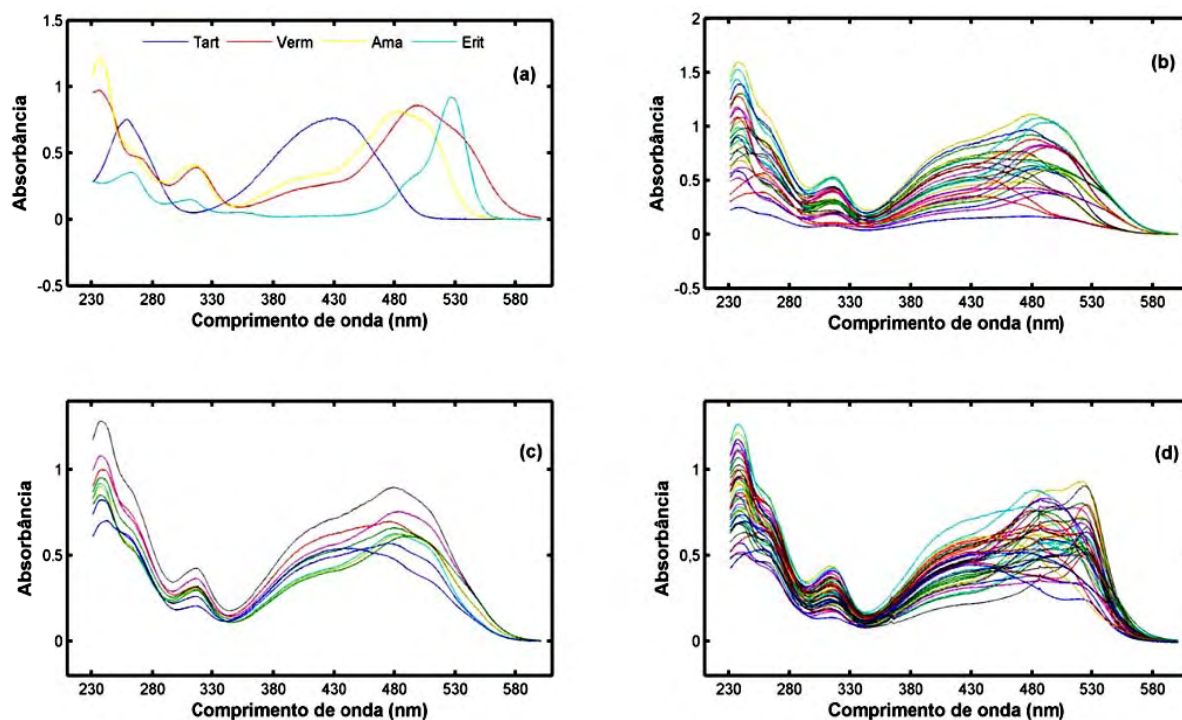


Figura 4.11 - a) Espectros dos corantes tartrazina, vermelho 40, amarelo crepúsculo e o interferente eritrosina puros e os espectros de mistura das amostras de: b) calibração, c) validação e d) previsão.

Na **Figura 4.11 (a)** percebe-se a sobreposição espectral e a grande semelhança entre os espectros dos corantes vermelho 40 e amarelo crepúsculo. Em Nunes et al.^[75] pode ser visto que a diferença em suas estruturas se deve à presença dos grupos (CH_3 e OCH_3) ligados a um anel aromático na molécula do corante vermelho 40. O grupo OCH_3 é o provável responsável pelo deslocamento da banda da direita para maiores valores de comprimentos de onda no espectro desse corante.

Tais problemas realçam a exigência quanto ao desempenho requerido na aplicação dos algoritmos utilizados para a seleção de variáveis em modelos de calibração MLR.

4.2.1. Determinações de corantes sem interferente

A determinação de corantes nesta etapa foi realizada sem a presença do corante eritrosina nos espectros. Os modelos MLR (construídos com as variáveis selecionadas) e os PLS (usando os espectros completos) foram obtidos e os RMSEPs encontram-se apresentados na **Tabela 4.5**.

Tabela 4.5 - RMSEPs obtidos para determinação dos corantes sem a presença de eritrosina (mg L^{-1}) pelos modelos SPA-SPE-MLR, SPA-MLR e PLS. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.

Corante	SPA-MLR	SPA-SPE-MLR	PLS
Amarelo	0,1 (4)	0,1 (3)	0,1 (3)
Tartrazina	0,1 (7)	0,1 (7)	0,1 (3)
Vermelho	0,2 (4)	0,2 (4)	0,1 (3)

Na **Tabela 4.5** percebe-se que em quase todas as determinações, o desempenho dos modelos foram similares apresentando RMSEPs iguais a $0,1 \text{ mg L}^{-1}$, os quais correspondem a precisão das concentrações dos corantes utilizados como método de referência. Apenas na determinação de vermelho 40 os valores de RMSEPs foram superiores. Provavelmente, isso se deve à forte sobreposição espectral decorrente da semelhança do perfil espectral entre o corante vermelho 40 e o amarelo crepúsculo.

Os gráficos dos valores de referência *versus* valores previstos para o SPA-MLR, SPA-SPE-MLR e PLS são apresentados nas **Figuras 4.12, 4.13 e 4.14**.

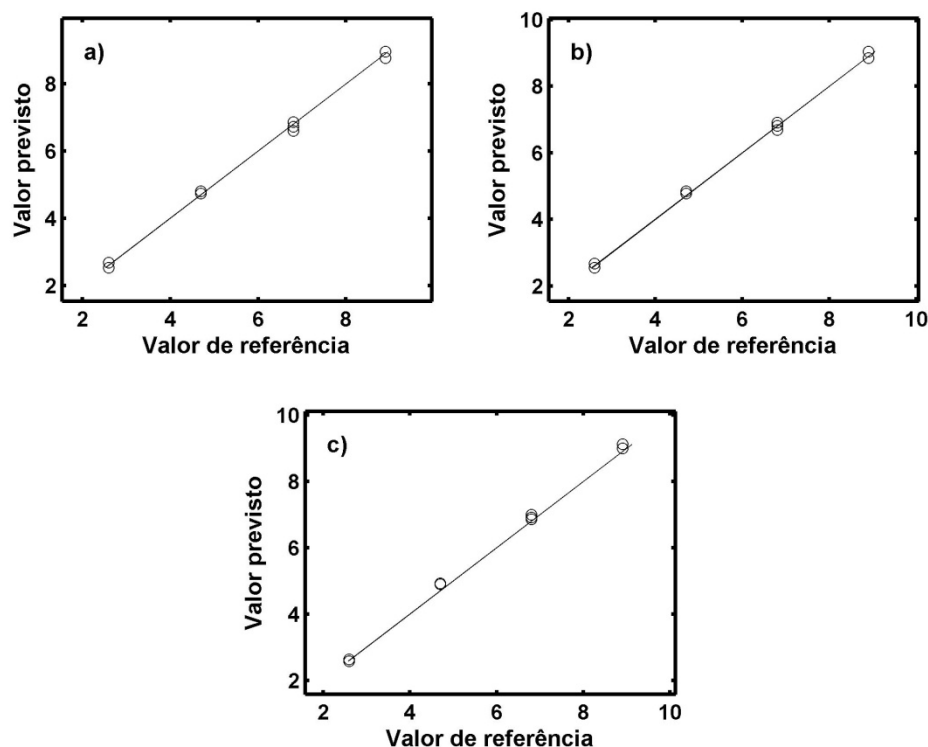


Figura 4.12 - Valores de referência do corante amarelo crepúsculo *versus* valores previstos sem a presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.

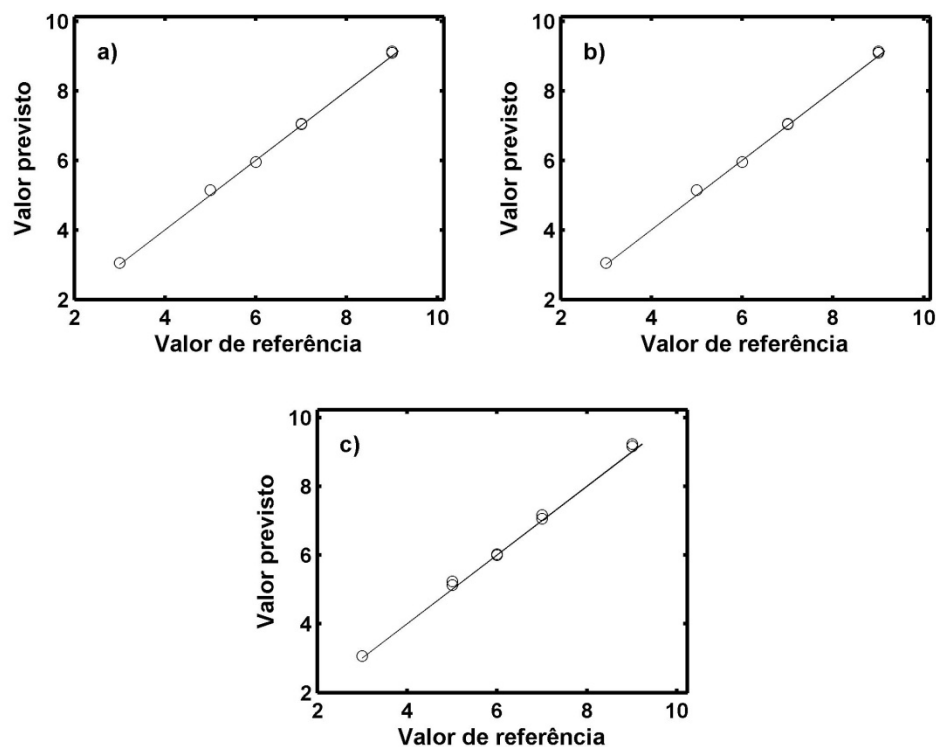


Figura 4.13 - Valores de referência do corante tartrazina *versus* valores previstos sem a presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.

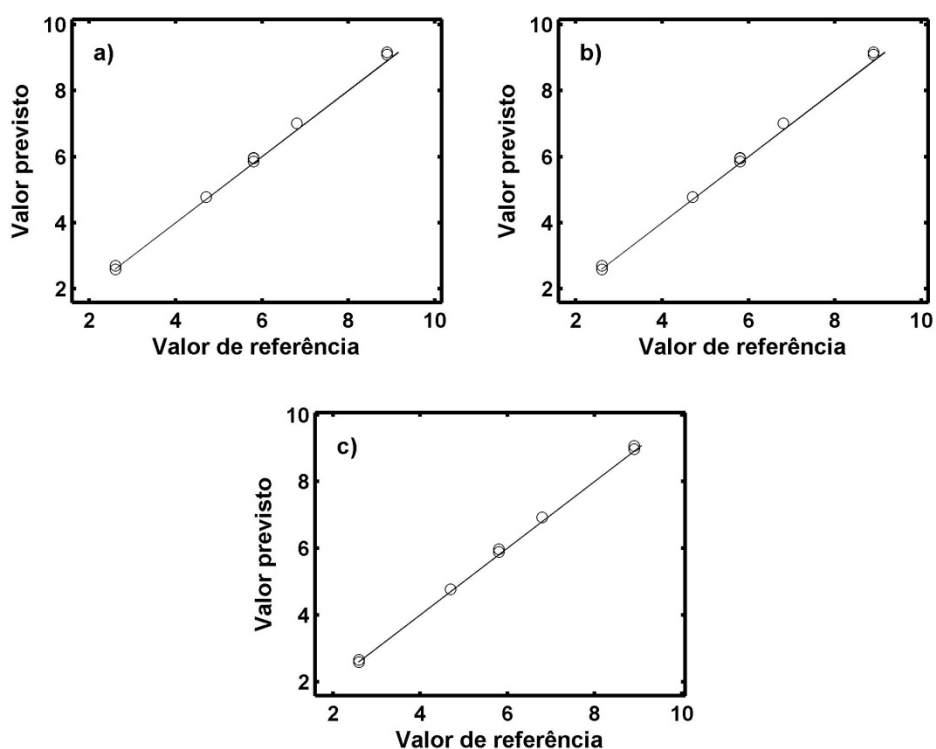


Figura 4.14 - Valores de referência do corante vermelho 40 *versus* valores previstos sem a presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.

Nas **Figuras 4.12, 4.13 e 4.14** as previsões realizadas por todos os modelos reforçam a igualdade observada na **Tabela 4.5**.

Os testes *t* e *F*, ao nível de 95% de confiança, foram realizados para comparação entre os modelos. Todos os modelos SPA-MLR e PLS quando comparados aos modelos SPA-SPE-MLR apresentaram desempenhos semelhantes.

Dessa maneira, sem interferentes, as determinações dos corantes amarelo crepúsculo, tartrazina e vermelho 40 são realizadas de maneira satisfatória pelo modelo SPA-SPE-MLR.

4.2.2. Determinações de corantes com interferente

Nesta seção no conjunto de previsão usado foi adicionado o corante eritrosina. Na **Figura 4.15** são apresentados os espectros puros de eritrosina, variando a concentração entre 1,0 e 10,0 mg L⁻¹.

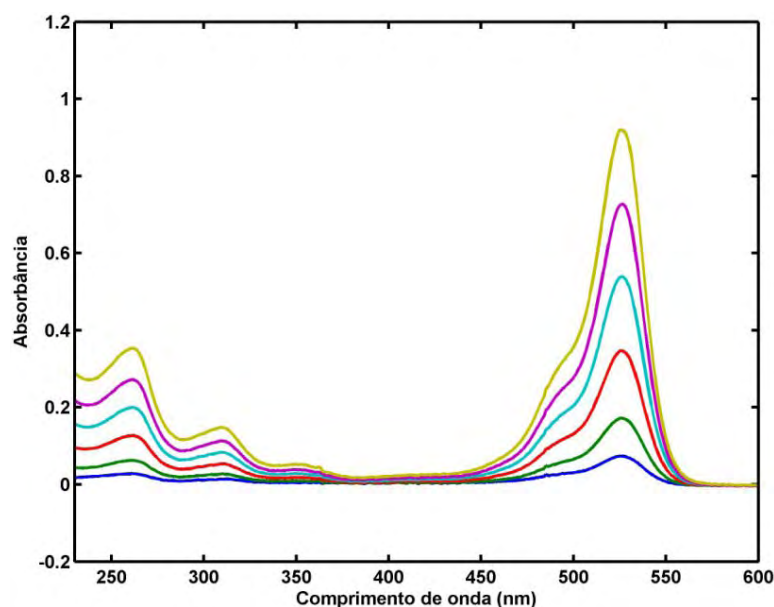


Figura 4.15-Espectros de eritrosina variando a concentração entre 1,0 e 10,0 mg L⁻¹.

No conjunto de previsão, a região entre 450 e 560 nm foi à região mais afetada pela ação do corante eritrosina. Desse modo espera-se que o SPA-SPE-MLR selecione variáveis que não se encontram principalmente nessa região.

A determinação dos corantes quando há a contribuição de eritrosina adicionada aos espectros de previsão (assim como descrito na **seção 3.2.2**) foi

realizada. Os valores de RMSEPs para a determinação dos três corantes são apresentados na **Tabela 4.6**.

Tabela 4.6 - RMSEPs obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS na presença do interferente eritrosina (mg L^{-1}). O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.

Corante	SPA-MLR	SPA-SPE-MLR	PLS
Amarelo	3,3 (4)	1,0 (3)	3,8 (3)
Tartrazina	2,2 (7)	0,6 (2)	1,1 (3)
Vermelho	4,4 (4)	0,9 (2)	4,9 (3)

Os resultados mostram que os modelos MLR e PLS, quando comparados aos resultados da aplicação anterior, diminuíram muito o seu desempenho de previsão, mostrando que a ação dos interferentes prejudica a previsão. De um modo geral na presença de interferentes os modelos SPA-SPE-MLR apresentaram menores RMSEPs que os modelos PLS e SPA-MLR.

As **Figuras 4.16, 4.17 e 4.18** apresentam os valores de referência *versus* valores previstos para os modelos SPA-MLR, SPA-SPE-MLR e PLS.

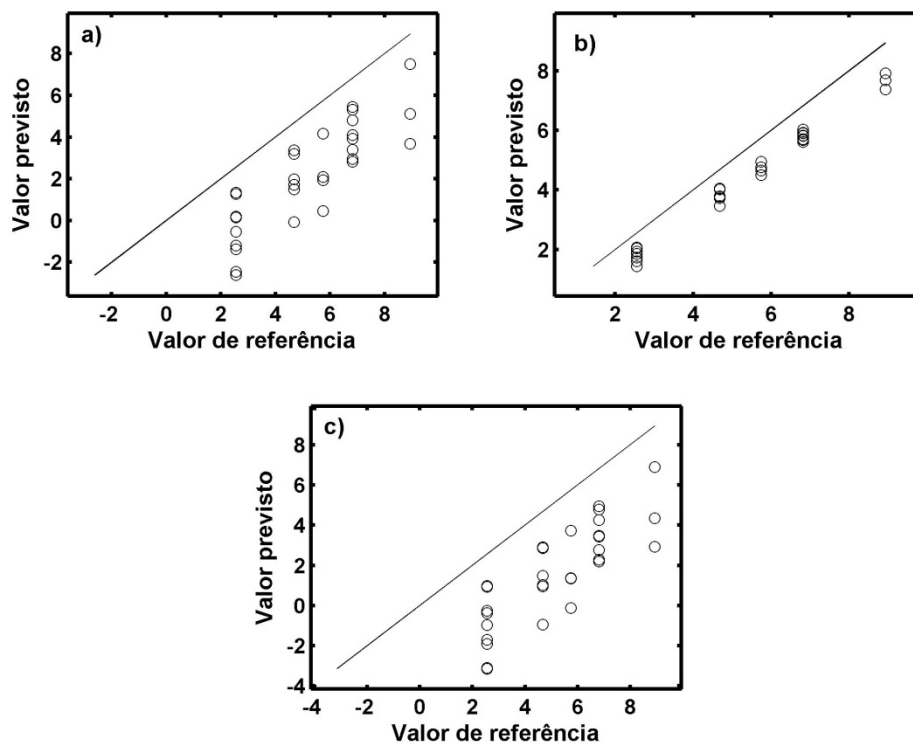


Figura 4.16 - Valores de referência do corante amarelo crepúsculo *versus* valores previstos na presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.

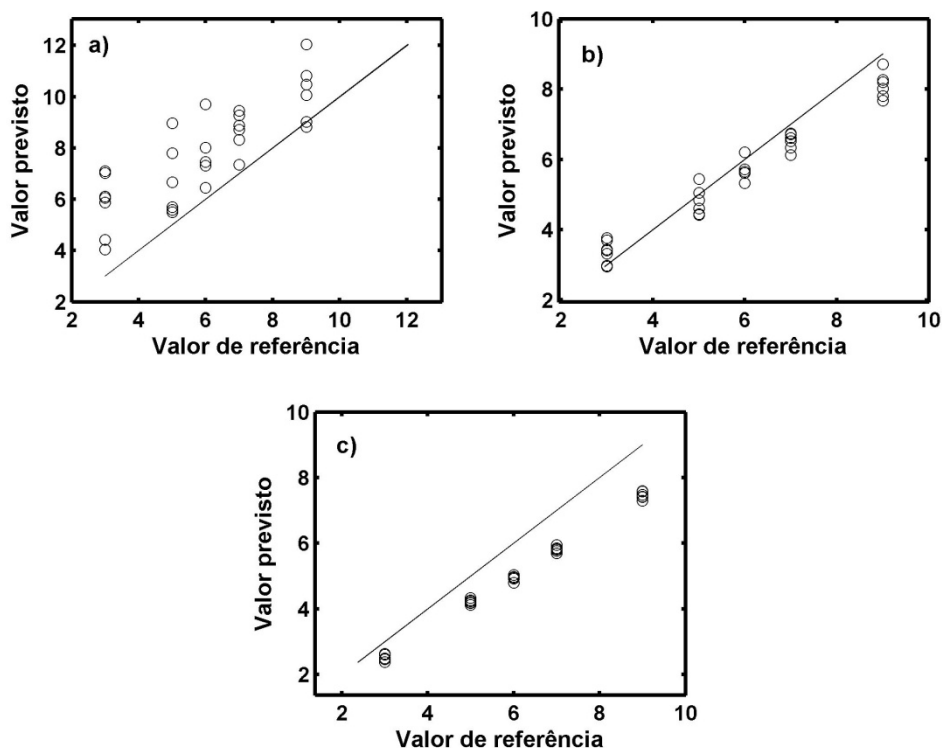


Figura 4.17 - Valores de referência do corante tartrazina *versus* valores previstos na presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.

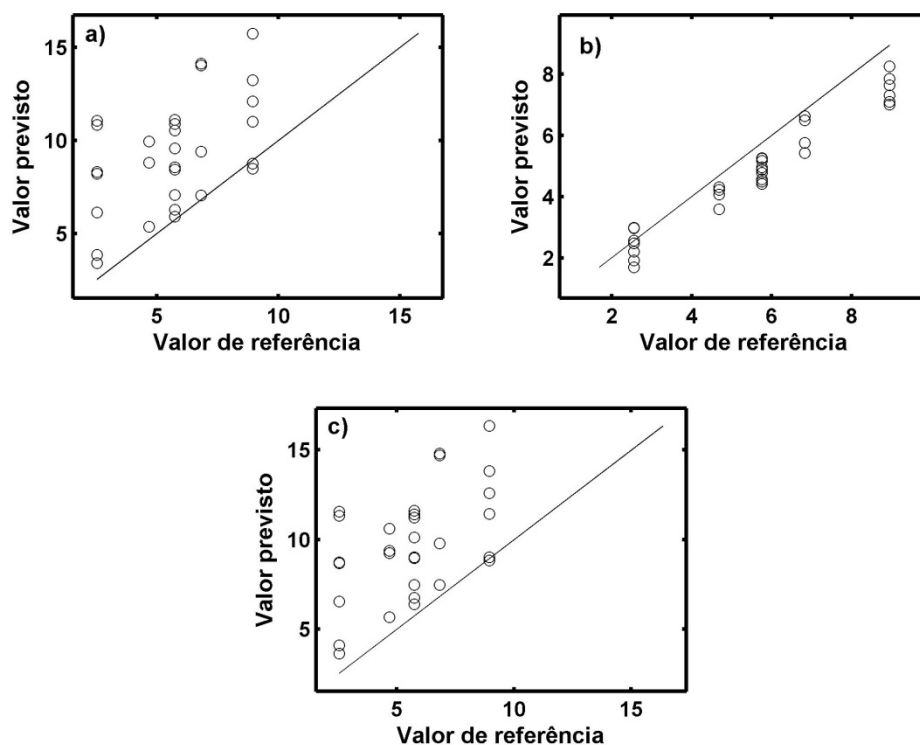


Figura 4.18 - Valores de referência do corante vermelho 40 *versus* valores previstos na presença de eritrosina pelos modelos: a) SPA-MLR, b) SPA-SPE-MLR e c) PLS.

As **Figuras 4.16, 4.17 e 4.18** mostram a superioridade dos modelos obtidos por meio do SPA-SPE-MLR. Na determinação de todos os corantes os modelos SPA-SPE-MLR se aproximaram mais da bissetriz.

Para avaliar se há diferença na previsão realizada pelos modelos, os testes t e F foram realizados com intuito de comparar os resultados dos modelos SPA-MLR e PLS com as previsões realizadas pelo SPA-SPE-MLR. Os valores de t_{cal} e t_{crit} encontram-se na **Tabela 4.7** e os valores de F_{cal} e F_{crit} na **Tabela 4.8**.

Tabela 4.7 - Valores de t_{cal} e t_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação dos corantes. O número de graus de liberdade encontra-se entre parêntesis.

	SPA-MLR	PLS	t_{crit} (29)
SPA-SPE-MLR	t_{cal}	t_{cal}	
Amarelo	10,04	11,9883	2,0452
Tartrazina	14,48	9,8712	
Vermelho 40	11,30	12,0054	

Observa-se na **Tabela 4.7** que os valores de t_{cal} para todos os casos foram maiores t_{crit} . Desse modo pode-se afirmar que os valores previstos pelos modelos SPA-SPE-MLR são estatisticamente diferentes dos previstos pelos modelos SPA-MLR e PLS, ao nível de 95% de confiança.

Tabela 4.8 - Valores de F_{cal} e F_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS) usados na determinação de corantes. Os números de graus de liberdade encontram-se entre parêntesis.

	SPA-MLR	PLS	F_{crit} (29,29)
SPA-SPE-MLR	F_{cal}	F_{cal}	
Amarelo	29,0063	32,8774	1,8608
Tartrazina	5.2617	1.9603	
Vermelho	19,4434	20.7316	

A **Tabela 4.8** mostra que para todos os modelos os valores de F_{cal} foram maiores que F_{crit} , dessa forma pode-se afirmar, com 95% de confiança, que todos os modelos SPA-MLR e PLS apresentaram precisões diferentes do SPA-SPE-MLR.

As **Figuras 4.19, 4.20 e 4.21** apresentam as variáveis selecionadas pelos modelos SPA-MLR e SPA-SPE-MLR, nos espectros dos corantes puros.

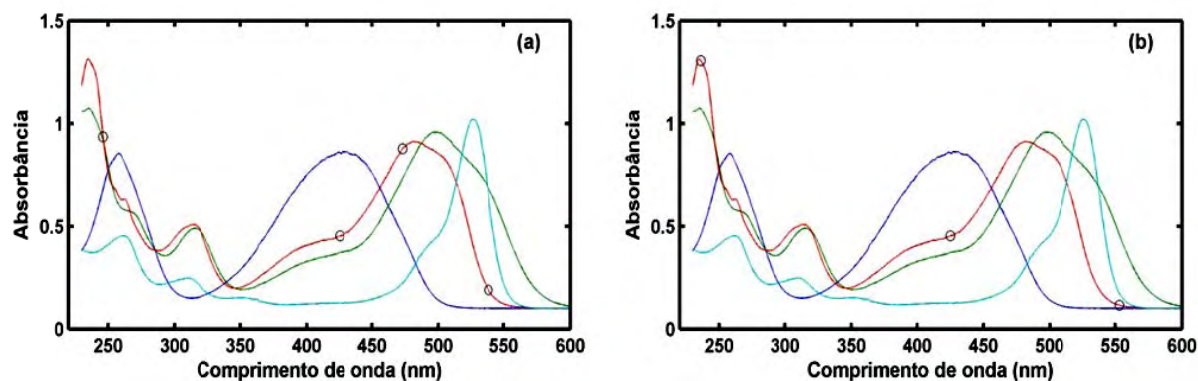


Figura 4.19 - Variáveis selecionadas para determinação do corante amarelo crepúsculo usando o conjunto de previsão com interferente pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.

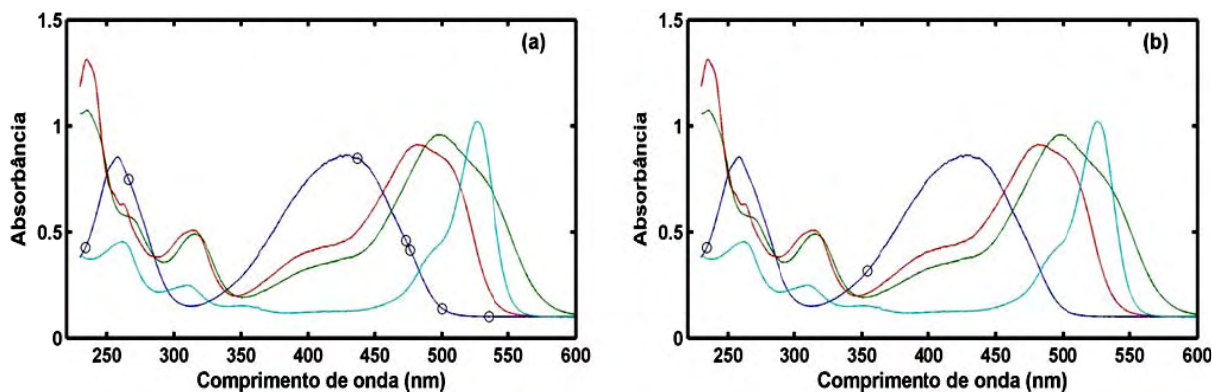


Figura 4.20 - Variáveis selecionadas para determinação do corante tartrazina usando o o conjunto de previsão com interferente pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.

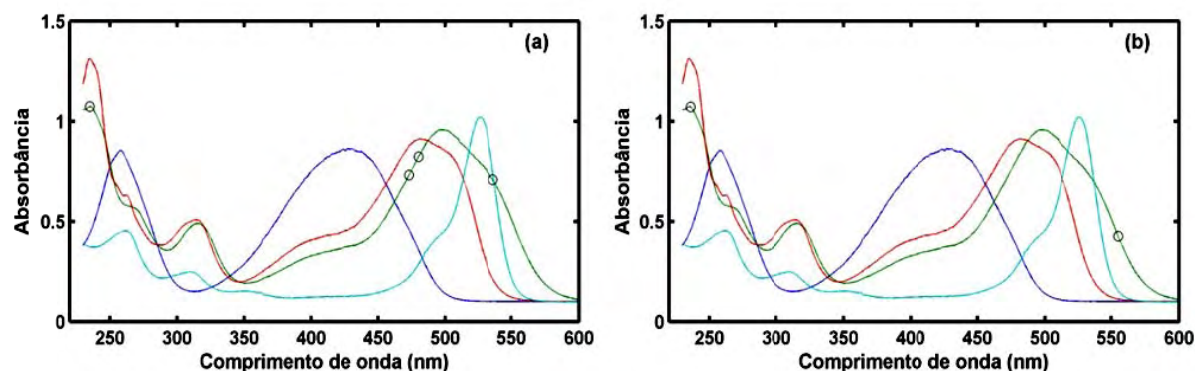


Figura 4.21 - Variáveis selecionadas para determinação do corante vermelho 40 usando o conjunto de previsão com interferente pelos modelos: a) SPA-MLR e b) SPA-SPE-MLR.

As Figuras 4.19, 4.20 e 4.21 mostram que o SPA-SPE foi capaz de evitar a região de maior absorção do interferente, selecionando variáveis em regiões de menor intensidade de absorção. A essa mudança, atribuímos a causa dos melhores resultados para os modelos SPA-SPE-MLR.

A similaridade entre o perfil espectral dos próprios corantes presentes na calibração também dificultam a modelagem. Isso é refletido nos modelos para determinação de amarelo crepúsculo e vermelho 40, esses são os corantes que possuem maiores semelhanças espectrais e apresentaram maiores RMSEPs (**Tabela 4.6**). Estes RMSEPs para determinações dos diferentes corantes podem ser comparados, pois as concentrações destes variam dentro da mesma escala (**Tabela 3.1**).

De um modo geral os resultados desta aplicação mostraram que o SPA-SPE-MLR foi estatisticamente superior, ao nível de 95% de confiança, para determinação de todos os corantes quando eritrosina estava presente nos espectros de previsão.

4.3. Determinação de álcool em gasolina

A **Figura 4.22** apresenta os espectros de 36 amostras de gasolinas com o teor de álcool variando entre 10 e 38% (v/v). Nesses espectros é perceptível as absorções variando principalmente na região entre 1400 e 1800 nm, que corresponde a bandas provenientes de vibrações das ligações O – H do álcool.

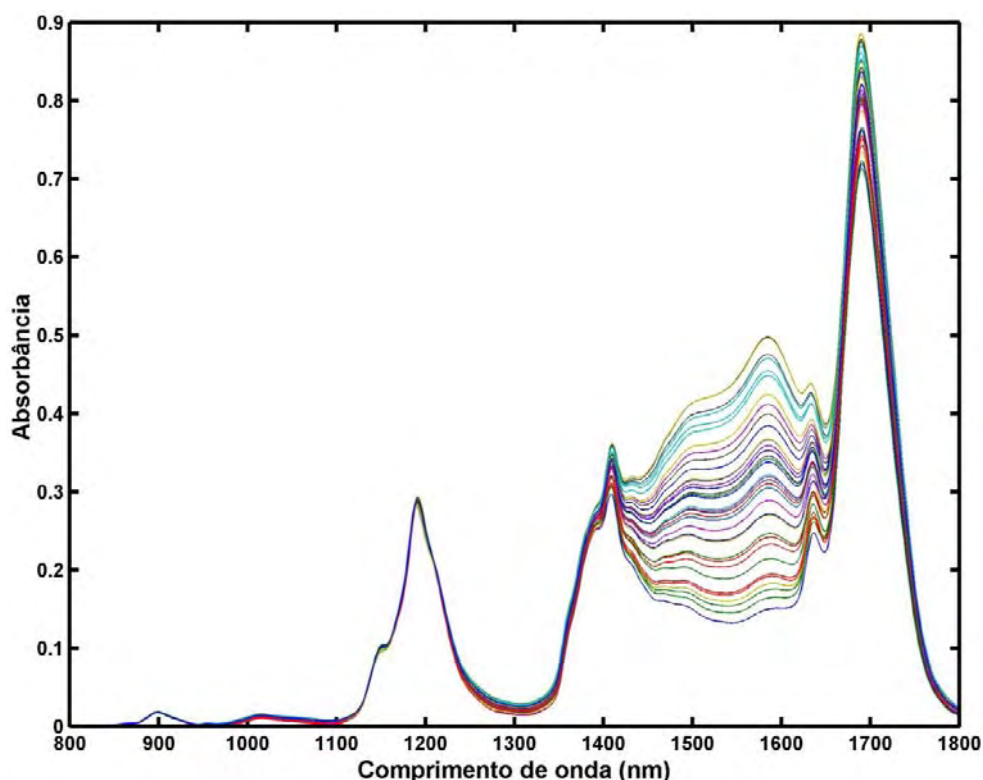


Figura 4.22 – Espectros NIR de 36 amostras de gasolina com a concentração de etanol variando entre 10 e 38% (v/v).

A partir dos espectros gerados, como descritos no **Capítulo 3**, o SPA-MLR, SPA-SPE-MLR e o PLS (aplicado aos espectros completos) foram usados. Nesta aplicação vale à pena lembrar que os modelos foram construídos usando a validação cruzada, diferentemente das outras aplicações. Amostras de previsão sem interferentes e interferidas por tolueno, hexano e Iso-octano foram previstas e os resultados das aplicações em termos de RMSEPs encontram-se apresentados na **Tabela 4.9**.

Tabela 4.9 - RMSEPs do teor de etanol (%v/v) em amostras de gasolinas obtidos pelos modelos SPA-MLR, SPA-SPE-MLR e PLS. O número de variáveis selecionadas e variáveis latentes encontram-se entre parêntesis.

Amostras	SPA-MLR	SPA-SPE-MLR	PLS
Sem interferentes	0,89 (9)	0,54 (1)	0,80 (1)
(10%, 5%) de Tolueno	34,13 (9)	1,07 (1)	4,00 (1)
(10%, 5%) de Hexano	5,16 (9)	0,79 (1)	0,49 (1)
(10%, 5%) de Iso-octano	2,76 (9)	0,66 (1)	0,61 (1)

Como podem ser vistos, os valores de RMSEPs apresentados pelo SPA-SPE-MLR foram menores em todos os casos quando comparado ao SPA-MLR, entretanto quando comparado ao PLS os modelos apresentaram menores RMSEPs em duas das quatro determinações. Os modelos SPA-SPE-MLR selecionaram apenas uma variável, em todas as determinações, demonstrando uma melhor parcimônia.

Uma explicação para a seleção de apenas uma variável pode ser atribuída ao fato das 36 amostras terem sido obtidas de apenas uma amostra de gasolina tipo A (amostra que não contém álcool em sua composição), o que nos revela que apenas o etanol provocou a variação nos espectros quando as amostras de previsão não estão com interferentes.

As **Figuras 4.23, 4.24, 4.25 e 4.26** mostram como os valores de referência *versus* valores previstos se comportaram em torno da bissetriz.

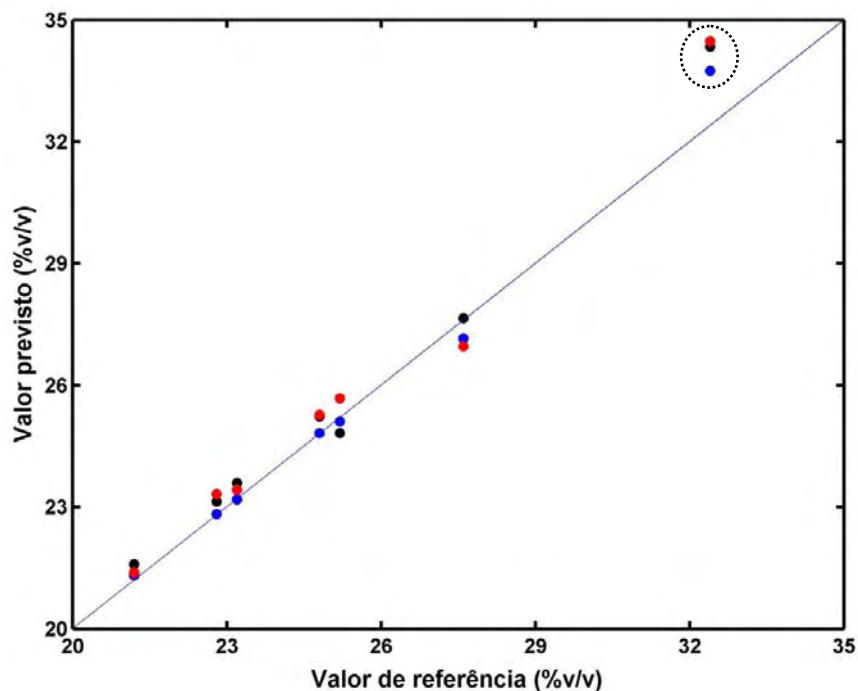


Figura 4.23 - Valores de referência de etanol *versus* valores previstos sem interferentes pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS.

Na **Figura 4.23** pode-se perceber que uma amostra se desvia muito da bissetriz. Não foi investigado se essa amostra é um outlier, pois neste trabalho pretende-se atenuar as variações que ocorrem nas amostras de previsão, sem ser necessário descartá-las.

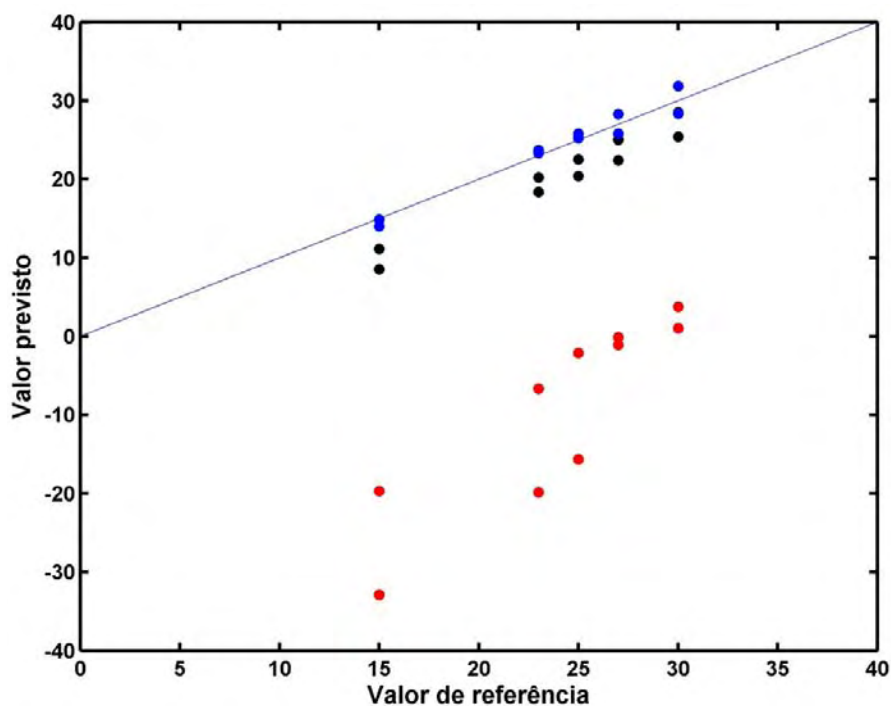


Figura 4.24 - Valores de referência de etanol *versus* valores previstos, na presença do interferente tolueno, pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS.

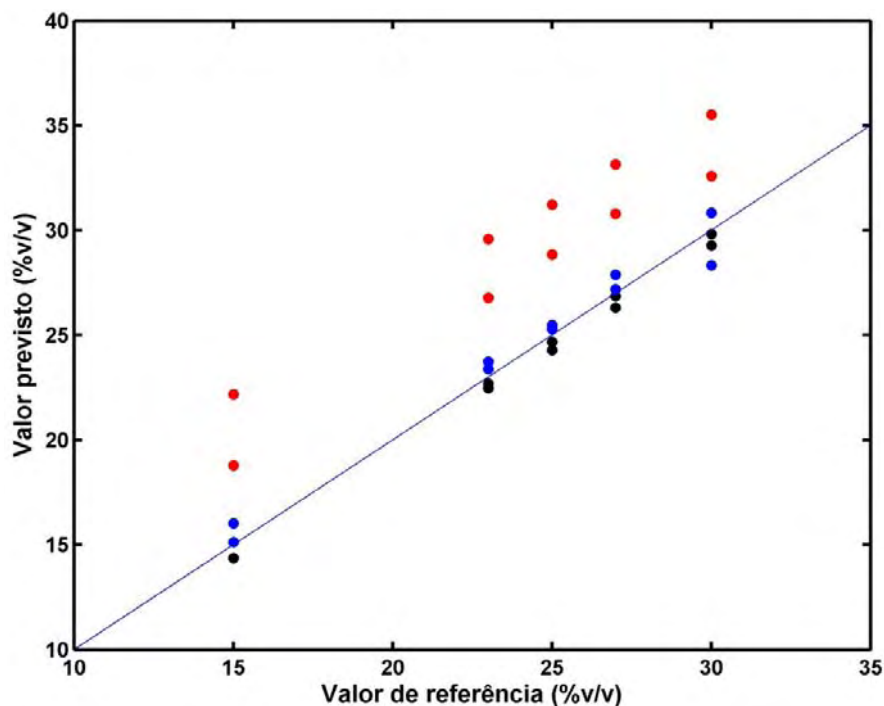


Figura 4.25 - Valores de referência de etanol *versus* valores previstos, na presença do interferente hexano, pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS.

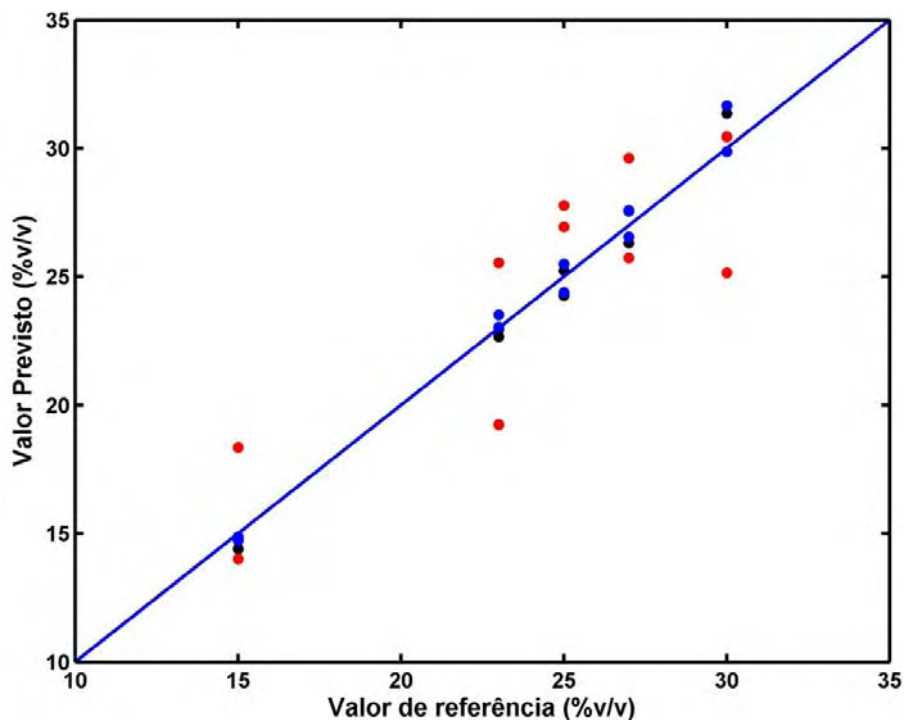


Figura 4.26 - Valores de referência de etanol *versus* valores previstos, na presença do interferente iso-octano, pelos modelos: ● SPA-MLR, ● SPA-SPE-MLR e ● PLS.

As Figuras 4.24, 4.25 e 4.26 mostram que em todos os casos a presença de um interferente nos espectros, prejudicou as previsões realizadas pelos modelos. Na determinação de etanol interferida por tolueno percebe-se

claramente que a perturbação ocorreu de forma mais intensa. As previsões realizadas pelo SPA-MLR foram as que mais se afastaram da bissetriz, confirmando o que já havia sido observado na **Tabela 4.9**.

Para verificar, assim como nas outras aplicações, se os menores RMSEPs são realmente diferentes, os testes t e F foram realizados. Os valores de t_{cal} e F_{cal} encontram-se apresentados nas **Tabelas 4.10 e 4.11**.

Tabela 4.10 - Valores de t_{cal} e t_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS). Os números de graus de liberdade encontram-se entre parêntesis.

	SPA-MLR	PLS	t_{crit}
SPA-SPE-MLR	t_{cal}	t_{cal}	t_{crit} (6)
Sem interferentes	2,78	2,92	2,45
(10%, 5%) de Tolueno	13,77	14,05	t_{crit} (9)
(10%, 5%) de Hexano	13,64	3,91	2,26
(10%, 5%) de Iso-octano	0,10	2,69	

Tabela 4.11 - Valores de F_{cal} e F_{crit} para comparação entre os modelos (SPA-SPE-MLR e SPA-MLR) e (SPA-SPE-MLR e PLS). Os números de graus de liberdade encontram-se entre parêntesis.

	SPA-MLR	PLS	F_{crit}
SPA-SPE-MLR	F_{cal}	F_{cal}	F_{crit} (6,6)
Sem interferentes	2,0816	1,6175	4,2839
(10%, 5%) de Tolueno	48,3331	1,8435	F_{crit} (9,9)
(10%, 5%) de Hexano	4,1345	6,8889	3,1789
(10%, 5%) de Iso-octano	18,8446	1,0604	

Com base nos valores de t e F apresentados nas **Tabelas 4.10 e 4.11** pode-se afirmar, ao nível de 95% de confiança, que os modelos SPA-MLR e PLS apresentam diferenças significativas quando comparados aos modelos SPA-SPE-MLR. Dessa forma os modelos que obtiveram menores RMSEPs podem ser considerados melhores e estatisticamente diferentes dos demais.

A **Figura 4.27** apresenta os espectros de tolueno, álcool e gasolina tipo C.

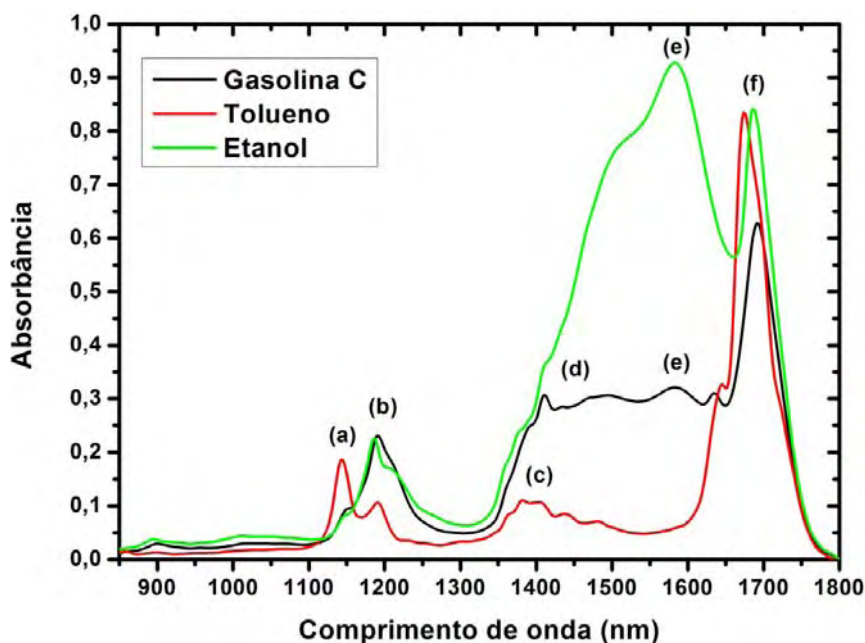


Figura 4.27 – Espectros de absorção de gasolina tipo C, tolueno e etanol. Regiões de absorção dos grupos: (a) C-H de aromático; (b) e (c) metila e metileno, (d) metila/OH e metileno/OH, (e) O-H e (f) OH/C-H de aromático (fonte: Pereira et al.^[76]).

Os maiores erros foram obtidos na previsão de amostras interferidas por tolueno. Uma justificativa seria a sobreposição de bandas de vibração do C-H de aromático e O-H de álcool, que ocorre na região de 1400 a 1800 nm.

Na **Figura 4.28** são apresentadas as variáveis selecionadas pelos modelos SPA-MLR e SPA-SPE-MLR.

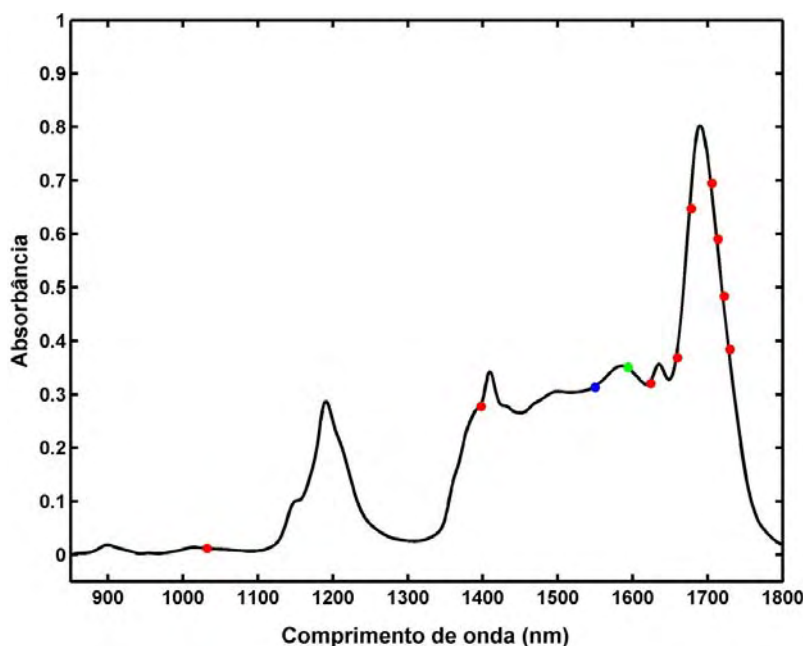


Figura 4.28 – Variáveis selecionadas para construção de modelos MLR para determinação de etanol usando: ● SPA-SPE-MLR sem interferentes e interferidas por hexano, ● SPA-SPE-MLR interferido por tolueno e iso-octano e ● SPA-MLR sem interferentes e interferidas por hexano, iso-octano e tolueno.

A **Figura 4.28** mostra que muitas das variáveis selecionadas pelo SPA-MLR encontram-se na região (1610 a 1800 nm) que ocorre a sobreposição entre as vibrações C-H de aromático e O-H de álcool, (**Figura 4.27**), o que justifica o grande valor de RMSEP obtido por esse modelo na determinação de amostras interferidas por tolueno. Maiores valores de RMSEPs para os modelos SPA-MLR na determinação de amostras interferidas por hexano e iso-octano, podem ser explicados pela seleção da variável em 1398 nm. Essa é uma região comum a absorções provenientes de estiramento C-H de metila e metileno presentes nos compostos hexano e iso-octano.

As variáveis selecionadas pelo SPA-SPE ocorreram principalmente nas regiões provenientes de ligações O-H de álcool, confirmando nesse caso a capacidade dessa estratégia de escolher melhores variáveis diretamente relacionadas ao parâmetro de interesse.

Como pôde ser visto, a estratégia proposta quando aplicada na determinação de álcool, produziu resultados de previsão menos afetados que os modelos SPA-MLR. Dentre todas as aplicações, o algoritmo proposto forneceu melhores resultados em duas das quatro aplicações, mostrando mais uma vez que essa estratégia é capaz de diminuir a influência dos interferentes.

Na determinação usando o conjunto de previsão interferido por tolueno, percebe-se claramente que o modelo foi capaz de evitar a região de maior sobreposição espectral provenientes do o álcool e do tolueno.

CAPÍTULO 5

Conclusões

5. Conclusões

Nesta dissertação foi apresentado uma nova maneira de selecionar variáveis considerando a estatística da matriz de previsão, de modo explorar possíveis problemas causados pela presença de constituintes que não estão presentes nos espectros das amostras de calibração.

A mudança aqui apresentada foi validada por meio de três aplicações sendo uma a dados simulados e a duas diferentes técnicas instrumentais (UV-VIS e NIR). Para resolver tal problema o SPE foi usado conjuntamente com o RMSE para a escolha das variáveis para modelos de calibração multivariada MLR, usando o SPA.

Os modelos SPA-SPE-MLR se mostraram superiores às técnicas de calibração multivariada PLS e ao SPA-MLR usando somente o RMSEV ou RMSECV para a escolha de variáveis. Os modelos MLR baseados nas variáveis selecionadas pelo SPA-SPE produziram, na presença de interferentes, modelos com menores valores de RMSEP na maioria das determinações realizadas, sendo esses estatisticamente diferentes, ao nível de 95% confiança.

Os modelos MLR obtidos com as variáveis selecionadas pelo SPA-SPE-MLR puderam ser obtidos satisfatoriamente a partir da validação por série de teste ou validação cruzada. Estes se mostraram de um modo geral mais parcimoniosos, selecionando em todos os casos um número de variáveis igual ou menor que o SPA.

O SPA-SPE-MLR foi capaz de selecionar variáveis em regiões onde a influência do interferente é minimizada, mostrando que o SPA-SPE atende à idéia pelo qual foi construído. Essa vantagem ficou bastante evidenciada na determinação de álcool contaminado por tolueno.

Dessa forma pode-se considerar o SPA-SPE, como uma melhoria significativa ao algoritmo das projeções sucessivas, incorporando esse novo método de escolha em seu código fonte para cálculos futuros.

5.1. Propostas futuras

Como proposta de possíveis melhorias e aplicações da nova estratégia, pode-se destacar:

- Testar, ao invés do conjunto, cada amostra de previsão separadamente de modo que cada modelo contemple a composição individual de cada amostra;
- Usar outras funções custo, de forma contemplar as informações das amostras de previsão;
- Avaliar o potencial dessa estratégia em outras técnicas analíticas.

CAPÍTULO 6

Referências Bibliográficas

6. Referencias Bibliográficas

1. SKOOG, D. A. E LEARY, J. J. *Principles of instrumental analysis*. 6. ed. New York : Saunders College Publishing, **1992**.
2. PONTES, M. J. C., et al. Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain. *Analytica Chimica Acta*. **642: 12, 2009**.
3. XIAOBO, Z., et al. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*. **667: 14, 2010**.
4. HIBBERT, D. B., et al. IUPAC project: A glossary of concepts and terms in chemometrics. *Analytica Chimica Acta*. **642: 3, 2009**.
5. FORINA, M, LANTERI, S e CASALE, M. Multivariate calibration. *Journal of Chromatography A*. **1158: 61, 2007**.
6. NAES, TORMOD, et al. *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester, UK : NIR Publications, **2002**.
7. MARTENS, H. e NAES, T. *Multivariate Calibration*. New York : John Wiley & Sons, **1989**.
8. HAIR, J. F., et al. *Análise multivariada de dados*. 5ª Ed. Porto Alegre : Bookmam, **2005**.
9. BEEBE, K. R., PELL, R. J. e SEASHOLTZ, M. B. *Chemometrics: A Practical Guide*. New York : John Wiley & Sons, **1998**.
10. FERREIRA, M. M. C., et al. Quimiometria I: Calibração Multivariada, Um Tutorial. *Química Nova*. **22: 724, 1999**.
11. KALIVAS, J. H., ROBERTS, N. e SUTTER, J. M. Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry. *Analytical Chemistry*. **61: 2024, 1989**.
12. WALMSLEY, A. D. Improved variabel selection procedure for multivariate linear regression. *Analytica Chimica Acta*. **354: 225, 1997**.

13. ARAÚJO, M. C. U., et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*. **57: 65, 2001.**
14. GALVÃO, R. K. H., et al. Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry. *Analytica Chimica Acta*. **443: 107, 2001.**
15. COELHO, C. J., et al. A Linear Semi-infinite Programming Strategy for Constructing Optimal Wavelet Transforms in Multivariate Calibration Problems. *J. Chem. Inf. Comput. Sci.* **43: 928, 2003.**
16. —. A solution to the wavelet transform optimization problem in multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*. **66: 205, 2003.**
17. BREITKREITZ, M. C., et al. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. *Analyst*. **128: 1204, 2003.**
18. DANTAS FILHO, H. A., et al. A strategy for selecting calibration samples for multivariate modelling. *Chemometrics and Intelligent Laboratory Systems*. **72: 83, 2004.**
19. GALVÃO, R. K. H., et al. Optimal wavelet filter construction using X and Ydata. *Chemometrics and Intelligent Laboratory Systems*. **70: 1, 2004.**
20. HONORATO, F. A., et al. Robust modeling for multivariate calibration transfer by the successive projections algorithm. *Chemometrics and Intelligent Laboratory Systems*. **76: 65, 2005.**
21. DANTAS FILHO, H. A., et al. Simultaneous Spectrometric Determination of Cu²⁺, Mn²⁺ and Zn²⁺ in Polivitaminic/Polimineral Drug Using SPA and GA Algorithms for Variable Selection. *J. Braz. Chem. Soc.* **16: 58, 2005.**
22. PONTES, M. J. C., et al. The successive projections algorithm for spectral variable selection in classification problems. *Chemometrics and Intelligent Laboratory Systems*. **78: 11, 2005.**

23. Galvão, R. K. H., et al. An application of subagging for the improvement of prediction accuracy of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*. **81: 60, 2006.**
24. AKHLAGHI, Y. E KOMPANY-ZAREH, M. Application of radial basis function networks and successive projections algorithm in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *Journal of Chemometrics*. **20: 1, 2006.**
25. CANECA, A. R., et al. Assessment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricating oils. *Talanta*. **70: 344, 2006.**
26. GALVÃO, R. K. H., et al. Cross-Validation for the Selection of Spectral Variables Using the Successive Projections Algorithm. *Journal of Brazilian Chemical Society*. **18: 1580, 2007.**
27. KOMPANY-ZAREH, M. E AKHLAGHI, Y. Correlation weighted successive projections algorithm as a novel method for variable selection in QSAR studies: investigation of anti-HIV activity of HEPT derivatives. *Journal of Chemometrics*. **21: 239, 2007.**
28. DI NEZIO, M. S., et al. Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water. *Microchemical Journal*. **85: 194, 2007.**
29. HONORATO, F. A., et al. Transferência de Calibração em Métodos Multivariados. *Química Nova*. **30: 1301, 2007.**
30. GALVÃO, R. K. H, et al. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. *Chemometrics and Intelligent Laboratory Systems*. **92: 83, 2008.**
31. RIBEIRO, M. P. A., et al. Multivariate calibration methods applied to the monitoring of the enzymatic synthesis of ampicilin. *Chemometrics and Intelligent Laboratory Systems*. **90: 169, 2008.**

32. GRÜNHUT, , M., et al. Flow-batch technique for the simultaneous enzymatic determination of levodopa and carbidopa in pharmaceuticals using PLS and successive projections algorithm. *Talanta*. **75: 950, 2008**.
33. PEREIRA, A. F. C., et al. NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection. *Food Research International*. **41: 341, 2008**.
34. YE, S., WANG, D. e MIN, S. Successive projections algorithm combined with uninformative variable elimination for spectral variable selection. *Chemometrics and Intelligent Laboratory Systems*. **91: 194, 2008**.
35. LI, LI-NA, LI, QING-BO e ZHANG, GUANG-JUN. A Weak Signal Extraction Method for Human Blood Glucose Noninvasive Measurement using Near Infrared Spectroscopy. *J Infrared Milli Terahz Waves*. **30: 1191, 2009**.
36. LIU, F. e HE, Y. Application of successive projections algorithm for variable selection to determine organic acids of plum vinegar. *Food Chemistry*. **115: 1430, 2009**.
37. FERRÃO, MARCO F., et al. LS-SVM: Uma Nova Ferramenta Quimiométrica para Regressão Multivariada. Comparação de Modelos de Regressão LS-SVM e PLS na Quantificação de Adulterantes em Leite em Pó Empregando NIR. *Química Nova*. **30: 852, 2007**.
38. SUYKENS, J. A. K., et al. *Least squares support vector machines*. Singapore : World Scientific, **2002**.
39. LIU, F., JIANG, Y. e HE, Y. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer. *Analytica Chimica Acta*. **635: 45, 2009**.
40. KHANMOHAMMADI, M., et al. Artificial neural network for quantitative determination of total protein in yogurt by infrared spectrometry. *Microchemical Journal*. **91: 47, 2009**.

41. GAMBARRA-NETO, F. F., et al. Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis. *Talanta*. **77**: 1660, 2009.
42. LIU, F., HE, Y. e SUN, G. Determination of Protein Content of Auricularia auricula Using Near Infrared Spectroscopy Combined with Linear and Nonlinear Calibrations. *J. Agric. Food Chem.* **57**: 4520, 2009.
43. WU, D., et al. Exploring Near and Midinfrared Spectroscopy to Predict Trace Iron and Zinc Contents in Powdered Milk. *J. Agric. Food Chem.* **57**: 1697, 2009.
44. GOODARZI, M., FREITAS, M. P. e JENSEN, R. Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3 Inhibitory Activities. *J. Chem. Inf. Model.* **49**: 824, 2009.
45. MOREIRA, E. D. T., et al. Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection. *Talanta*. **79**: 1260, 2009.
46. GOUDARZI, N., et al. QSPR Modeling of Soil Sorption Coefficients (Koc) of Pesticides Using SPA-ANN and SPA-MLR. *J. Agric. Food Chem.* **57**: 7153, 2009.
47. CHEN, X., et al. Application of a hybrid variable selection method for the classification of rapeseed oils based on ¹H NMR spectral analysis. *Eur Food Res Technol.* **230**: 981, 2010.
48. SANTOS, E. O., et al. Determination of degree of polymerization of insulating paper using near infrared spectroscopy and multivariate calibration. *Vibrational Spectroscopy*. **52**: 154, 2010.
49. WU, D., et al. Hybrid variable selection in visible and near-infrared spectral analysis for non-invasive quality determination of grape juice. *Analytica Chimica Acta*. **659**: 229, 2010.
50. SOARES, A. S., et al. Improving the Computational Efficiency of the Successive Projections Algorithm by using a Sequential Regression Implementation: A Case Study Involving NIR Spectrometric Analysis of Wheat Samples. *J. Braz. Chem. Soc.* **21**: 760, 2010.

51. SOARES, A. S., et al. Multi-Core Computation in Chemometrics: Case Studies of Voltammetric and NIR Spectrometric Analyses. *J. Braz. Chem. Soc.* **21: 1626, 2010.**
52. MARTINS, M. N., GALVÃO, R. K. H. e PIMENTEL, M. F. Multivariate Calibration Transfer Employing Variable Selection and Subagging. *J. Braz. Chem. Soc.* **21: 127, 2010.**
53. SOUTO, U. T. C. P., et al. UV-Vis spectrometric classification of coffees by SPA-LDA. *Food Chemistry.* **119: 368, 2010.**
54. PEREIRA, C. F. e PASQUINI, C. A Flow System for Generation of Concentration Perturbation in Two-Dimensional Correlation Near-Infrared Spectroscopy: Application to Variable Selection in Multivariate Calibration. *Applied Spectroscopy.* **64: 507, 2010.**
55. BRO, RAMUS. Multivariate calibration What is in chemometrics for the analytical chemist ? *Analytica Chimica Acta.* **500: 185, 2003.**
56. ZEAITER, M., ROGER, J. M. e BELLON-MAUREL, V. Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods. *Trends in Analytical Chemistry.* **24:5, 2005.**
57. BRERETON, R. G. Introduction to multivariate calibration in analytical chemistry. *Analyst.* **125: 2125, 2000.**
58. WILLIAMS, P. e NORRIS, K. *Near-Infrared Technology in the Agricultural and Food Industries.* St. Paul, USA : Amer Assn of Cereal Chemists, **2001.**
59. PIMENTEL, M. F., GALVÃO, R. K. H. e ARAÚJO, M. C. U. Recomendações para calibração em Química Analítica parte 2. Calibração Multianálito. *Química Nova.* **31: 462, 2008.**
60. CHARNET, R., et al. *Análise de modelos de regressão linear com aplicações.* Campinas : Editora da Unicamp, **1999.**
61. JOCHUM, C., JOCHUM, P. e KOWALSKI, B. Error propagation and optimal performance in multicomponent analysis. *Analytical Chemistry.* **53: 85, 1981.**

62. WOLD, S., SJOSTRON, M. e ERIKSSON, L. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. **58: 109, 2001.**
63. NADLER, B. e COIFMAN, R. R. The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration. *J. Chemometrics*. **19: 107, 2005.**
64. GALVÃO, R. K. H. e ARAÚJO, M. C. U. variable selection. in: B. WALCZAK, R. T. FERRÉ e S. BROWN. *Linear regression modelling*. Comprehensive chemometrics, **2009.**
65. Centner, V., et al. Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry*. **68: 3851, 1996.**
66. CHEN, D., et al. A Background and noise elimination method for quantitative calibration of near infrared spectra. *Analytica Chimica Acta*. **511: 37, 2004.**
67. NORGAARD, L., et al. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy*. **54: 413, 2000.**
68. LEARDI, R. e NORGAARD, L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *J. Chemometrics*. **18: 486, 2004.**
69. WESTAD, F. e KERMIT, M. Independent Component Analysis. in: B. WALCZAK, R. T. FERRÉ e S. BROWN. *Linear regression modelling*. Comprehensive chemometrics Linear regression modelling, **2009.**
70. LEARDI, R., SEASHOLTZ, M. B. e PELL, R. J. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Analytica Chimica Acta*. **461: 189, 2002.**
71. HAALAND, D. M. e THOMAS, E. V. Partial Least-Squares Methods for Spectral Analyses. 1. to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. *Anal. Chem.* **60: 1193, 1988.**

72. SKOOG, D. A., et al. *Fundamentos de química analítica*. São Paulo : Thomson, **2006**.
73. JOHN, STEPHEN HASWELL. *Practical Guide to Chemometrics*. New York : Marcel Dekker, **1992**.
74. MAESSCHALCK, R., JOUAN-RIMBAUD, D. e MASSART, D. L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*. **50: 1, 2000**.
75. Nunes, P. G. A. Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às espectrometrias UV-VIS e NIR. *Tese de doutorado*. João Pessoa : UFPB, **2008**.
76. Pereira, C. F. Uso da espectroscopia de correlação bidimensional (2D) e construção e avaliação de um espectropolarímetro para a região do infravermelho próximo (NIR). *Tese de doutorado*. Campinas, SP : UNICAMP, **2006**.
77. ASTM E 1655-00; Standards Practices for Infrared Multivariate.
78. ROGGO, Y., et al. Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data. *Journal of Molecular Structure*. **654: 253, 2003**.