

**JÚLIA TRISTÃO DO CARMO ROCHA**

**EMPREGO DE ESPECTROMETRIA NO  
INFRAVERMELHO E MÉTODOS QUIMIOMÉTRICOS  
PARA A IDENTIFICAÇÃO E QUANTIFICAÇÃO DE  
PETRÓLEOS A PARTIR DE MISTURAS DE FRAÇÕES  
DE DIESEL**

Dissertação apresentada ao Programa de Pós-Graduação em Química do Centro de Ciências Exatas da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Mestre em Química, na área de concentração em Química de Produtos Naturais.

Orientador: Prof. Dr. Reginaldo Bezerra dos Santos.

Co-orientador: Prof. Dr. Eustáquio Vinícius Ribeiro de Castro

VITÓRIA

2009

**Emprego de Espectrometria no Infravermelho e Métodos  
Quimiométricos para a Identificação e Quantificação de Petróleos a  
Partir de Misturas de Frações de Diesel**

Júlia Tristão do Carmo Rocha

Dissertação de mestrado submetida ao Programa de Pós-Graduação em Química do Centro de Ciências Exatas da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do Grau de Mestre em Química, área de Química de Produtos Naturais.

Aprovada em 25 de Setembro de 2009

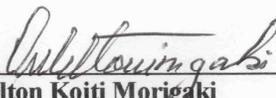
COMISSÃO EXAMINADORA



**Prof. Dr. Reginaldo Bezerra dos Santos**  
Universidade Federal do Espírito Santo  
Orientador



**Prof.<sup>a</sup> Dr.<sup>a</sup> Denise Rocco de Sena**  
Instituto Federal do Espírito Santo



**Prof. Dr. Milton Koiti Morigaki**  
Universidade Federal do Espírito Santo



**Prof. Dr. Eustáquio Vinicius Ribeiro de Castro**  
Universidade Federal do Espírito Santo  
Co-Orientador

À Letícia,  
que chegou trazendo muita alegria e dando um novo  
sentido à minha vida

Ao Hugo,  
pelo companheirismo e amor

Aos meus pais,  
minha maior inspiração e meu maior incentivo

## ÍNDICE

AGRADECIMENTOS.....	i
LISTA DE ABREVIATURAS OU SIGLAS .....	ii
ÍNDICE DE FIGURAS .....	iii
ÍNDICE DE TABELAS .....	vii
ABSTRACT.....	x
RESUMO.....	xii
<b>1 INTRODUÇÃO</b> .....	1
<b>2 FUNDAMENTOS TEÓRICOS</b> .....	4
2.1 O PETRÓLEO .....	5
<b>2.1.1 Diesel</b> .....	6
2.2 ESPETROMETRIA NO INFRAVERMELHO .....	8
<b>2.2.1 Reflexão total atenuada</b> .....	9
2.3 ANÁLISE MULTIVARIADA: QUIMIOMETRIA.....	11
<b>2.3.1 Análise por componentes principais (PCA)</b> .....	13
2.3.1.1 Conceitos matemáticos .....	16
<b>2.3.2 Métodos de calibração multivariada</b> .....	17
2.3.2.1 Regressão por componentes principais – PCR .....	18
2.3.2.2 Mínimos quadrados parciais – PLS .....	19
2.3.2.3 Validação dos modelos de regressão.....	22
2.3.2.3.1 Escolha do número de variáveis latentes (VLs) .....	23
2.3.2.3.2 Cálculo dos erros .....	24
2.3.2.3.2.1 Auto-predição .....	25
2.3.2.3.2.2 Validação Cruzada .....	25
2.3.2.3.2.3 Previsão de amostras externas .....	27
2.3.2.3.2.4 O problema do overfitting .....	28
2.3.2.3.3 Detecção de amostras anômalas – outliers .....	29
2.3.2.4 Método de seleção de variáveis .....	30
2.3.2.4.1 PLS por intervalos (i-PLS).....	31
<b>3 OBJETIVOS</b> .....	33
3.1 OBJETIVOS GERAIS.....	34
3.2 OBJETIVOS ESPECÍFICOS.....	34
<b>4 METODOLOGIA</b> .....	35
4.1 PREPARO DAS AMOSTRAS.....	36
<b>4.1.2 Misturas</b> .....	36
4.2 DENSIDADE E VISCOSIDADE .....	39
4.3 INFRAVERMELHO .....	39
4.4 APLICATIVO PARA CRIAÇÃO DOS MODELOS .....	40
4.5 SELEÇÃO DE AMOSTRAS.....	40
4.6 SELEÇÃO DA FAIXA ESPECTRAL .....	40
4.7 ANÁLISE POR COMPONENTES PRINCIPAIS (PCA).....	41
<b>4.7.1 Separação de amostras</b> .....	42
<b>4.7.2 Seleção das componentes principais</b> .....	42
<b>4.7.3 Cálculos das componentes principais das amostras de validação</b> .....	42
4.8 ELABORAÇÃO DOS MODELOS DE CALIBRAÇÃO .....	43
4.9 AVALIAÇÃO DOS MODELOS DE CALIBRAÇÃO .....	43
5.1 IDENTIFICAÇÃO DAS FRAÇÕES .....	46
5.2 QUANTIFICAÇÃO DAS FRAÇÕES.....	48

<b>5.2.1 Quantificação da fração A</b> .....	48
5.2.1.1 Aplicação do método PLS para a quantificação da fração A .....	48
5.2.1.1.2 <i>Definição dos melhores modelos obtidos por PLS na quantificação da fração A e predição de amostras externas</i> .....	49
5.2.1.2 Aplicação do método PCR para a quantificação da fração A .....	54
5.2.1.2.1 <i>Definição dos melhores modelos obtidos por PCR na quantificação da fração A e predição de amostras externas</i> .....	55
<b>5.2.2 Quantificação da fração B</b> .....	58
5.2.2.1 Aplicação do método PLS para a quantificação da fração B .....	58
5.2.2.1.1 <i>Definição dos melhores modelos obtidos por PLS na quantificação da fração B e predição de amostras externas</i> .....	60
<b>5.2.2.2 Aplicação do método PCR para a quantificação da fração B</b> .....	64
5.2.2.2.1 <i>Definição dos melhores modelos obtidos por PCR na quantificação da fração B e predição de amostras externas</i> .....	65
<b>5.2.3 Quantificação da fração C</b> .....	67
5.2.3.1 Aplicação do método PLS para a quantificação da fração C .....	67
5.2.3.1.1 <i>Definição dos melhores modelos obtidos por PLS na quantificação da fração C e predição de amostras externas</i> .....	69
5.2.3.2 Aplicação do método PCR para a quantificação da fração C .....	73
5.2.3.2.1 <i>Definição dos melhores modelos obtidos por PCR na quantificação da fração C e predição de amostras externas</i> .....	73
<b>5.2.4 Quantificação das frações A, B e C simultaneamente</b> .....	75
5.2.4.1 Aplicação do método PLS para a quantificação das frações A, B e C .....	75
5.2.4.1.1 <i>Definição dos melhores modelos obtidos por PLS na quantificação das frações A, B e C e predição de amostras externas</i> .....	77
5.2.4.2 Aplicação do método PCR para a quantificação das frações A, B e C .....	80
5.2.4.2.1 <i>Definição dos melhores modelos obtidos por PCR na quantificação simultânea das frações A, B e C e predição de amostras externas</i> .....	81
<b>5.2.5 Predição da densidade</b> .....	83
5.2.5.1 Aplicação do método PLS para a predição da densidade .....	83
5.2.5.1.1 <i>Definição dos melhores modelos obtidos por PLS na determinação da densidade e predição de amostras externas</i> .....	85
5.2.5.2 Aplicação do método PCR para a predição da densidade .....	88
5.2.5.2.1 <i>Definição dos melhores modelos obtidos por PCR na determinação da densidade e predição de amostras externas</i> .....	88
<b>5.2.6 Predição da viscosidade</b> .....	90
5.2.6.1 Aplicação do método PLS para a predição da viscosidade .....	90
5.2.6.1.1 <i>Definição dos melhores modelos obtidos por PLS na determinação da densidade e predição de amostras externas</i> .....	91
5.2.6.2 Aplicação do método PCR para a predição da viscosidade .....	93
5.2.6.2.1 <i>Definição dos melhores modelos obtidos por PCR na determinação da viscosidade e predição de amostras externas</i> .....	94
<b>6 CONCLUSÕES</b> .....	96
6.1 SUGESTÕES PARA TRABALHOS FUTUROS .....	97
<b>REFERÊNCIAS BIBLIOGRÁFICAS</b> .....	99
<b>ANEXO A</b> .....	104

## **AGRADECIMENTOS**

Agradeço a Deus, meu guia, pela providencia de todas as coisas, por minha existência, alegria, força e saúde.

Ao meu orientador, professor Reginaldo Bezerra dos Santos, pela paciência em me conduzir na execução deste trabalho, pelas valiosas críticas e sugestões, pelos ensinamentos, pelo exemplo de profissionalismo, pelas oportunidades de crescimento profissional e pessoal, pela amizade e pelo apoio constante.

Ao professor Eustáquio Vinícius Ribeiro de Castro, pela ajuda na análise estatística, pelas sugestões, pela amizade e apoio durante todo o período de execução deste trabalho.

À Letícia, pela sua existência.

Ao Hugo, pelo amor, incentivo, paciência e atenção.

Aos meus pais, Rita e Paulo, pela educação, carinho e apoio.

Aos meus irmãos, Carmem e Lucas, pela ajuda e compreensão em cada momento difícil.

À Universidade Federal do Espírito Santo, em especial ao programa de pós-graduação em química da UFES, por todos os recursos que permitiram a elaboração deste trabalho. Em especial; ao LabPetro pelo suporte prestado na execução deste trabalho.

A CENPES/ PETROBRAS pelo apoio financeiro prestado através do projeto e da bolsa.

Aos amigos do LabPetro-UFES pela ajuda e em especial a Cristina Sad pela dedicação e auxílio e a Renzo Correa Silva pela grande ajuda na análise estatística e pela amizade.

Aos amigos do CENPES/PETROBRAS, especialmente a Lílian Carmem Medina, pelo apoio no desenvolvimento deste trabalho.

Aos amigos do LEC/UFMG pela ajuda.

Aos amigos do laboratório de Química Orgânica, pela amizade e colaboração.

Aos amigos do mestrado.

Agradeço a todos aqueles que, de alguma forma, contribuíram para a conclusão deste trabalho.

## LISTA DE ABREVIATURAS OU SIGLAS

ATR – Reflectância total atenuada, do inglês *Attenuated total reflectance*

CENPES – Centro de Pesquisas e Desenvolvimento Leopoldo A. Miguez de Mello

CV – Validação cruzada

FTIR – ATR – Espectroscopia no infravermelho com transformada de Fourier e reflectancia total atenuada, do inglês, *Fourier transform infrared attenuated total reflectance*.

HCA - Análise por Agrupamentos Hierárquicos

iPLS – Mínimos Quadrados Parciais por Intervalos

IV – Infravermelho

LABPETRO – Laboratório de Pesquisa e Desenvolvimento de Metodologias para Análises do Petróleo

LEC – Laboratório de Ensaio de Combustível

MIR – Infravermelho Médio

MLR - Regressão Linear Múltipla, do inglês, *Multilinear Regression*

PC – Componentes Principais, do inglês, *Principal Components*

PCA – Análise por Componentes Principais, do inglês *Principal Components Analysis*

PCR – Regressão por Componentes Principais

PETROBRAS – Petróleo Brasileiro S. A.

PLS – Mínimos Quadrados Parciais, do inglês *Partial least squares*

PLSR – Regressão por Mínimos Quadrados Parciais

$R^2$  – Coeficiente de Correlação Linear

RMSE – Raiz Quadrada do Erro Quadrático Médio

RMSEC - Raiz Quadrada do Erro quadrático Médio de Calibração

RMSECV – Raiz Quadrada do Erro Quadrático Médio de Validação Cruzada

RMSEP - Raiz Quadrada do Erro Quadrático Médio de Previsão

siPLS – Mínimos Quadrados Parciais por Sinergismo de Intervalos

UFES – Universidade Federal do Espírito Santo

UFMG – Universidade Federal de Minas Gerais

VL – Variável Latente

## ÍNDICE DE FIGURAS

Figura 2.1 Representação da ATR.....	10
Figura 2.2 Representação da decomposição da matriz de dados X.....	16
Figura 2.3 Ilustração teórica da validação de um modelo .....	24
Figura 2.4 Comparação entre os erros de modelagem e de predição .....	28
Figura 5.1 Gráfico dos <i>scores</i> para o intervalo 2 relacionando as PCs 1, 2 e 3.....	46
Figura 5.2 Gráfico dos <i>scores</i> para o intervalo 3 relacionando as PCs 2, 3 e 4.....	46
Figura 5.3 Perfil do RMSECV para a fração A com modelos construídos para as cinco melhores regiões.....	48
Figura 5.4 Perfil do RMSECV da fração A para a associação de duas regiões para as 5 melhores combinações.....	49
Figura 5.5 Perfil do RMSECV, RMSEC e RMSEP variando-se o número de variáveis latentes para a fração A na associação dos intervalos 2 e 3 .....	50
Figura 5.6 Perfil do RMSECV, RMSEC e RMSEP variando-se o número de variáveis latentes para a fração A na região I .....	51
Figura 5.7 Gráfico com o resultado de predição da fração A em amostras de calibração no intervalo 2 .....	52
Figura 5.8 Gráfico com o resultado de predição da fração A em amostras de validação no intervalo 2.....	52
Figura 5.9 Gráfico com o resultado de predição da fração A em amostras de calibração na associação dos intervalos 2 e 3 .....	53
Figura 5.10 Gráfico com o resultado de predição da fração A em amostras de validação na associação dos intervalos 2 e 3 .....	53
Figura 5.11 Gráfico com o resultado de predição da fração A em amostras de validação no intervalo 2 por PCR.....	57
Figura 5.12 Gráfico com o resultado de predição da fração A em amostras de validação na associação dos intervalos 2 e 3 por PCR.....	57
Figura 5.13 Perfil do RMSECV para a fração B com modelos construídos para as cinco melhores regiões.....	58
Figura 5.14 Perfil do RMSECV da fração B para a associação de duas regiões para as 5 melhores combinações.....	59
Figura 5.15 Perfil do RMSECV da fração B para a associação de três regiões para as 5 melhores combinações.....	59

Figura 5.16 Gráfico com o resultado de predição da fração B em amostras de calibração no intervalo 2 .....	61
Figura 5.17 Gráfico com o resultado de predição da fração B em amostras de validação no intervalo 2.....	61
Figura 5.18 Gráfico com o resultado de predição da fração B em amostras de calibração na associação dos intervalos 2 e 13 .....	62
Figura 5.19 Gráfico com o resultado de predição da fração B em amostras de validação na associação dos intervalos 2 e 13 .....	62
Figura 5.20 Gráfico com o resultado de predição da fração B em amostras de calibração na associação dos intervalos 2, 3 e 13 .....	62
Figura 5.21 Gráfico com o resultado de predição da fração B em amostras de validação na associação dos intervalos 2, 3 e 13 .....	63
Figura 5.22 Gráfico com o resultado de predição da fração B em amostras externas no intervalo 2 por PCR .....	66
Figura 5.23 Gráfico com o resultado de predição da fração B em amostras externas na associação dos intervalos 2 e 13 por PCR .....	66
Figura 5.24 Gráfico com o resultado de predição da fração B em amostras externas na associação dos intervalos 2, 3 e 13 por PCR .....	66
Figura 5.25 Perfil do RMSECV para a fração C com modelos construídos para as cinco melhores regiões.....	68
Figura 5.26 Perfil do RMSECV da fração C para a associação de duas regiões para as 5 melhores combinações.....	68
Figura 5.27 Perfil do RMSECV da fração C para a associação de três regiões para as 5 melhores combinações.....	69
Figura 5.28 Gráfico com o resultado de predição da fração C em amostras de calibração no intervalo 2 .....	71
Figura 5.29 Gráfico com o resultado de predição da fração C em amostras de validação no intervalo 2.....	71
Figura 5.30 Gráfico com o resultado de predição da fração C em amostras de calibração na associação dos intervalos 2 e 13 .....	71
Figura 5.31 Gráfico com o resultado de predição da fração C em amostras de validação na associação dos intervalos 2 e 13 .....	72
Figura 5.32 Gráfico com o resultado de predição da fração C em amostras externas no intervalo 2 por PCR .....	74

Figura 5.33 Perfil do RMSECV para as frações A, B e C com modelos construídos para as cinco melhores regiões .....	76
Figura 5.34 Perfil do RMSECV para as frações A, B e C com modelos construídos para a associação de 2 regiões para as cinco melhores regiões .....	76
Figura 5.35 Gráfico com o resultado de predição da fração A, para o modelo global, em amostras de calibração no intervalo 2 .....	77
Figura 5.36 Gráfico com o resultado de predição da fração B, para o modelo global, em amostras de calibração no intervalo 2 .....	78
Figura 5.37 Gráfico com o resultado de predição da fração C, para o modelo global, em amostras de calibração no intervalo 2 .....	78
Figura 5.38 Gráfico com o resultado de predição da fração A, para o modelo global, em amostras de validação no intervalo 2.....	78
Figura 5.39 Gráfico com o resultado de predição da fração B, para o modelo global, em amostras de validação no intervalo 2.....	79
Figura 5.40 Gráfico com o resultado de predição da fração C, para o modelo global, em amostras de validação no intervalo 2.....	79
Figura 5.41 Gráfico com o resultado de predição da fração A, para o modelo global, em amostras de validação no intervalo 2 por PCR.....	82
Figura 5.42 Gráfico com o resultado de predição da fração B, para o modelo global, em amostras de validação no intervalo 2 por PCR.....	82
Figura 5.43 Gráfico com o resultado de predição da fração C, para o modelo global, em amostras de validação no intervalo 2 por PCR.....	82
Figura 5.44 Perfil do RMSECV para a densidade com modelos construídos para as cinco melhores regiões.....	84
Figura 5.45 Perfil do RMSECV para a densidade para a associação de duas regiões para as 5 melhores combinações .....	84
Figura 5.46 Gráfico com o resultado de predição da densidade em amostras de calibração no intervalo 2 .....	86
Figura 5.47 Gráfico com o resultado de predição da densidade em amostras de calibração na associação dos intervalos 2 e 13 .....	86
Figura 5.48 Gráfico com o resultado de predição da densidade em amostras de validação no intervalo 2.....	87
Figura 5.49 Gráfico com o resultado de predição da densidade em amostras de validação na associação dos intervalos 2 e 13 .....	87

Figura 5.50 Gráfico com o resultado de predição da densidade em amostras de validação no intervalo 2 por PCR .....	89
Figura 5.51 Gráfico com o resultado de predição da densidade em amostras de validação na associação dos intervalos 2 e 13 por PCR.....	90
Figura 5.52 Perfil do RMSECV para a viscosidade com modelos construídos para as cinco melhores regiões.....	91
Figura 5.53 Perfil do RMSECV, RMSEC e RMSEP variando-se o número de variáveis latentes para a viscosidade no intervalo 2 .....	92
Figura 5.54 Gráfico com o resultado de predição da viscosidade em amostras de calibração na associação no intervalo 2.....	93
Figura 5.55 Gráfico com o resultado de predição da viscosidade em amostras de validação no intervalo 2.....	93
Figura 5.56 Gráfico com o resultado de predição da viscosidade em amostras de validação no intervalo 2 por PCR.....	95
Figura A1.1 Espectros obtidos das frações A, B e C.....	104

## ÍNDICE DE TABELAS

Tabela 2.1 Frações de petróleo e suas faixas de temperatura .....	6
Tabela 2.2 Regiões espectrais do infravermelho .....	8
Tabela 2.3 Principais características das diferentes composições dos elementos de ATR disponíveis .....	11
Tabela 4.1 Temperaturas de destilação das frações do petróleo I.....	37
Tabela 4.2 Temperaturas de destilação das frações do petróleo II.....	37
Tabela 4.3 Temperaturas de destilação das frações do petróleo III.....	37
Tabela 4.4 Proporção, em volume, das frações A, B e C nas misturas formadas.....	38
Tabela 4.5 Faixa de absorvância dos 20 intervalos formados .....	41
Tabela 5.1 Covariância acumulada para as 5 primeiras PCs nos intervalos 2 e 3 .....	47
Tabela 5.2 Correlações entre os volumes das frações A, B e C nas misturas e as 5 primeiras PCs nos intervalos 2 e 3 .....	47
Tabela 5.3 Seleção do número de variáveis latentes para cada modelo proposto para a quantificação da fração A.....	51
Tabela 5.4 Valores de $R^2$ e RMSEP para os modelos construídos para a quantificação da fração A.....	51
Tabela 5.5 Modelos selecionados para a quantificação da fração A.....	52
Tabela 5.6 Médias e Desvio Padrões para os modelos obtidos na quantificação da Fração A por PLS.....	54
Tabela 5.7 Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação da Fração A.....	55
Tabela 5.8 Modelos, da fração A, que apresentaram $R^2$ maior que 0,90.....	56
Tabela 5.9 Modelos, da fração A, que apresentaram $R^2$ maior que 0,90 e RMSEP menor que 0,020 por PCR.....	56
Tabela 5.10 Média e desvio padrões para os modelos obtidos na quantificação da Fração A por PCR .....	57
Tabela 5.11 Seleção do número de variáveis latentes para cada modelo proposto para a quantificação da fração B.....	60
Tabela 5.12 Valores de $R^2$ e RMSEP para os modelos construídos para a quantificação da fração B.....	60
Tabela 5.13 Modelos selecionados para a quantificação da fração B.....	61

Tabela 5.14 Média e desvio padrões para os modelos obtidos na quantificação da Fração B por PLS.....	63
Tabela 5.15 Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação da Fração B.....	64
Tabela 5.16 Modelos, da fração B, que apresentaram $R^2$ maior que 0,90.....	65
Tabela 5.17 Modelos, da fração B, que apresentaram $R^2$ maior que 0,90 e RMSEP menor que 0,020 por PCR.....	65
Tabela 5.18 Média e desvio padrões para os modelos obtidos na quantificação da Fração B por PCR.....	67
Tabela 5.19 Temperaturas Seleção do número de variáveis latentes para cada modelo proposto para a quantificação da fração C.....	70
Tabela 5.20 Valores de $R^2$ e RMSEP para os modelos construídos para a quantificação da fração C.....	70
Tabela 5.21 Modelos selecionados para a quantificação da fração C por PLS.....	70
Tabela 5.22 Média e desvio padrões para os modelos obtidos na quantificação da Fração C por PLS.....	72
Tabela 5.23 Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação da Fração C.....	73
Tabela 5.24 Modelos, da fração C, que apresentaram $R^2$ maior que 0,90.....	74
Tabela 5.25 Média e desvio padrões para o modelo obtido na quantificação da Fração C por PCR.....	75
Tabela 5.26 Valores de $R^2$ e RMSEP para o modelo obtido para a quantificação das frações A, B e C por PLS.....	77
Tabela 5.27 Média e desvio padrões para o modelo obtido na quantificação das Frações A, B e C por PLS.....	79
Tabela 5.28 Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação das frações A, B e C.....	80
Tabela 5.29 Modelos, das frações A B e C, que apresentaram $R^2$ maior que 0,90 por PCR.....	81
Tabela 5.30 Modelos, das frações A, B e C, que apresentaram $R^2$ maior que 0,90 e RMSEP menor que 0,02 por PCR.....	81
Tabela 5.31 Média e desvio padrões para o modelo obtidos na quantificação das Frações A, B e C por PCR.....	83

Tabela 5.32 Seleção do número de variáveis latentes para cada modelo proposto para a predição da densidade .....	85
Tabela 5.33 Valores de $R^2$ e RMSEP para os modelos construídos para a predição da densidade .....	85
Tabela 5.34 Modelos selecionados para a predição da densidade por PLS .....	86
Tabela 5.35 Média e desvio padrões para os modelos obtidos na predição da densidade por PLS .....	87
Tabela 5.36 Número de PCs utilizadas na construção dos modelos, por PCR, na predição da densidade .....	88
Tabela 5.37 Modelos que apresentaram $R^2$ maior que 0,90, por PCR, na predição da densidade .....	89
Tabela 5.38 Média e desvio padrões para o modelo obtido na predição da densidade por PCR .....	90
Tabela 5.39 Valores de $R^2$ e RMSEP para o modelo construído para a predição da viscosidade.....	92
Tabela 5.40 Número de PCs utilizadas na construção dos modelos, por PCR, na predição da viscosidade .....	94
Tabela 5.41 Modelos que apresentaram $R^2$ maior que 0,90, por PCR, na predição da viscosidade .....	94
Tabela 5.42 Média e Desvio Padrão para o modelo selecionado para a predição da viscosidade por PCR .....	95

---

## ABSTRACT

### Use of Infrared Spectrometry and Chemometrics Methods for the Identification and Quantification of the Petroleum based in Mixtures of Diesel Fractions

---

**Keywords:** Infrared, Chemometrics, Diesel, Petroleum

The knowledge of the petroleum's composition and its products is an essential necessity in a refinery for the adjustment of the process conditions. The infrared spectrometry, associated to chemometrics tools, has been presented as a useful tool to take care of this necessity for its great potentiality and applications.

In such a way, due to applicability of the chemometrics tools, in this work the discrimination for PCA (principal components analysis) and the quantification for intervals partial least square (iPLS), for synergism of intervals (siPLS) and for principal components regression (PCR) of three fractions of distillation are proposals, proceeding from three different petroleums (A, B and C), in the band of diesel attainment (210°C 260° C approximately) from the absorption in the medium infrared (MIR).

For that, 150 mixtures of the 3 fractions were produced, varying it enters 0 and 1,33% between a sample and another one. Later, specters were gotten, in the region of the MIR, of all the produced mixtures and after that the data were submitted to the chemometrics analyses. 100 samples were used for the calibration of model and 50 for its validation.

First it was carried through the analysis for principal components and from it it was possible to discriminate the samples in accordance with its predominant fraction.

After the identification of the samples, it was continued in the attempt of quantifying the 3 fractions, individually and simultaneously, in each one of them through the application of the methods iPLS, siPLS and PCR. The gotten models had been, then, evaluated for its the  $R^2$ , the RMSEP and the number of latent variable (VL) or principal components (PC) used.

iPLS, siPLS and PCR models were also developed for the determination of the samples density and viscosity, which were evaluated in the same way.

All the generated models had gotten resulted sufficiently satisfactory. However comparing the models siPLS and iPLS, it was verified that it did not have significant difference between them not justifying, therefore, the use of the models siPLS. About the models generated for PCR, the same ones had also gotten resulted sufficiently next to supplied for the method iPLS not having a justifiable preference in the choice between a method and another one.

With this, it demonstrates, therefore, that it is possible to discriminate and to quantify petroleums of distinct origins from its fractions using spectrometry in the region of the MIR and chemometrics techniques

---

## RESUMO

### Emprego de Espectrometria no Infravermelho e Métodos Quimiométricos para a Identificação e Quantificação de Petróleos a partir de Misturas de Frações de Diesel

---

**Palavras-chave:** Infravermelho, Quimiometria, Diesel, Petróleo

O conhecimento da composição do petróleo e de seus produtos é uma necessidade imprescindível numa refinaria para o ajuste das condições do processo. A espectrometria no infravermelho associada a ferramentas quimiométricas tem se apresentado como uma ferramenta útil para atender esta necessidade, por sua grande potencialidade e aplicações.

Desta forma, devido à aplicabilidade das ferramentas quimiométricas, neste trabalho são propostas a discriminação por PCA (análise por componentes principais) e a quantificação por mínimos quadrados parciais por intervalos (iPLS) e por sinergismo de intervalos (siPLS) e regressão por componentes principais (PCR) de três frações de destilação, provenientes de 3 petróleos diferentes (A, B e C), na faixa de obtenção do diesel (210°C a 260° C aproximadamente) a partir da absorção no infravermelho médio (MIR).

Para isso, foram produzidas 150 misturas das 3 frações, variando-se o entre 0 e 1,33% v/v entre uma amostra e outra. Posteriormente, foram obtidos espectros, na região do MIR, de todas as misturas produzidas e em seguida os dados foram submetidos às análises quimiométricas. Foram utilizadas 100 amostras para a calibração do modelo e 50 para a sua validação.

Primeiramente foi realizada a análise por componentes principais (PCA) e a partir dela foi possível discriminar as amostras de acordo com a fração predominante em cada uma delas.

Após a identificação das amostras prosseguiu-se na tentativa de se quantificar as 3 frações, individualmente e simultaneamente, em cada uma delas através da aplicação dos métodos iPLS, siPLS e PCR. Os modelos obtidos foram, então, avaliados quanto ao  $R^2$ , ao RMSEP (raiz quadrada do erro quadrático médio de previsão) e ao número de variáveis latentes (VL) ou componentes principais (PC) utilizados.

Foram desenvolvidos, também, modelos por iPLS, siPLS e PCR para a determinação da densidade e viscosidade das amostras, os quais foram avaliados da mesma maneira.

Todos os modelos gerados obtiveram resultados bastante satisfatórios. Entretanto comparando-se os modelos siPLS e iPLS, verificou-se que não havia diferença significativa entre eles não justificando, portanto, o uso dos modelos siPLS. Quanto aos modelos gerados por PCR, os mesmos também obtiveram resultados bastante próximos aos fornecidos pelo método iPLS não havendo uma preferência justificável na escolha entre um método e outro.

Com isso demonstra-se, portanto, que é possível discriminar e quantificar de petróleos de origens distintas a partir de suas frações, utilizando espectrometria na região do MIR e técnicas quimiométricas.

# 1 INTRODUÇÃO

Comercialmente, a composição do diesel varia muito, devido a diferentes origens dos petróleos utilizados como matéria prima e diferentes processos de refino. Além disso, sua qualidade tem mudado constantemente desde sua introdução no mercado como combustível.

Para a determinação das propriedades do diesel, normalmente são empregadas análises de laboratório ou analisadores de processo. Contudo, alguns desses métodos são bastante demorados e susceptíveis a erros. Sob este aspecto, a utilização de uma metodologia simples e, principalmente, de menor custo e rápida, pode antecipar a detecção dessas propriedades.

Como técnica que pode agregar essas características, a espectrometria no infravermelho oferece a possibilidade de obtenção de espectros com relativa rapidez, além de fornecer informações do ponto de vista analítico, qualitativo ou quantitativo. Esta técnica tem sido utilizada cada vez mais para fins quantitativos, aumentando o seu uso que, antigamente, era restrito somente à análises qualitativas.

Tem sido possível através de métodos quimiométricos, mais especificamente métodos de calibração multivariada, a determinação de propriedades químicas e físicas de amostras a partir de dados obtidos por infravermelho. As técnicas espectrométricas, quando associadas a ferramentas quimiométricas, são relativamente simples, permitem uma análise rápida, precisa, com razoável exatidão, fazem uso de pequenas amostras e não são destrutivas, razões pelas quais tem se mostrado adequadas para obter diversos tipos de informações em petróleos e derivados.

A quimiometria, associada à espectrometria no infravermelho, é uma ferramenta que cada vez mais torna-se essencial para a química analítica. A análise quimiométrica considera que espectros de infravermelho, assim como qualquer outro tipo de análise espectrométrica, fornece mais informações do que apenas as absorbâncias medidas. Esta ferramenta pode fornecer resultados satisfatórios e interessantes para análises de rotina.

Como método quimiométrico, a calibração multivariada, através do método dos mínimos quadrados parciais (PLS) e regressão por componentes principais (PCR), pode ser empregada para a construção de modelos que permitem a predição de determinados constituintes em amostras, independentemente do estado físico na qual se encontram, com erros mínimos e com rapidez satisfatória. Os modelos construídos podem ser facilmente adaptados às análises de rotinas em laboratórios de controle de qualidade. Assim, modelos de calibração multivariada com infravermelho podem representar boas alternativas para o monitoramento da linha de produção.

Alguns métodos foram descritos na literatura para implementar seleção de região espectral para melhorar significativamente o desempenho dos métodos de calibração de espectros inteiros. Esses métodos são chamados métodos de seleção de variáveis, que selecionam regiões específicas do espectro que não são importantes, enquanto geram modelos mais estáveis, robustos e mais simples de interpretar. Na prática, a metodologia está baseada na identificação de um subconjunto dos dados inteiros que produzirão menores erros de previsão. Assim, em espectros vibracionais, por exemplo, podem ser eliminados os comprimentos de onda que apenas representam ruídos, informações irrelevantes ou não-linearidades.

Na literatura, ainda, podem ser encontrados trabalhos que utilizam a espectrometria na região do infravermelho e as ferramentas quimiométricas, para determinar diversos parâmetros do petróleo e seus produtos, como a determinação da composição, da densidade e a estabilidade das emulsões.

Desta forma, face às suas aplicabilidades, foi proposto neste trabalho a utilização das ferramentas quimiométricas, com seleção de variáveis (iPLS e siPLS), associada à espectrometria no infravermelho com dispositivo ATR, para quantificação de diversos petróleos em diversas misturas, a partir das frações de diesel produzidos pelos mesmos, avaliando também os modelos obtidos, bem como as técnicas utilizadas na seleção de variáveis.

## **2 FUNDAMENTOS TEÓRICOS**

## 2.1 O PETRÓLEO

Petróleo é uma palavra derivada do Latim *petra* e *oleum*, que significa literalmente “óleo de pedra” e refere-se a líquidos ricos em hidrocarbonetos que acumularam-se em reservatórios subterrâneos. O petróleo (igualmente chamado de óleo cru) varia drasticamente nas propriedades da cor, do odor e da densidade e viscosidade que refletem a diversidade de sua origem<sup>1</sup>.

Do ponto de vista químico o petróleo é uma mistura complexa composta em sua maioria de hidrocarbonetos, podendo possuir, em menor parte, compostos de oxigênio, nitrogênio e enxofre, combinados de forma variável, conferindo características diferenciadas aos diversos tipos de óleos crus encontrados na natureza<sup>1, 2, 3</sup>.

Em estado bruto o petróleo não tem aplicabilidade prática, mas quando refinado ele fornece combustíveis líquidos valiosos, solventes, lubrificantes, e muitos outros produtos. Os combustíveis derivados do petróleo contribuem com aproximadamente um terço a um meio do suprimento total de energia no mundo<sup>1</sup>.

O processo físico básico para a separação dos derivados do petróleo em uma refinaria é a destilação. O petróleo contém muitos milhares de compostos diferentes que variam em massa molar de 16 g/mol (metano, CH<sub>4</sub>), a mais de 2000 g/mol<sup>1, 3</sup>. Esta larga escala nas massas molares conduz aos pontos de ebulição que variam de -160° C (ponto de ebulição do metano) a mais do que 600° C, que é o ponto de ebulição de compostos pesados no óleo cru. Um grupo de hidrocarbonetos, portanto, pode ser separado com a destilação de acordo com o ponto de ebulição dos compostos mais leves e mais pesados nas misturas.

De fato, durante a destilação um petróleo bruto é convertido em uma série de frações do petróleo onde cada uma é uma mistura de um número limitado de hidrocarbonetos com uma escala específica do ponto de ebulição. As frações com uma escala mais larga de pontos de ebulição contêm o maior número de hidrocarbonetos. Todas as frações de uma coluna de destilação têm uma escala de ebulição conhecida, exceto o resíduo para o qual o ponto de ebulição superior

geralmente não é conhecido<sup>3</sup>. Porém, a natureza infinitamente variável de fatores composicionais faz com que todos os óleos crus e produtos de petróleo processados numa refinaria sejam diferentes entre si. Essa variabilidade representa uma característica química do tipo *fingerprint* ou impressão digital para cada petróleo e fornece uma base para caracterizá-lo<sup>1, 3</sup>. Na tabela 2.1<sup>3</sup> apresentam-se os principais derivados do petróleo com seus respectivos pontos de ebulição e hidrocarbonetos contidos.

**Tabela 2.1** Frações de petróleo e suas faixas de temperatura

Fração	Hidrocarbonetos Contidos	Faixa de Temperatura (°C)
Gás	C <sub>2</sub> -C <sub>4</sub>	-90 – 1
Gasolina	C <sub>4</sub> -C <sub>10</sub>	-1 – 200
Naftas	C <sub>4</sub> -C <sub>11</sub>	-1 – 205
Combustível de Aviação	C <sub>9</sub> -C <sub>14</sub>	150 – 255
Querosene	C <sub>11</sub> -C <sub>14</sub>	205 – 255
Diesel	C <sub>11</sub> -C <sub>16</sub>	205 – 290
Óleo Combustível Leve	C <sub>14</sub> -C <sub>18</sub>	255 – 315
Óleo Combustível Pesado	C <sub>18</sub> -C <sub>28</sub>	315 – 425
Graxa	C <sub>18</sub> -C <sub>36</sub>	315 – 500
Óleo Lubrificante	>C <sub>25</sub>	> 400
Óleo Combustível de Vácuo	C <sub>28</sub> -C <sub>55</sub>	425 – 600
Resíduo	>C <sub>55</sub>	> 600

A seguir será feita uma breve abordagem a respeito das frações de diesel, tendo em vista que as mesmas foram utilizadas para a realização deste trabalho.

### 2.1.1 Diesel

Os produtos petrolíferos em geral são altamente complexos, e é exigido um esforço considerável para a caracterização de suas propriedades químicas e físicas. Certamente, a análise desses produtos é necessária para se determinar as

propriedades que podem ajudar a resolver um problema do processo assim como as propriedades que indicam a função e o desempenho do produto em questão<sup>1</sup>.

Comercialmente a composição do diesel varia muito devido a diferentes origens do petróleo utilizado como matéria prima e diferentes processos de refino. Além disso, a qualidade do diesel tem mudado constantemente desde sua introdução no mercado como combustível<sup>4</sup>.

Muitos testes de caracterização dos combustíveis diesel dependem da natureza do óleo cru original, dos processos de refino pelos quais o combustível é produzido, e do aditivo (eventualmente) usado. Além disso, a especificação para o combustível diesel pode existir em várias combinações de características.

Assim, como para todos os combustíveis, as propriedades do diesel definem sua habilidade em servir a uma certa finalidade. Uma vez que as propriedades exigidas são determinadas, elas são controladas por testes e por análises apropriadas. Dentre esses testes podemos citar: índice de acidez, aparência e odor, cinzas, resíduo de carbono, calor de combustão, número de cetana, composição, densidade, índice de diesel, estabilidade, viscosidade, volatilidade, água, sedimentos, entre outros<sup>1</sup>.

Os motores diesel são máquinas básicas que geram energia para veículos utilizados principalmente em aplicações que precisem de elevada potência, o que inclui ônibus, grandes caminhões, tratores e máquinas para mineração e dragagem. Atualmente, os veículos diesel vem atraindo uma porção crescente do mercado mundial de veículos de carga leve, incluindo os utilitários, e ainda 20% dos carros para transporte de passageiros, o que inclui as vans, são movidos a diesel.

A popularidade dos veículos a diesel deve-se principalmente à eficiência do diesel como combustível em relação à gasolina, ou mesmo com relação a outros combustíveis simples ou misturados, como o metanol por exemplo, o que chega a conferir uma economia relativa de 25 a 45%. Os motores a diesel apresentam ainda uma excepcional durabilidade. É comum um motor diesel para veículos de carga pesada ter um tempo de vida superior a um milhão de quilômetros, ou seja, cerca de

dez vezes mais que a durabilidade apresentada por um motor à gasolina. Assim, as vantagens oferecidas pelos motores a diesel, tais como durabilidade, segurança e eficiência, justificam sua utilização em vários tipos de máquinas, apesar dos problemas relacionados com a composição de sua emissão e ainda, com os níveis de poluentes sob regulamentação<sup>4</sup>.

## 2.2 ESPETROMETRIA NO INFRAVERMELHO

A região do espectro eletromagnético correspondente ao infravermelho se estende de, aproximadamente, 12800 a 200  $\text{cm}^{-1}$  (0,78  $\mu\text{m}$  a 1000  $\mu\text{m}$ ). Esta região é dividida em: infravermelho distante ou longínquo (FIR, do inglês, *Far IR*), infravermelho médio (MIR, do inglês, *Mid IR*) e infravermelho próximo (NIR, do inglês, *Near IR*). A tabela 2.2 apresenta os limites aproximados para cada região<sup>5</sup>.

**Tabela 2.2** Regiões espectrais do infravermelho.

Região	Intervalo de número de onda ( $\nu$ ), $\text{cm}^{-1}$	Região em comprimento de onda ( $\lambda$ ), $\mu\text{m}$	Região de frequência ( $\nu$ ), Hz
<b>Próximo (NIR)</b>	12.800 a 4.000	0,78 a 2,5	$3,8 \times 10^{14}$ a $1,2 \times 10^{14}$
<b>Médio (MIR)</b>	4.000 a 200	2,5 a 50	$1,2 \times 10^{14}$ a $6,0 \times 10^{12}$
<b>Distante (FIR)</b>	200 a 10	50 a 1000	$6,0 \times 10^{12}$ a $3,0 \times 10^{11}$

Via de regra, as análises feitas na região do infravermelho não geram subprodutos tóxicos, apresentam simplicidade na preparação de amostras além de ser um método não destrutivo.

O infravermelho médio é a região mais usada, tanto para análises qualitativas como quantitativas. Essa região é provavelmente onde se encontra a maioria das pesquisas desenvolvidas e o maior número de aplicações. Ainda hoje, a maioria das aplicações consiste na identificação de compostos orgânicos, pois nessa região ocorrem essencialmente transições fundamentais e existe uma faixa espectral conhecida como região de impressão digital (1.200 a 700  $\text{cm}^{-1}$ ). Nessa região, pequenas diferenças na estrutura e na constituição de uma molécula resultam em

mudanças significativas na distribuição das bandas de absorção. Em consequência, uma semelhança estreita entre dois espectros nesta região, bem como nas outras, constitui forte evidência da identidade dos compostos que produziram os espectros<sup>6</sup>.

Os métodos mais usados para obtenção de espectros no infravermelho envolvem transmissão e reflexão. Em ambos os casos, existem vários acessórios disponíveis comercialmente para a obtenção dos espectros de amostras sólidas, líquidas e gasosas.

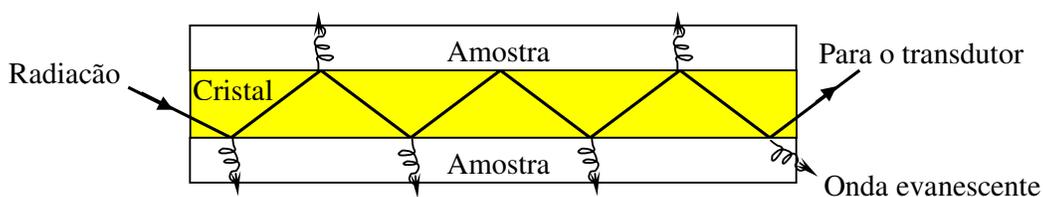
Nos últimos anos, a maioria dos fabricantes de instrumentos oferece adaptadores que cabem nos compartimentos de amostra dos instrumentos de espectrometria no infravermelho e que tornam possível a obtenção imediata de espectros de reflexão<sup>5</sup>.

As técnicas de reflexão vem aumentando sua aplicabilidade particularmente porque elas não envolvem processos morosos de preparo de amostras, sendo úteis tanto para análises qualitativas quanto quantitativas, facilitando o estudo de amostras de difícil preparo para análise por transmissão. Elas permitem a obtenção de espectros da maior parte das substâncias sólidas e líquidas. As técnicas de reflexão mais utilizadas para obtenção de espectros são: reflexão especular, reflexão difusa e reflexão total atenuada<sup>7</sup>. Tendo em vista, que para a realização deste trabalho foi utilizado o acessório de reflexão total atenuada, a seguir serão abordadas características relevantes a esta técnica.

### **2.2.1 Reflexão total atenuada**

A reflexão total atenuada (ATR), também conhecida por espectrometria de reflexão interna, é uma técnica de amostragem rápida que requer mínima preparação da amostra. Além disso, permite a obtenção de espectros de materiais espessos e fortemente absorventes, difíceis de serem analisados por espectrometria por transmissão, possibilitando a obtenção de espectros por reflexão com boa qualidade e tornando possível sua utilização em diversas aplicações analíticas<sup>8</sup>.

A espectrometria de ATR baseia-se no fato de que quando um feixe de radiação passa de um meio mais denso (cristal de ATR) para um menos denso (amostra) ocorre uma reflexão. A fração do feixe incidente que é refletida aumenta com o ângulo de incidência; além de um certo ângulo crítico, a reflexão é total. No processo de reflexão, o feixe se comporta como se penetrasse um pouco no meio menos denso antes que a reflexão ocorra. A profundidade de penetração, que varia de uma fração até vários comprimentos de onda, depende do comprimento de onda da radiação incidente, dos índices de refração dos dois materiais e do ângulo do feixe incidente em relação à interface. A radiação penetrante é chamada onda evanescente. Se o meio menos denso absorve essa onda evanescente, ocorre atenuação do feixe nos comprimentos de onda das bandas de absorção. Esse fenômeno é conhecido como reflectância total atenuada (ATR), mostrado na Figura 2.1<sup>5,9</sup>.



**Figura 2.1** Representação da ATR.

Os espectros de ATR são semelhantes, mas não iguais aos espectros comuns de absorção. Em geral, enquanto as mesmas bandas são observadas, suas intensidades relativas diferem. As absorbâncias, embora dependam do ângulo de incidência, são independentes da espessura da amostra, uma vez que a radiação penetra apenas alguns micrômetros na mesma<sup>5</sup>.

O uso deste tipo de acessório facilita e agiliza a obtenção dos espectros, pois a amostra sofre menor pré-tratamento, sendo, para fins quantitativos, somente pesada ou medida antes da leitura espectrométrica.

Existem inúmeros elementos de ATR e uma variedade de acessórios desenvolvidos para obter espectros de materiais líquidos, sólidos e viscosos para uma larga faixa de tipos de amostras, incluindo filmes, resíduos, papel, revestimentos sobre papel, pós, tintas, tecidos, espumas, minerais, vidros, etc. Alguns possibilitam o

aquecimento da amostra e os mais recentes permitem a análise em fluxo. Na tabela 2.3 são apresentadas as diferentes composições dos elementos de ATR disponíveis, e características a eles relacionadas, sendo determinante para sua escolha a amostra que se deseja analisar. Um dos materiais mais comumente utilizados é o seleneto de zinco<sup>5,7</sup>.

**Tabela 2.3** Principais características das diferentes composições dos elementos de ATR disponíveis

Composição	Dureza kg/mm <sup>2</sup>	Faixa espectro, cm <sup>-1</sup>	Índice de reflexão	Profundidade de penetração 45°, 1000cm <sup>-1</sup> , μ	Faixa de pH
AMTIR	170	11000-630	2,5	1,7	1 a 9
Diamante / KRS-5	5700	30000-250	2,4	2	1 a 14
Diamante / ZnSe	5700	30000-525	2,4	2	1 a 14
Ge	70	5500-570	4	0,66	1 a 14
Si / ZnSe	1150	8900-550	3,4	0,85	1 a 12
Si	1150	8900-1500 475-40	3,4	0,85	1 a 12
ZnSe	120	15000-520	2,4	2	5 a 9

## 2.3 ANÁLISE MULTIVARIADA: QUIMIOMETRIA

A moderna instrumentação de análises química é capaz de gerar uma quantidade considerável de dados, sobre uma única amostra, em um curto espaço de tempo. Um espectrômetro pode registrar sinais provenientes de mais de mil comprimentos de onda ou, um único cromatograma pode apresentar mais de cem picos. Assim, para que uma informação útil seja obtida deste grande volume de dados é necessária a utilização de técnicas matemáticas adequadas, sendo a quimiometria um dos campos de estudo da química que fornece tais ferramentas<sup>5,10</sup>.

A quimiometria pode ser definida como uma área da química na qual métodos matemáticos e estatísticos são aplicados a dados de origens distintas para a obtenção de uma informação química desejada. Consiste de um conjunto de técnicas de cálculo com o objetivo de promover a obtenção de informação útil de um conjunto complexo de dados, englobando conceitos de planejamento experimental, pré-processamento de dados e análise estatística multivariada<sup>11, 12</sup>. A quimiometria

é, portanto, uma ferramenta de automação laboratorial e está relacionada com a análise multivariada. De um modo geral, a análise multivariada refere-se aos métodos estatísticos e matemáticos que analisam, simultaneamente, múltiplas medidas de um objeto sob investigação, seja ele de caráter químico ou não<sup>13</sup>.

Até meados da década de 1970 os métodos de análise eram construídos basicamente pela regressão de uma propriedade química qualquer, mais usualmente a concentração de um analito, sobre uma única variável, geralmente um único canal de instrumento analítico.

Os métodos de análise univariada exigiam, e ainda exigem, considerável seletividade instrumental porque o único canal selecionado deve apresentar sinal exclusivamente dependente da propriedade de interesse, sendo livre de interferência devido a outras substâncias. Deste modo, a aplicação destas técnicas requer o desprendimento de quantidades consideráveis de recursos na determinação de procedimentos de tratamento de amostras para isolamento do analito de interesse e/ou na construção de instrumentos cada vez mais seletivos.

Com o surgimento dos métodos de análise multivariada o cenário em análise química começou a mudar. Ao invés de desenvolver métodos analíticos seletivos, passou-se a investir recursos no desenvolvimento de técnicas matemáticas capazes de transformar sinais oriundos de misturas de componentes, em informação útil sobre uma ou várias propriedades de interesse. Estes métodos ganharam força com o desenvolvimento de computadores pessoais equipados com maiores recursos e mais acessíveis aos pesquisadores e profissionais da química.

Outra característica dos métodos multivariados na determinação de uma propriedade de interesse é a chamada *vantagem multicanal*. Quando apenas um canal ou variável é utilizado, a calibração torna-se extremamente sensível a ruídos nesta variável. A utilização de diversos canais minimiza ou mesmo elimina esta interferência, tornando a calibração mais robusta<sup>10, 24</sup>.

Existem diversos tipos de técnicas de estatística multivariada, com as mais variadas aplicações. Tais métodos podem ser classificados em dois tipos principais: os métodos de *análise exploratória* e os métodos de *calibração multivariada*.

A aplicação de um ou outro método, ou até mesmo da combinação dos dois depende da natureza do problema que se deseja resolver, ou do tipo de informação que se deseja obter<sup>14</sup>.

Ao se deparar com um conjunto de dados, principalmente aquele muito extenso, é necessária uma análise prévia inicial para se avaliar a qualidade das informações disponíveis. Os métodos de análise exploratória têm por objetivo fazer a avaliação inicial dos dados para descobrir que tipo de informação pode-se extrair deles, e assim definir as diretrizes para um tratamento mais aprofundado. Isto é feito através da utilização de algoritmos que permitem reduzir a dimensão dos dados, ou organizá-los numa estrutura que facilite a visualização de todo o conjunto, de forma global. Desta forma é possível descrever e identificar grupos de amostras dentro dos dados, agrupando-os de modo a permitir a identificação das semelhanças. Entre os métodos de análise exploratória, podem ser citados a análise por componentes principais (PCA) e análise por agrupamentos hierárquicos (HCA)<sup>15</sup>.

Por outro lado, os métodos de calibração multivariada, relacionam-se ao desenvolvimento de modelos matemáticos que permitem estimar alguma propriedade de interesse com os sinais instrumentais. Entre os métodos de calibração, podem ser citados a regressão por componentes principais (PCR- *Principal Components Regression*), o método de mínimos quadrados parciais (PLS- *Partial Least Square*) e o da regressão linear múltipla (MLR – *Multilinear Regression*)<sup>16</sup>.

Tendo em vista que os métodos de análise por componentes principais (PCA), regressão por componentes principais (PCR) e mínimos quadrados parciais (PLS) foram utilizados no presente trabalho, estes serão abordados de forma preferencial.

### **2.3.1 Análise por componentes principais (PCA)**

A análise por componentes principais (PCA) tem como principal aplicação a redução e a interpretação de dados, sendo frequentemente utilizada como passo intermediário para maiores investigações<sup>17</sup>.

A PCA pode ser resumida como sendo uma transformação de um espaço dimensional com  $m$  variáveis (medidas), para um espaço com  $i$  novas variáveis, podendo  $i$  ser igual ou menor que  $m$ , de forma que o novo conjunto, forme um conjunto de vetores ortonormais. Como resultado desta transformação as novas variáveis obtidas, chamadas de componentes principais (PCs), são combinações lineares das variáveis medidas originais e são ortogonais entre si, ou seja, são completamente não correlacionadas. As PCs estão ordenadas em seqüência que vai daquela com maior explicação de variação dos dados (primeira componentes principal) para aquela com menor explicação da variação dos dados (última componentes principal). Isto cria a possibilidade de decomposição da matriz de dados em *estrutura* e *ruído*, ou em outras palavras, em variáveis significantes e não significantes, para a explicação da variação dos dados<sup>18, 19</sup>.

As PCs, por serem combinações lineares das variáveis originais, superam os problemas de seletividade (não eliminando nenhuma variável) e colinearidade (muitas variáveis contém tipos de informações similares)<sup>20</sup>.

De forma geral e bem simplificada, as combinações lineares das  $m$ -variáveis originais que geram cada componente principal podem ser representadas pela equação 2.1<sup>19, 21</sup>.

$$PC_i = x_1 p_{1i} + x_2 p_{2i} + \dots + x_j p_{ji} \quad (2.1)$$

Nesta equação,  $x_j$  (para  $j = 1, 2, \dots, m$ ) são as variáveis originais e  $p_{ji}$  (para  $j = 1, 2, \dots, m$ ) são os coeficientes que medem a importância de cada variável na  $i$ -ésima componente principal ( $PC_i$ ), ou seja, o peso que cada variável tem naquela combinação linear. Estes pesos ou *loadings*, nada mais são do que o cosseno do ângulo entre o eixo da componente principal e o eixo da variável original e seu valor,

portanto, estará sempre entre +1 e -1. Quanto mais próximo de  $\pm 1$ , maior a influência que esta determinada variável tem na descrição desta componente principal, e quanto mais próximo de zero este coeficiente estiver, menor a influência da variável naquela componente principal<sup>16, 19, 32</sup>. Se o conjunto de dados contiver três variáveis, este irá possuir três PC's<sup>16, 22, 32</sup>. Cabe salientar que uma variável pode possuir um peso considerável para uma dada PC, mas desprezível para outra<sup>17, 19</sup>.

A projeção de cada amostra, de um universo de n amostras, neste novo sistema de eixos fornece os *scores*, e cada amostra terá então um valor de *score* para cada um dos novos eixos (as PCs). Por exemplo, se um conjunto de dados contém dez objetos, o mesmo número de *scores* é obtido para cada componente principal<sup>19, 32</sup>.

O gráfico dos *scores* poderá revelar agrupamentos ou tendências das amostras analisadas, que poderiam ser de difícil visualização no caso das variáveis originais.

Como dito anteriormente, os pesos são capazes de mostrar quais variáveis originais tem maior importância na combinação linear de cada componente principal. Através disto pode-se saber, por exemplo, que variáveis contribuem para a descrição deste ou daquele conjunto de amostras<sup>17, 21</sup>.

Normalmente, as últimas PCs modelam ruído inerente aos dados e sendo assim, a eliminação dessas PCs freqüentemente aumenta a relação sinal/ruído. Um procedimento básico para melhorar a análise de modo a reduzir a dimensão dos dados é a escolha do número de componentes principais a serem utilizadas na descrição do sistema, mantendo apenas aquelas mais significativas, ou seja, aquelas que carregam maior variância dos dados<sup>30, 32</sup>. Este assunto será melhor discutido nos próximos tópicos.

A análise de componentes principais também é largamente utilizada como método de análise exploratória de dados. Gráficos de *loadings* e *scores* revelam padrões característicos do comportamento de amostras em função de um conjunto de variáveis que dificilmente seriam reconhecidos pela observação de valores

tabelados<sup>24</sup>. Neste trabalho, uma análise exploratória utilizando gráficos de PCA será apresentada.

### 2.3.1.1 Conceitos Matemáticos

Em termos matemáticos, a PCA corresponde à decomposição de uma matriz de dados  $\mathbf{X}$ , de dimensão  $n \times m$ , no produto de duas matrizes: a matriz dos *scores*  $\mathbf{T}$  e a transposta da matriz dos pesos  $\mathbf{P}^T$ , como mostra a equação 2.2, que pode ser representada pelo esquema da figura 2.2<sup>19, 32</sup>.

$$X = TP^T \quad (2.2)$$

The diagram shows three rectangular boxes representing matrices. The first box on the left is labeled 'X' and has 'n' written at the bottom-left corner and 'm' written above its top-right corner. To its right is an equals sign. The second box is labeled 'T' and has 'n' written at the bottom-left corner and 'e' written above its top-right corner. To its right is a multiplication sign 'x'. The third box is labeled 'P<sup>T</sup>' and has 'q' written at the bottom-left corner and 'm' written above its top-right corner.

**Figura 2.2** Representação da decomposição da matriz de dados  $\mathbf{X}$ .

Os vetores linha da matriz dos *scores*  $\mathbf{T}$ , correspondem às projeções das  $n$  amostras, ou objetos, da matriz original  $\mathbf{X}$ , nos novos eixos formados, ou seja, cada vetor coluna da matriz dos *scores* corresponde às coordenadas das amostras em cada *componente principal* gerada.

Desta forma, o número  $n$  de linhas da matriz original é igual ao número de linhas da matriz dos *scores* e o número de colunas corresponde ao número  $e$  de componentes geradas.

Na matriz dos pesos  $\mathbf{P}$ , cada vetor coluna corresponde aos pesos que cada variável possui na combinação linear das  $m$  componentes geradas. Então esta matriz possui

$q$  linhas (o número de componentes principais armazenadas) e  $m$  colunas (as variáveis originais).

A cada componente modelada, certa quantidade de informação permanece sem ser explicada, ou seja, uma quantidade de variância continua não descrita. A esta quantidade de variância não descrita chamamos de *resíduos* e podem ser organizados na forma da matriz dos resíduos **E**. Podemos, então, reescrever a equação 2.2 como descrito na equação 2.3<sup>19, 22</sup>.

$$X = TP^T + E \quad (2.3)$$

Depois que a  $PC_1$  é calculada, a próxima PC é obtida com a matriz residual E, que contém a variância não explicada pela  $PC_1$ , e assim sucessivamente, conforme a equação 2.4<sup>19, 22</sup>:

$$X = t_1p_1 + t_2p_2 + \dots + t_h p_h + E \quad (2.4)$$

### 2.3.2 Métodos de calibração multivariada

Os métodos de calibração multivariada permitem o tratamento de dados complexos do ponto de vista matemático e estatístico, correlacionando medidas instrumentais e valores para uma propriedade de interesse correspondente<sup>16, 23</sup>.

O desenvolvimento do método de calibração consiste de duas etapas: a calibração e a validação. A etapa de construção do modelo de calibração começa com a seleção de um conjunto de amostras cuidadosamente escolhidas para que sejam representativas de toda a região a ser modelada. Estas amostras (conjunto de calibração) serão utilizadas na construção de um modelo apropriado para relacionar as respostas instrumentais com a informação desejada. Durante essa etapa dois fatores são considerados cruciais: o número de componentes principais ou de variáveis latentes, e a detecção de amostras anômalas (*outliers*). Após, deve ser realizada a etapa de validação, verificando a capacidade preditiva do modelo. A validação consiste em testar o modelo com amostras externas, das quais se tem

conhecimento prévio das propriedades (ou concentrações) que se desejam medir<sup>15, 16</sup>.

Um modelo de calibração, na verdade, é uma função matemática (**f**), que relaciona dois grupos de variáveis, uma delas denominada dependente (**Y**) e a outra denominada independente (**X**)<sup>14</sup>:

$$Y = f(X) = X * b \quad (2.5)$$

,onde **b** corresponde à matriz dos coeficientes de regressão do modelo, e são determinados matematicamente a partir de dados experimentais<sup>32</sup>.

#### 2.3.2.1 Regressão por componentes principais – PCR

A Regressão por Componentes Principais, ou Principal Component Regression - PCR, utiliza a regressão para converter os *scores* resultantes da Análise dos Componentes Principais (*Principal Components Analysis* - PCA) em concentrações<sup>25, 32</sup>.

Como as componentes principais estão ordenadas em seqüência que vai daquela com maior explicação para aquela com menor explicação da variação dos dados, para a construção dos modelos por PCR, utiliza-se apenas as componentes de maior importância, ou melhor, as componentes que estão associadas com uma maior variância dos dados. Isso ocorre, pois as componentes que possuem pequena variância, tendem a possuir piores relações sinal / ruído. Remover estas direções é desejável para diminuir a sensibilidade a ruídos do modelo de predição<sup>25</sup>.

Pela compressão por PCA, o modelo matemático da regressão linear (PCR) pode ser expresso como:

$$y = T^A b + E \quad (2.6)$$

em que  $T^A$  indica que a matriz de *scores* original foi truncada nas  $A$  primeiras componentes principais. Alguns autores sugerem que os *loadings* também sejam considerados no modelo, obtendo-se assim um vetor de regressão  $\mathbf{b}'$  diretamente relacionado com as variáveis originais<sup>24</sup>.

Um aspecto característico do modelo PCR é a construção das componentes principais utilizando unicamente as respostas instrumentais ( $\mathbf{X}$ ) sem levar em consideração informações provenientes das concentrações ( $\mathbf{Y}$ ). Isto pode se constituir numa fragilidade do método no caso em que o analito de interesse tem um sinal muito fraco e, portanto, não influencia fortemente nas primeiras componentes principais, fazendo com que um número maior delas seja necessário para a construção do modelo<sup>10</sup>.

#### 2.3.2.2 Mínimos quadrados parciais – PLS

O método dos mínimos quadrados parciais – PLS (do inglês “Partial Least Squares”) ou a regressão por PLS (PLSR) é o método de regressão mais utilizado para a construção de modelos de calibração multivariada a partir dos dados de primeira ordem. Este método, assim como a PCR, não requer um conhecimento exato de todos os componentes presentes nas amostras, podendo realizar a previsão de amostras mesmo na presença de interferentes, desde que estes também estejam presentes por ocasião da construção do modelo<sup>21, 32, 34</sup>.

A PLS contorna a dificuldade característica da PCR descrita anteriormente usando a informação das concentrações na obtenção dos fatores, o que só é justificável se tais concentrações tiverem valores confiáveis. A primeira componente, neste caso chamado de variável latente (LV-Latent Variable), descreve a direção de máxima variância que também se correlaciona com a concentração. Estas variáveis latentes são na realidade combinações lineares das componentes principais calculadas pelo método PCR.

A preferência entre um dentre esses dois métodos não pode ser aconselhada de uma forma genérica uma vez que ambos são em geral igualmente eficientes e as pequenas variações dependem de caso para caso<sup>10</sup>.

A regressão por PLS é uma extensão da PCA, pois a construção de um modelo PLS ocorre de maneira similar à PCA. Porém, além da matriz de dados independentes  $X$  ser projetada num novo sistema de coordenadas, também a matriz dependente  $Y$  é decomposta da mesma maneira, simultaneamente. Portanto, a PLS utiliza as respostas analíticas, bem como as informações de interesse, para capturar a variância dos dados da matriz  $X$  e da matriz dependente  $Y$ , através de suas decomposições sucessivas e simultâneas, correlacionando-as<sup>16, 26</sup>.

O modelo PLS é obtido através de um processo iterativo, no qual se otimiza ao mesmo tempo a projeção das amostras sobre o(s) peso(s), para a determinação dos *scores*, e o ajuste por uma função linear dos *scores* da matriz  $X$  aos *scores* da matriz  $Y$  de modo a minimizar os desvios. Essa otimização simultânea ocasiona pequenas distorções nas direções dos pesos de modo que, rigorosamente eles perdem a ortogonalidade, levando a pequenas redundâncias de informação. Porém são essas pequenas redundâncias que otimizam a relação linear entre os *scores*, e estas distorções da ortogonalidade entre os componentes principais na PLS fazem com que não sejam mais componentes principais (que são ortogonais) e sim variáveis latentes (VLs)<sup>21, 27, 28</sup>.

Portanto, a construção de um modelo PLS consiste de uma regressão entre os *scores* das matrizes  $X$  e  $Y$  em uma soma de “h” variáveis latentes. O modelo PLS pode ser definido através de relações externas, que correlacionam, individualmente, as matrizes  $X$  e  $Y$  (conforme as equações 2.7 e 2.8), enquanto as internas correlacionam ambas as matrizes<sup>26</sup>.

$$X = TP^T + E_X = \sum t_h p_h^T + E_X \quad (2.7)$$

$$Y = UQ^T + E_Y = \sum u_h q_h^T + E_Y \quad (2.8)$$

onde  $X$  é a matriz de dados (medida instrumental),  $Y$  é a matriz de resposta (concentração, por exemplo),  $T$  e  $U$  são os scores de  $X$  e  $Y$  respectivamente; e os elementos de  $P$  e  $Q$  são os “pesos”. As matrizes  $E_x$  e  $E_y$  correspondem aos resíduos. A relação interna dos *scores* das matrizes  $X$  e  $Y$  é obtida através do coeficiente de regressão linear, como descrito na equação 2.9:

$$u_h = b_h t_h \quad (2.9)$$

em que,  $b_h$  é o vetor de coeficientes de regressão do modelo linear para cada VL, obtido através de:

$$b_h = \frac{u_h^T t_h}{t_h^T t_h} \quad (2.10)$$

sendo  $u$  e  $t$  os elementos das matrizes  $U$  e  $T$ , respectivamente<sup>14, 30</sup>.

os valores de  $b_h$  são agrupados na matriz diagonal  $B$  (matriz identidade), que contém os coeficientes de regressão entre a matriz de *scores*  $U$  de  $Y$  e a matriz de *scores*  $T$  de  $X$ . A melhor relação linear possível entre os *scores* das matrizes é obtida através de pequenas rotações das variáveis latentes das matrizes  $X$  e  $Y$ . A matriz  $Y$  pode ser calculada através das informações contidas em  $u_h$  (equação 2.9), conforme a equação 2.11:

$$Y = TBQ' + F \quad (2.11)$$

Onde  $T$  são os *scores* da matriz  $X$ ,  $B$  é a matriz identidade de  $b_h$ ,  $Q'$  são os *loadings* da matriz  $Y$  e  $F$  é a matriz residual de  $Y$ . Desta forma, a concentração das novas amostras é prevista a partir dos novos *scores* de  $X$ , dado por  $T^*$ , substituídos na equação 2.12:

$$Y = T^*BQ' \quad (2.12)$$

O modelo final consiste das matrizes dos “scores” **T** e **U** que são linearmente relacionadas por um coeficiente **B**. No final do processo, a variância explicada pela primeira VL será maior que a variância explicada pela segunda VL e a terceira VL explicará uma variância menor que a segunda VL, e assim sucessivamente até o número de VLs definido <sup>21, 26</sup>.

Os métodos PCR e PLS são consideravelmente mais eficientes do que os métodos tradicionais de calibração (MLR, por exemplo) para lidar com ruídos experimentais, colinearidade e não linearidades. Todas as variáveis relevantes são incluídas nos modelos via PCR ou PLS, o que implica que a calibração pode ser realizada eficientemente mesmo na presença de interferentes, não havendo necessidade do conhecimento do número e natureza dos mesmos. Os métodos PCR e PLS são robustos, isto é, seus parâmetros praticamente não se alteram com a inclusão de novas amostras no conjunto de calibração<sup>10</sup>.

### 2.3.2.3 Validação dos modelos de regressão

Antes da aplicação do modelo construído, o mesmo deve ser validado com o objetivo de se testar a sua capacidade preditiva; sem esta etapa não há sentido em prosseguir. A validação consiste em testar o modelo prevendo concentrações de amostras, para estabelecer se ele de fato irá refletir o comportamento do analito de interesse <sup>10</sup>.

Durante essa etapa de validação algumas perguntas devem ser respondidas, tais como:

- 1 – Quais amostras do conjunto serão utilizadas na construção do modelo?
- 2 – Quantas componentes principais são necessárias para se caracterizar uma série de dados?
- 3 – Uma amostra desconhecida possui uma boa previsão pelo modelo?

Existe uma vasta literatura descrevendo métodos de como desenvolver um modelo, sua validação, o projeto experimental a ser seguido e como medir os erros. A maioria desses métodos guiam o leitor a respeito de quantas componentes significativas se deve reter <sup>32</sup>.

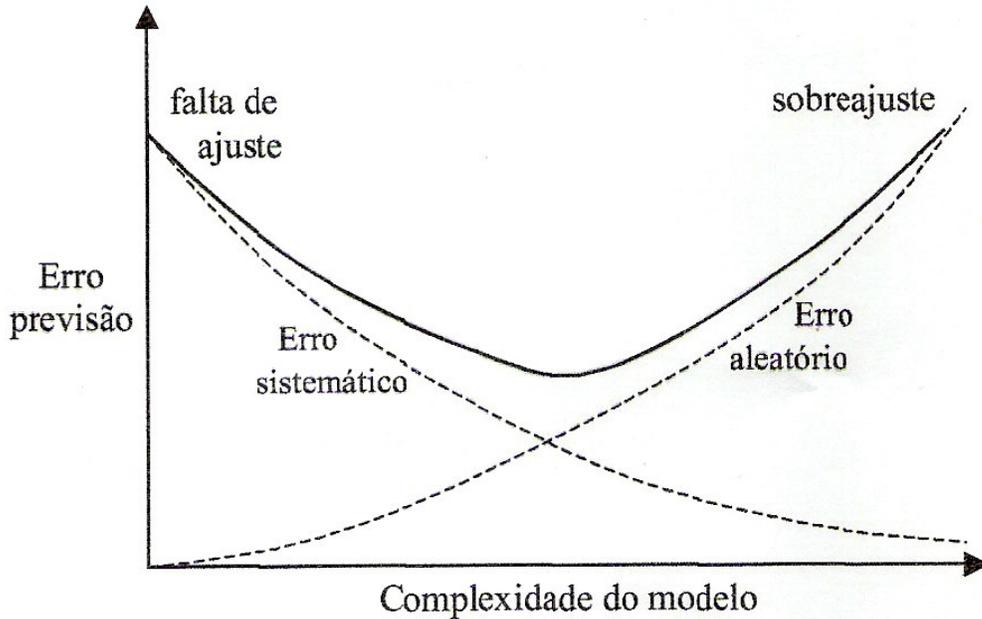
É possível se obter uma medição dos dados cada vez mais próxima da desejada utilizando mais e mais componentes principais, até que os dados sejam medidos exatamente; entretanto, as últimas componentes geralmente são pouco significativas, pois modelam ruídos inerentes aos dados, sendo assim, a eliminação dessas PCs freqüentemente aumenta a relação sinal/ruído. Um procedimento básico para melhorar a análise de modo a reduzir a dimensão dos dados é a escolha do número de componentes principais a serem utilizadas na descrição do sistema. A escolha do número de componentes principais permite descrever até mais de 90% da variância dos dados com um número muito menor de fatores do que das variáveis originais. A escolha das componentes principais a serem utilizadas na descrição dos dados é feita levando-se em conta a porcentagem da variância descrita pelas PCs e a variância residual <sup>18, 30</sup>.

A seguir serão abordados aspectos importantes referentes à seleção do número de componentes a se utilizar na construção de um modelo e à validação dele. Os métodos mencionados abaixo tomam como referência a PLS, porém os princípios se aplicam a todos os métodos de calibração.

#### *2.3.2.3.1 Escolha do número de variáveis latentes (VLs)*

O número de variáveis latentes utilizadas em um modelo é de fundamental importância nos resultados a serem obtidos. Um modelo contendo um número de VLs inferior ao ideal resultará em subajuste, pois não irá considerar na totalidade a informação contida no conjunto de dados. Por outro lado, quanto maior o número de VLs, também aumentará o ruído e os erros de modelagem, resultando em um sobreajuste. Deste modo, o número ótimo de VLs que corresponda ao ponto no qual a diminuição do erro (produzido pelo aumento da complexidade do modelo) é compensado pelo aumento de erro de superavaliação. Na figura 2.2, este ponto é

representado pelo ponto mínimo na curva, e para o qual um erro mínimo de previsão é produzido <sup>7</sup>.



**Figura 2.3** Ilustração teórica da validação de um modelo

### 2.3.2.3.2 Cálculo dos erros

A eficiência dos modelos de calibração multivariada pode ser avaliada pelo cálculo dos valores da raiz quadrada do erro quadrático médio (RMSE – do inglês, *root mean square error*). Esses valores são igualmente utilizados para a definição do número de VLs a ser utilizado no modelo.

Tais valores expressam a exatidão do modelo, ou seja, a proximidade entre o valor calculado pelo modelo ( $y_{prev}$ ) e o valor verdadeiro ou obtido por um método de referência ( $y_{real}$ ). Os erros são definidos como:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_{prev} - y_{real})^2}{n}} \quad (2.13)$$

sendo  $n$  o número de amostras.

Esses erros podem ser calculados a partir da auto-predição das amostras de calibração (raiz quadrada do erro quadrático médio de calibração - RMSEC), da validação cruzada das mesmas (raiz quadrada do erro quadrático médio de validação cruzada - RMSECV) e da predição de amostras externas (raiz quadrada do erro quadrático médio de previsão - RMSEP).

#### *2.3.2.3.2.1 Auto-predição*

O método mais simples de se determinar o número de variáveis latentes é o da auto-predição, na qual o erro (RMSEC) é determinado através da previsão, de uma só vez, do conjunto de amostras utilizadas na construção do modelo. Este método possui a desvantagem de que o valor de RMSEC depende do número de variáveis latentes utilizadas e, por isso, o poder de previsão do modelo também fica condicionado ao maior número de variáveis latentes utilizadas <sup>32</sup>.

#### *2.3.2.3.2.2 Validação Cruzada*

Para a determinação do número correto de variáveis latentes (LV) e validação do modelo, também pode ser utilizado o método de Validação Cruzada (CV – do inglês, *Cross Validation*). Tal método é baseado na habilidade de previsão de um modelo construído por parte de um conjunto de dados, seguido pela previsão do restante do conjunto de dados, que é realizada pelo modelo construído <sup>10, 14, 32</sup>.

A validação cruzada pode ser realizada em blocos, ou seja, um número determinado de amostras é deixado de fora no processo de construção do modelo e a seguir essas amostras são previstas pelo modelo construído, ou ainda por um caso conhecido como “*leave-one-out*” (deixe uma fora), em que uma amostra é deixada de fora no processo de construção do modelo e a seguir essa amostra é prevista pelo modelo construído. Em ambos os casos, o processo é repetido até que todas as amostras tenham sido previstas e a raiz quadrada do erro quadrático médio da validação cruzada (RMSECV) é calculada através da equação 2.13 (página 24) <sup>10, 32</sup>.

O cálculo é realizado para o número de componentes de 1 até A, e os resultados de RMSECV são apresentados em um gráfico em função do número de LV. O comportamento típico para esses gráficos é a observação de um mínimo ou um patamar, que indica a melhor dimensionalidade do modelo de regressão, ou seja, o melhor número de VL que produziu o menor erro de previsão sem perda significativa da variância dos dados <sup>14</sup>.

Uma fraqueza significativa do método de validação cruzada é que ele depende da dimensão dos dados gerados pelas amostras originais utilizadas na construção do modelo. Estas amostras são frequentemente chamadas de amostras de treinamento.

Considere-se uma situação em que dois compostos quaisquer, A e B, estão presentes em diferentes quantidades em diversas misturas dos mesmos e deseja-se, através da aplicação de métodos quimiométricos associados à espectrometria no infravermelho, quantificá-los. Porém, as concentrações de ambos estão correlacionadas e são diretamente proporcionais, de modo que quando há uma elevada concentração do composto A, também há uma elevada concentração do composto B e vice-versa. Para isso uma série de espectros da mistura é gerada e, posteriormente, um modelo da calibração pode ser obtido desses dados analíticos, com boa habilidade na previsão de ambas as concentrações. O método de validação cruzada poderia sugerir que o modelo fosse bom, com valores baixos de RMSECV. Entretanto, se fosse necessário prever a concentração dos compostos A e B em misturas nas quais eles se encontram em concentrações inversamente proporcionais, os resultados não seriam muito satisfatórios visto que o modelo criado não foi “treinado” para lidar com esta nova situação.

A validação cruzada é muito útil para remover a influência de fatores internos tais como o ruído instrumental a partir de erros da diluição, mas não pode ajudar muito nos casos em que há correlações na concentração das amostras de treinamento e inevitavelmente como discutido no exemplo anterior <sup>32</sup>.

### 2.3.2.3.2.3 Previsão de amostras externas

Ao invés de se validar as previsões internamente, é possível testá-las com amostras externas, frequentemente chamadas de “amostras de validação”.

Utilizando o exemplo dos dados produzidos por amostras submetidas à análise na região do infravermelho, a etapa de previsão do modelo desenvolvido é feita da seguinte forma: forma-se uma nova matriz com os valores das intensidades em cada número de onda dispostos em coluna, e cada linha desta matriz representando uma amostra (matriz  $\mathbf{X}_{teste}$ )<sup>10, 14</sup>.

Formada tal matriz deve-se aplicar a este conjunto os mesmos pré-tratamentos matemáticos aplicados ao conjunto de calibração.

O passo seguinte consiste em obter os vetores de *scores* ( $\mathbf{t}$ ) e os resíduos ( $\mathbf{e}$ ) da matriz  $\mathbf{X}$  teste utilizando os pesos ( $\mathbf{p}$ ), calculados na fase de treinamento.

$$\mathbf{X}_{teste} = \sum_{a=1}^A t_a \mathbf{p}_a^T + \mathbf{e} \quad (2.14)$$

para “A” variáveis latentes na modelagem

Com os valores dos *scores* ( $\mathbf{t}$ ) da matriz  $\mathbf{X}$  teste, além dos coeficientes de regressão  $\mathbf{b}$  e os pesos  $\mathbf{q}$  do bloco  $\mathbf{Y}$  calculados na fase de calibração, pode-se fazer a estimativa das concentrações da seguinte forma:

$$\mathbf{Y} = \sum_{a=1}^A t_a \mathbf{b}_a \mathbf{q}_a^T \quad (2.15)$$

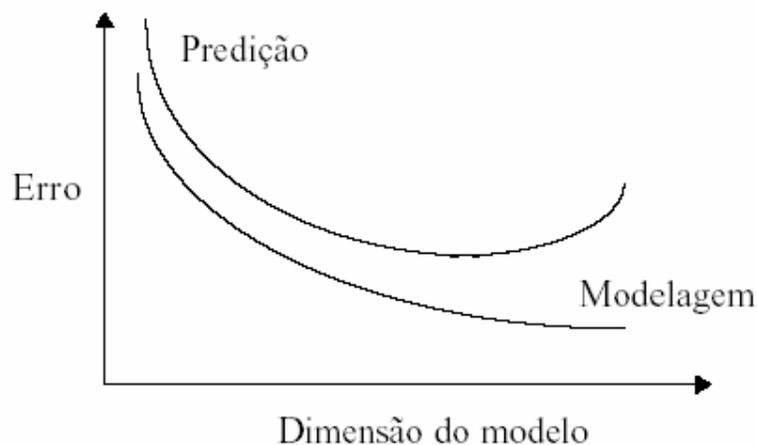
para “A” variáveis latentes<sup>14</sup>.

Após ter sido realizada a estimativa das concentrações, é calculada a raiz quadrada do erro quadrático médio de previsão (RMSEP) através da equação 2.13 (página 24) da mesma forma como para a validação cruzada, porém o modelo de calibração é calculado apenas uma vez.

Normalmente o erro mínimo calculado para as amostras de validação é maior do que o da validação cruzada, porém se a estrutura das amostras de validação abrange as de treinamento, esses dois erros serão bastante similares <sup>32</sup>.

#### 2.3.2.3.2.4 O problema do overfitting

Um problema que pode ocorrer quando são utilizados dados de processo na confecção de modelos é o *overfitting*. Normalmente, o conjunto de dados é subdividido em dois subconjuntos, sendo um para a obtenção do modelo (referência) e outro para a avaliação, ou teste, do modelo obtido. Aumentando-se a quantidade de variáveis de entrada do modelo, o erro da modelagem diminui, porém o desempenho do mesmo, avaliado através do erro de predição, pode piorar. Resultados típicos de *overfitting* em um conjunto de dados de um modelo de grande dimensão (muitas entradas) é mostrado na figura a seguir <sup>25</sup>.



**Figura 2.4** Comparação entre os erros de modelagem e de predição<sup>25</sup>

### 2.3.2.3.3 Detecção de amostras anômalas – outliers

Outro critério, tão importante quanto a determinação de VLS para avaliar um modelo de regressão, é a detecção de amostras anômalas (do inglês, *outliers*). Ao verificar a qualidade do conjunto de calibração, deve-se assegurar de que as amostras formam um conjunto homogêneo, removendo-se aquelas amostras que são anômalas.

Anomalias são elementos muito diferentes ou que apresentam erros grosseiros quando comparados à maioria dos dados portanto, é necessária a identificação e eliminação destes elementos já no processo de calibração, pois caso contrário pode-se obter um modelo não representativo. Estas amostras anômalas também podem ser encontradas nos dados utilizados para a previsão do modelo.

A ocorrência de anomalias pode ser devida a várias razões: erros instrumentais, experimentais, presença de compostos químicos não pertencentes ao grupo que está sendo analisado ou de diferentes composições químicas, além de outras fontes. Assim, tem-se que as anomalias podem ser encontradas nas amostras, nas variáveis e na relação entre amostra e variáveis.

As formas mais importantes de anomalias correspondem àquelas provenientes das amostras. Uma amostra difere da outra por ter uma composição diferente ou por algum erro experimental, instrumental, etc. Uma variável anômala é proveniente de dados de alguma amostra que diferem do conjunto total de calibração por causa de ruídos, erros na medição e/ou métodos inadequados para análise de determinada amostra ou porque esta reflete alguma propriedade única.

Nas últimas décadas, métodos estatísticos robustos têm sido desenvolvidos a partir da identificação e remoção automática das anomalias. No entanto, tal procedimento deve ser tomado com alguma cautela, pois algumas vezes a presença de uma amostra diferente das demais pode conter, ao invés de erros, informações importantes que não são encontradas nos outros dados e, dessa forma, sua presença poderá contribuir muito para o desenvolvimento do modelo.

Amostras com alto ou baixo nível de analito são, geralmente, amostras muito diferentes do restante do conjunto, porém, em casos onde o modelo de calibração é linear, amostras deste tipo devem ser mantidas no modelo, pois correspondem a amostras informativas. Outros tipos de anomalias que não podem ser explicadas devem ser eliminadas. A identificação do problema pode estar na análise experimental ou instrumental, sendo necessário repetir os procedimentos.

Um método de identificação de *outliers* pode ser feito através da análise de gráficos gerados através da análise exploratória dos dados.

#### 2.3.2.4 Método de seleção de variáveis

A construção de modelos por calibração multivariada pode ser realizada utilizando a informação de toda faixa espectral de trabalho para construir um modelo de regressão correlacionando com a propriedade de interesse. No entanto, considerando o grande número de variáveis fornecidas por toda a faixa espectral, algumas destas variáveis podem interferir na modelagem, além de tornar o tratamento dos dados mais lento. Portanto, para melhorar o desempenho de técnicas de calibração multivariada, têm sido utilizados procedimentos apropriados para a seleção das regiões espectrais associadas<sup>29, 34</sup>.

Alguns métodos têm sido descritos recentemente na literatura para implementar seleção de região espectral para melhorar significativamente o desempenho dos métodos de calibração de espectros totais. Esses métodos são chamados métodos de seleção de variáveis, que escolhem regiões específicas do espectro (um comprimento de onda ou um conjunto de comprimentos de onda) em que a colinearidade não é tão importante, enquanto gera modelos mais estáveis robustos e mais simples de interpretar. Na prática, a filosofia está baseada na identificação de um subconjunto dos dados inteiros que produzirão erros de previsão mais baixos<sup>30,31, 34</sup>.

Existem vários critérios para a escolha da região espectral. Normalmente, a escolha depende da experiência do analista (das regiões habitualmente excluídas) e do seu

conhecimento do sistema sob investigação, ou das correlações existentes entre as regiões escolhidas e as propriedades de interesse.

Os métodos de seleção de variáveis existentes se diferem com relação ao procedimento realizado para a seleção da região espectral. Dentre os métodos utilizados atualmente pode-se destacar o método de mínimos quadrados parciais por intervalos (iPLS – do inglês, *Interval Partial Least Square*)<sup>29,30,31</sup>, o método de eliminação de variáveis não informativas por mínimos quadrados parciais (UVE-PLS – do inglês, *Elimination of Uninformative Variables in Partial Least Square*)<sup>21,30</sup> e algoritmo genético (GA – do inglês, *Genetic Algorithm*)<sup>30,33</sup>. A técnica de seleção de variáveis permite a eliminação de informações não relevantes, como por exemplo, bandas que não contenham nenhuma informação das espécies ou propriedades a serem analisadas e a amplitude da relação sinal-ruído<sup>30, 31</sup>.

Somente será discutido em detalhes o método iPLS, utilizado no desenvolvimento deste trabalho

#### 2.3.2.4.1 PLS por intervalos (i-PLS)

O método iPLS é uma extensão do PLS, que desenvolve modelos locais PLS em subintervalos equidistantes de toda a região do espectro.

Seu principal objetivo é prever informação relevante nas diferentes subdivisões do espectro global, de forma a remover as regiões espectrais cujas variáveis se apresentam como supostamente de menor relevância e/ou interferentes. A partir deste ponto, um novo modelo PLS é construído a partir das variáveis selecionadas<sup>29</sup>.

No iPLS são realizadas regressões por mínimos quadrados parciais em subintervalos, de igual peso, de todo o espectro.

O espectro é dividido em tantas partes quanto se desejar, até que, através de tentativa e erro, chega-se a uma divisão ótima, ou seja, obtêm-se regiões do espectro com menores valores de RMSECV. Outro parâmetro utilizado é o  $R^2$

(coeficiente de correlação), que é a inclinação da reta do gráfico dos valores reais e previstos pelo modelo. Amostras e/ou medidas anômalas detectadas pelo PLS devem ser geralmente removidas antes da aplicação do iPLS.

Os novos modelos construídos são avaliados igualmente, como em um modelo PLS convencional. A diferença consiste, apenas, na divisão do conjunto de dados em intervalos iguais. O método é planejado para dar uma visão geral dos dados e pode ser útil para selecionar as variáveis mais representativas na construção de um modelo de calibração adequado. Porém, o método iPLS indica a região que está contida a informação, sendo uma aproximação univariada, pois não fornece sinergismo das regiões espectrais envolvidas <sup>7</sup>.

Para selecionar os subintervalos, a fim de obter melhores habilidades preditivas, pode ser utilizado, também, o algoritmo dos mínimos quadrados parciais por sinergismo de intervalos (siPLS), uma extensão do iPLS. Este possibilita selecionar a melhor combinação de intervalos, combinando 2 a 2, 3 a 3 e até 4 a 4 sub-regiões do espectro, fornecendo geralmente melhores coeficientes de determinação e os menores erros de predição que o iPLS <sup>7</sup>.

## **3 OBJETIVOS**

### 3.1 OBJETIVOS GERAIS

Este trabalho tem por objetivo geral desenvolver uma metodologia analítica capaz de identificar e quantificar o petróleo de origem à partir de suas misturas de frações de diesel, bem como estimar a densidade e a viscosidade das mesmas, utilizando quimiometria.

### 3.2 OBJETIVOS ESPECÍFICOS

A estratégia adotada para o alcance do objetivo geral deste trabalho é fazer uso da espectrometria no infravermelho associada a ferramentas quimiométricas onde, nesse processo, podem se destacar alguns objetivos específicos:

- Realizar estudos exploratórios utilizando a PCA tentando identificar, através da análise de gráficos, a fração predominante em misturas de frações de diesel de diferentes origens;
- Construir modelos de calibração multivariada por iPLS e siPLS, comparando-os quanto à habilidade de quantificar os petróleos de origem nas misturas de frações de diesel através de espectrometria no infravermelho médio.
- Construir modelos de calibração multivariada por PCR comparando-os, quanto ao desempenho na quantificação dos petróleos de origem nas misturas de frações de diesel, com os modelos construídos por iPLS e siPLS
- Construir modelos de calibração multivariada por iPLS e siPLS, comparando-os quanto à habilidade em estimar a densidade e a viscosidade em misturas de frações de diesel através de espectrometria no infravermelho médio.
- Construir modelos de calibração multivariada por PCR comparando-os quanto ao desempenho na determinação da densidade e da viscosidade em misturas de frações de diesel, com os modelos construídos por iPLS e siPLS

## **4 METODOLOGIA**

A metodologia empregada nesse trabalho se divide em três partes: o preparo das amostras, a medição dos espectros, da densidade e da viscosidade do conjunto de amostras (parte experimental), e a implementação de algumas das principais técnicas quimiométricas existentes para a construção dos modelos de calibração (parte computacional).

#### 4.1 PREPARO DAS AMOSTRAS

Utilizou-se para o desenvolvimento deste trabalho frações obtidas por destilação de três petróleos diferentes, as quais foram geradas pela destilação realizada no CENPES/Petrobras e, devido à política interna da empresa e o contrato entre as partes envolvidas no projeto, os dados fornecidos – tanto das frações como do petróleo – são sigilosos, sendo assim, serão referenciados por códigos (petróleos I, II e III).

Foram selecionadas as frações geradas de cada um dos três petróleos na faixa da destilação de 210°C a 260°C aproximadamente, região esta que compreende a faixa do diesel. Portanto, como pode ser observado nas tabelas 4.1, 4.2 e 4.3, a fração 3 do petróleo I foi selecionada para a realização deste trabalho visto que corresponde à faixa de temperatura de interesse, enquanto que do petróleo III, foi selecionada a fração 2, sendo que estas frações foram identificadas como frações A e C respectivamente. Já quanto ao petróleo II, foi necessária uma combinação igualitária, feita em balão volumétrico, das frações 8, 9 e 10 para que se obtivesse uma nova fração na faixa de interesse, a qual foi identificada como fração B.

##### 4.1.2 Misturas

A partir das frações A, B e C foram produzidas 147 combinações das mesmas entre si, com teores de cada uma variando entre 0 e 65,33% (v/v, conforme pode ser observado na tabela 4.4. Foram preparados 30 mL de cada amostra, em temperatura ambiente e utilizando buretas graduadas de diversos volumes.

**Tabela 4.1** Temperaturas de destilação das frações do petróleo I

<b>FRAÇÃO</b>	<b>TEMPERATURA (°C)</b>
0	15
1	148
2	211
3	252
4	284
5	305
6	342
7	370
8	400
9	440
10	448
11	457
12	462
13	476
14	488
15	506
16	518
17	533
18	550

**Tabela 4.3** Temperaturas de destilação das frações do petróleo III

<b>FRAÇÃO</b>	<b>TEMPERATURA (°C)</b>
0	15
1	223
2	259
3	281
4	315
5	328
6	336
7	360
8	381
9	400
10	444
11	466
12	510
13	519

**Tabela 4.2** Temperaturas de destilação das frações do petróleo II

<b>FRAÇÃO</b>	<b>TEMPERATURA (°C)</b>
0	15
1	86
2	111
3	134
4	157
5	179
6	201
7	210
8	226
9	243
10	260
11	277
12	293
13	308
14	330
15	345
16	360
17	377
18	393
19	400
20	426
21	444
22	447
23	465
24	475
25	497
26	514
27	530
28	551
29	570

**Tabela 4.4** Proporção, em volume, das frações A, B e C nas misturas formadas

<b>AM</b>	<b>A</b>	<b>B</b>	<b>C</b>												
1	0,50	0,50	0,00	38	0,15	0,52	0,33	75	0,33	0,32	0,35	112	0,33	0,49	0,17
2	0,50	0,00	0,50	39	0,16	0,51	0,33	76	0,65	0,01	0,33	113	0,33	0,48	0,19
3	0,00	0,50	0,50	40	0,17	0,49	0,33	77	0,64	0,03	0,33	114	0,33	0,47	0,20
4	0,01	0,33	0,65	41	0,19	0,48	0,33	78	0,63	0,04	0,33	115	0,33	0,45	0,21
5	0,03	0,33	0,64	42	0,20	0,47	0,33	79	0,61	0,05	0,33	116	0,33	0,44	0,23
6	0,04	0,33	0,63	43	0,21	0,45	0,33	80	0,60	0,07	0,33	117	0,33	0,43	0,24
7	0,05	0,33	0,61	44	0,23	0,44	0,33	81	0,59	0,08	0,33	118	0,33	0,41	0,25
8	0,07	0,33	0,60	45	0,24	0,43	0,33	82	0,57	0,09	0,33	119	0,33	0,40	0,27
9	0,08	0,33	0,59	46	0,25	0,41	0,33	83	0,56	0,11	0,33	120	0,33	0,39	0,28
10	0,09	0,33	0,57	47	0,27	0,40	0,33	84	0,55	0,12	0,33	121	0,33	0,37	0,29
11	0,11	0,33	0,56	48	0,28	0,39	0,33	85	0,53	0,13	0,33	122	0,33	0,36	0,31
12	0,12	0,33	0,55	49	0,29	0,37	0,33	86	0,52	0,15	0,33	123	0,33	0,35	0,32
13	0,13	0,33	0,53	50	0,31	0,36	0,33	87	0,51	0,16	0,33	124	0,65	0,33	0,01
14	0,15	0,33	0,52	51	0,32	0,35	0,33	88	0,49	0,17	0,33	125	0,64	0,33	0,03
15	0,16	0,33	0,51	52	0,33	0,01	0,65	89	0,48	0,19	0,33	126	0,63	0,33	0,04
16	0,17	0,33	0,49	53	0,33	0,03	0,64	90	0,47	0,20	0,33	127	0,61	0,33	0,05
17	0,19	0,33	0,48	54	0,33	0,04	0,63	91	0,45	0,21	0,33	128	0,60	0,33	0,07
18	0,20	0,33	0,47	55	0,33	0,05	0,61	92	0,44	0,23	0,33	129	0,59	0,33	0,08
19	0,21	0,33	0,45	56	0,33	0,07	0,60	93	0,43	0,24	0,33	130	0,57	0,33	0,09
20	0,23	0,33	0,44	57	0,33	0,08	0,59	94	0,41	0,25	0,33	131	0,56	0,33	0,11
21	0,24	0,33	0,43	58	0,33	0,09	0,57	95	0,40	0,27	0,33	132	0,55	0,33	0,12
22	0,25	0,33	0,41	59	0,33	0,11	0,56	96	0,39	0,28	0,33	133	0,53	0,33	0,13
23	0,27	0,33	0,40	60	0,33	0,12	0,55	97	0,37	0,29	0,33	134	0,52	0,33	0,15
24	0,28	0,33	0,39	61	0,33	0,13	0,53	98	0,36	0,31	0,33	135	0,51	0,33	0,16
25	0,29	0,33	0,37	62	0,33	0,15	0,52	99	0,35	0,32	0,33	136	0,49	0,33	0,17
26	0,31	0,33	0,36	63	0,33	0,16	0,51	100	0,33	0,65	0,01	137	0,48	0,33	0,19
27	0,32	0,33	0,35	64	0,33	0,17	0,49	101	0,33	0,64	0,03	138	0,47	0,33	0,20
28	0,33	0,33	0,33	65	0,33	0,19	0,48	102	0,33	0,63	0,04	139	0,45	0,33	0,21
29	0,01	0,65	0,33	66	0,33	0,20	0,47	103	0,33	0,61	0,05	140	0,44	0,33	0,23
30	0,04	0,63	0,33	67	0,33	0,21	0,45	104	0,33	0,60	0,07	141	0,43	0,33	0,24
31	0,05	0,61	0,33	68	0,33	0,23	0,44	105	0,33	0,59	0,08	142	0,41	0,33	0,25
32	0,07	0,60	0,33	69	0,33	0,24	0,43	106	0,33	0,57	0,09	143	0,40	0,33	0,27
33	0,08	0,59	0,33	70	0,33	0,25	0,41	107	0,33	0,56	0,11	144	0,39	0,33	0,28
34	0,09	0,57	0,33	71	0,33	0,27	0,40	108	0,33	0,55	0,12	145	0,37	0,33	0,29
35	0,11	0,56	0,33	72	0,33	0,28	0,39	109	0,33	0,53	0,13	146	0,36	0,33	0,31
36	0,12	0,55	0,33	73	0,33	0,29	0,37	110	0,33	0,52	0,15	147	0,35	0,33	0,32
37	0,13	0,53	0,33	74	0,33	0,31	0,36	111	0,33	0,51	0,16				

## 4.2 DENSIDADE E VISCOSIDADE

Utilizou-se para esta determinação um Analisador Viscosímetro Digital Stabinger modelo SVM 3000/G2 para medir a viscosidade cinemática e a densidade.

Foram feitas análises em duas temperaturas: a 20<sup>o</sup> C e a 40<sup>o</sup> C. Para cada amostra foram realizadas duas medidas em cada temperatura, verificando-se a repetibilidade dos valores obtidos seqüencialmente, caso estivessem fora dos limites pré-estabelecidos, novas medidas eram feitas até que os mesmos se enquadrassem.

Para a construção dos modelos quimiométricos, foi utilizada a média entre os dois últimos valores encontrados de densidade a 20<sup>o</sup> C e de viscosidade a 40<sup>o</sup> C de cada amostra.

Estas determinações foram realizadas no LabPetro – Laboratório de Petróleo da UFES (Universidade Federal do Espírito Santo).

## 4.3 INFRAVERMELHO

O espectrômetro utilizado foi o ABB Bomen modelo MB 102 equipado com um detector de sulfato de triglicerina deuterado (DGTS). Os espectros foram obtidos em absorbância, na região do MIR, na faixa de 4000 a 600 cm<sup>-1</sup> com uma varredura de 32 scans e uma resolução de 4 cm<sup>-1</sup> em um cristal de seleneto de zinco com ângulo de incidência de 45<sup>o</sup>C. Ao todo obteve-se 1764 comprimentos de onda (variáveis).

Anteriormente à obtenção do espectro de cada amostra foi feito o *background* utilizando-se o ar como referência para a correção da linha de base. Essas medições foram realizadas em todas as misturas produzidas, bem como nas suas frações de origem.

Estes ensaios foram realizados no LEC – Laboratório de Ensaio de Combustíveis da UFMG (Universidade Federal de Minas Gerais). Para aquisição dos espectros foi utilizado o programa GRAMS/AI 7.00.

#### 4.4 APLICATIVO PARA CRIAÇÃO DOS MODELOS

Todos os dados obtidos das amostras (densidade, viscosidade e espectros de infravermelho) foram trabalhados no programa MINITAB 14, para criação dos modelos de calibração. Foram feitas Análise por Componentes Principais (PCA), Regressão por Mínimos Quadrados Parciais (PLSR) e Regressão por Componentes Principais (PCR) a fim de se verificar a separação entre as 3 frações, bem como, quantificá-las. A PLSR e a PCR também foram realizadas com o intuito de se estimar a densidade e a viscosidade de cada mistura formada. Alguns cálculos foram realizados com o Aplicativo Excel 2003, da Microsoft. O programa Origin 7.0 foi utilizado no desenho de alguns gráficos.

#### 4.5 SELEÇÃO DE AMOSTRAS

Dentre as 150 amostras analisadas (3 frações puras e 147 misturas de frações), 100 foram selecionadas para o conjunto de calibração de maneira que representassem todo o conjunto. As 50 restantes foram utilizadas apenas na etapa de validação do modelo.

#### 4.6 SELEÇÃO DA FAIXA ESPECTRAL

Para a criação do modelo foi selecionada a região de 1319,2 a 1546,7  $\text{cm}^{-1}$  e de 2756 a 3062,7  $\text{cm}^{-1}$ , pois a mesma mostrou-se bastante interessante por se tratar de uma região onde ocorrem várias sobreposições de bandas e percebe-se uma maior diferença visual dos espectros (o perfil dos espectros gerados das frações A, B e C pode ser observado no anexo A). Essa região foi identificada como “Região I”.

Além disso, foi empregado o método de seleção de variáveis iPLS. As 1764 absorvâncias dos espectros originais foram divididas em 20 intervalos. Os pontos iniciais e finais de cada intervalo estão descritos na tabela 4.5. Para se verificar o sinergismo entre os intervalos foi aplicado o método siPLS.

A seleção dos intervalos a serem utilizados foi feita, inicialmente, pelo valor da Raiz Quadrada do Erro Médio Quadrático de Validação Cruzada (RMSECV) e pelo número de variáveis latentes. A partir dos valores obtidos foram selecionados os intervalos mais correlacionados e estes foram utilizados na construção dos modelos.

**Tabela 4.5** Faixa de absorvância dos 20 intervalos formados

INTERVALO	INÍCIO (cm <sup>-1</sup> )	FINAL (cm <sup>-1</sup> )
1	599,81	769,54
2	771,47	941,19
3	943,12	1112,8
4	1114,7	1284,4
5	1286,4	1454,2
6	1456,1	1623,9
7	1625,8	1793,6
8	1795,5	1963,3
9	1965,3	2133,1
10	2135	2302,8
11	2304,7	2472,5
12	2474,4	2642,2
13	2644,2	2812
14	2813,9	2981,7
15	2983,6	3151,4
16	3153,3	3321,1
17	3323,1	3490,9
18	3492,8	3660,6
19	3662,5	3830,3
20	3832,2	4000

#### 4.7 ANÁLISE POR COMPONENTES PRINCIPAIS (PCA)

A análise por componentes principais foi feita com os seguintes objetivos:

- Identificar amostras anômalas (*outliers*);
- Tentar separar, através dos gráficos dos *scores*, as amostras de acordo com as frações de maior predominância em cada mistura;
- Calcular as componentes principais das amostras de calibração e validação;
- Verificar a correlação entre cada componente e as frações A, B e C;
- Verificar a correlação entre cada componente e os valores de densidade e viscosidade;

#### **4.7.1 Separação de amostras**

Foi realizada a análise de componentes principais, em cada um dos 20 intervalos, utilizando todas as amostras (validação e calibração) com o objetivo de, a partir dos gráficos dos *scores*, tentar identificá-las de acordo com a fração predominante.

Foram calculadas as correlações entre as quantidades das frações A, B e C e as componentes principais de cada intervalo individualmente. Posteriormente, foram *plotados* gráficos 3D dos *scores*, utilizando as componentes com maior correlação com as 3 frações e feitas comparações buscando-se a melhor separação entre elas. A partir desses gráficos foi verificada, também, a existência de *outliers* ou amostras anômalas.

#### **4.7.2 Seleção das componentes principais**

Foram calculadas novas componentes, porém utilizando apenas as amostras de calibração. Após o cálculo das mesmas em cada faixa espectral pré-selecionada, foi calculada a correlação entre as quantidades das frações A, B e C e as componentes individualmente para cada uma dessas faixas. Esse mesmo cálculo também foi feito entre os valores de densidade e viscosidade e as componentes.

A partir dos valores obtidos foram selecionadas as componentes que carregavam maior informação relevante (com maior correlação) em relação aos dados a serem previstos para utilização na PCR.

#### **4.7.3 Cálculos das componentes principais das amostras de validação**

A partir dos *loadings* obtidos no cálculo das PCs das amostras de calibração, foram calculados os *scores* das amostras de validação no modelo inicial para que se obtivesse, assim, as coordenadas das mesmas no modelo criado. Esse procedimento foi realizado com o objetivo de se obter as componentes de validação do modelo a serem utilizadas na PCR.

#### 4.8 ELABORAÇÃO DOS MODELOS DE CALIBRAÇÃO

Por fim, foram construídos os modelos de calibração através de 3 diferentes métodos: iPLS, siPLS e PCR para a determinação quantitativa das frações A, B e C em cada uma das amostras analisadas e, também, da densidade e viscosidade de cada uma delas.

Foram desenvolvidos modelos de calibração para a determinação das frações A, B e C nas amostras, individualmente, bem como de suas respectivas densidades e viscosidades, ou seja, buscou-se o melhor modelo individual para cada fração e para os valores das densidades e viscosidades. Desenvolveu-se, também, um modelo global para a quantificação simultânea das frações A, B e C.

Primeiramente foram gerados os modelos por mínimos quadrados parciais nos intervalos selecionados e na região I. Posteriormente, foram desenvolvidos os modelos por mínimos quadrados parciais por sinergismo de intervalos (si-PLS). Os modelos gerados foram, então, testados com as amostras de validação.

Em seguida foram desenvolvidos e testados modelos por Regressão por Componentes Principais tanto na região I quanto nos intervalos (simples e combinados) selecionados para geração e criação dos modelos por PLS com o objetivo de se comparar o desempenho desses dois métodos.

#### 4.9 AVALIAÇÃO DOS MODELOS DE CALIBRAÇÃO

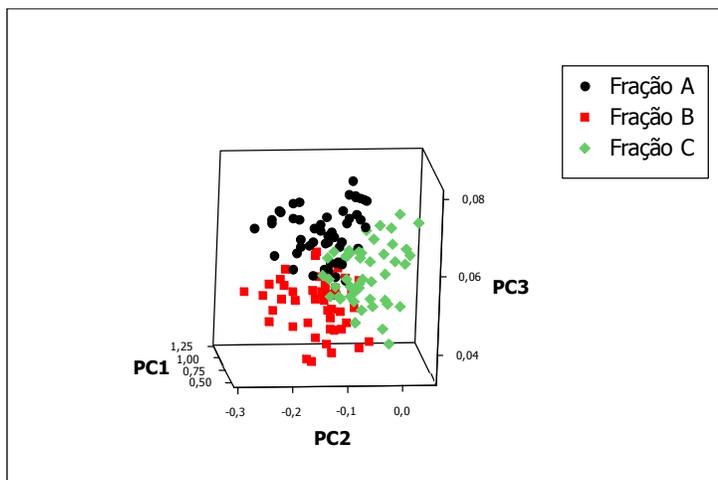
Para avaliação dos modelos de calibração criados foram calculados os Coeficientes de Correlação Linear ( $R^2$ ), a Raiz Quadrada do Erro Quadrático Médio de Validação Cruzada (RMSECV), Raiz Quadrada do Erro Quadrático Médio de Calibração (RMSEC) e a Raiz Quadrada do Erro Quadrático Médio de Predição (RMSEP). Foi verificada também a presença de pontos discrepantes ou *outliers*. Após a retirada dos pontos discrepantes e a avaliação da relação entre os erros de previsão e a performance do modelo construído, o mesmo foi recalibrado.

Os modelos foram avaliados, também, quanto ao número de variáveis latentes (VLs) ou componentes principais (CP) utilizados na construção dos modelos por mínimos quadrados parciais e regressão por componentes principais, respectivamente.

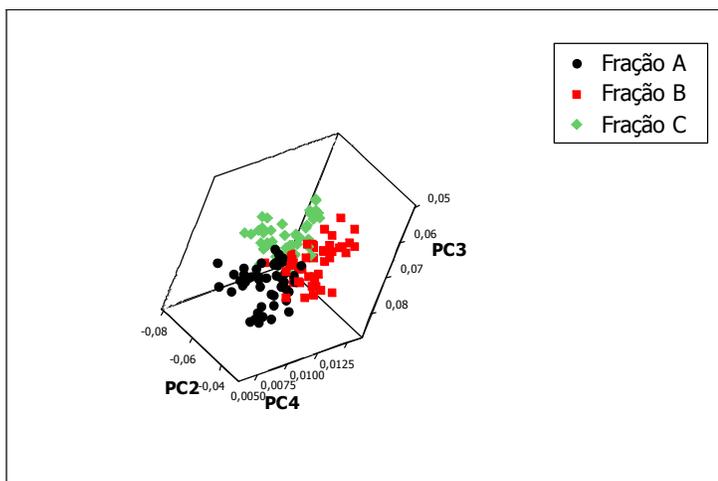
## **5 RESULTADOS E DISCUSSÕES**

## 5.1 IDENTIFICAÇÃO DAS FRAÇÕES

Inicialmente foi feita uma pré-seleção dos intervalos formados (tabela 4.5, pág. 41) observando a correlação entre as componente principais de cada um deles e as concentrações das frações A, B e C. Posteriormente foram *plotados* gráficos 3D dos *scores*, nos intervalos selecionados, relacionando as componentes principais que possuíam maior correlação com as frações A, B e C em cada intervalo, verificando se ocorria uma separação visual que permitisse a identificação das amostras de acordo com a fração predominante em cada uma. Os melhores resultados foram observados nos intervalos 2 e 3 e as figuras 5.1 e 5.2 trazem os respectivos gráficos gerados.



**Figura 5.1** Gráfico dos *scores* para o intervalo 2 relacionando as PCs 1, 2 e 3.



**Figura 5.2** Gráfico dos *scores* para o intervalo 3 relacionando as PCs 2, 3 e 4.

Em ambos os intervalos observa-se uma boa separação entre as frações. A tabela 5.1 mostra os respectivos valores de variância acumulada para as 5 primeiras componentes principais nesses intervalos.

**Tabela 5.1** Covariância acumulada para as 5 primeiras PCs nos intervalos 2 e 3

<b>COMPONENTE</b>	<b>INTERVALO 2</b>	<b>INTERVALO 3</b>
PC1	84,63	99,34
PC2	99,42	99,72
PC3	99,82	99,96
PC4	99,95	99,98
PC5	99,98	99,99

Analisando-se a tabela 5.1 e os gráficos 5.1 e 5.2 verifica-se que, embora as 3 primeiras PCs do intervalo 3 expliquem 99,96% dos dados, sendo 99,34% apenas da primeira PC, o gráfico gerado relacionando essas 3 PCs não resultou em uma boa separação dos dados. O melhor resultado, para este intervalo, foi encontrado relacionando as PCs 2, 3 e 4. Observa-se, na tabela 5.2, que essa 1ª. PC não possui uma boa correlação com as proporções em volumes das frações A, B e C nas misturas. Ao contrário desse resultado, embora a 1ª. PC do intervalo 2 explique apenas 84,66% dos dados, ela possui uma boa correlação com os volumes das frações e por isso o gráfico dos *scores* relacionando as 3 primeiras componentes (99,82% da variância acumulada) obteve um bom desempenho.

**Tabela 5.2** Correlações entre os volumes das frações A, B e C nas misturas e as 5 primeiras PCs nos intervalos 2 e 3

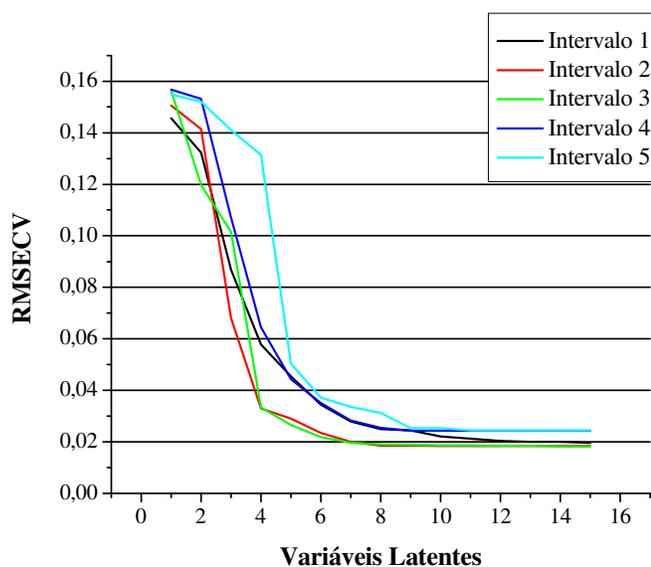
<b>COMPONENTE</b>	<b>INTERVALO 2</b>			<b>INTERVALO 3</b>		
	<b>A</b>	<b>B</b>	<b>C</b>	<b>A</b>	<b>B</b>	<b>C</b>
PC1	0,186	0,256	-0,436	-0,008	-0,07	0,078
PC2	-0,316	-0,542	0,847	0,103	0,688	-0,781
PC3	0,801	-0,765	-0,035	0,562	0,058	-0,612
PC4	-0,433	0,178	0,252	-0,803	0,664	0,137
PC5	-0,013	-0,06	0,072	-0,248	0,266	-0,018

## 5.2 QUANTIFICAÇÃO DAS FRAÇÕES

### 5.2.1 Quantificação da fração A

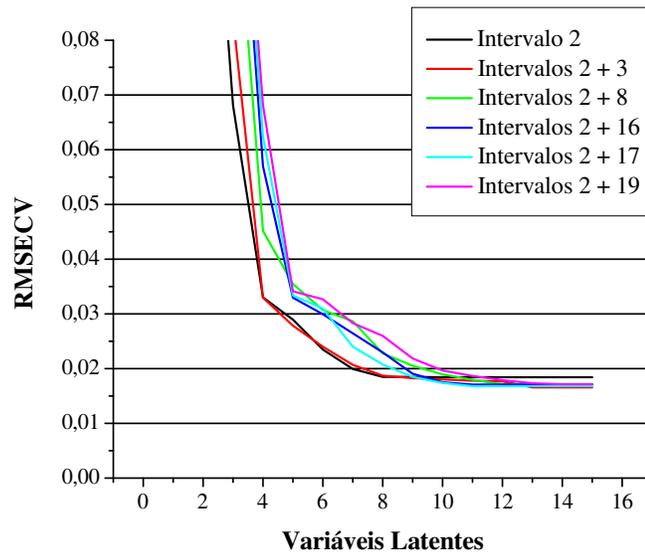
#### 5.2.1.1 Aplicação do método PLS para a quantificação da fração A

Na tentativa de se quantificar a fração A nas misturas, inicialmente calculou-se os valores de RMSECV e VL para cada um dos 20 intervalos em que o espectro foi dividido (tabela 4.5, pág. 41). A figura 5.3 apresenta o perfil do RMSECV para as 5 melhores regiões. Esses valores foram calculados, também, para a região I (1319,2 a 1546,7 e 2756 a 3062,7  $\text{cm}^{-1}$ ).



**Figura 5.3** Perfil do RMSECV para a fração A com modelos construídos para as cinco melhores regiões.

Como se pode observar, os dados contidos nos intervalos 2 e 3 resultaram nos melhores modelos e com valores bem próximos de RMSECV, porém o intervalo 2 obteve um desempenho ligeiramente melhor. Com isso, manteve-se o intervalo 2 fixo e adicionou-se a ele cada intervalo em separado, para verificar se ocorria uma melhora significativa no modelo, sempre em relação ao RMSECV. A figura 5.4 mostra os valores de RMSECV para os 5 melhores modelos obtidos combinando-se o intervalo 2 com os demais.



**Figura 5.4** Perfil do RMSECV da fração A para a associação de duas regiões para as 5 melhores combinações.

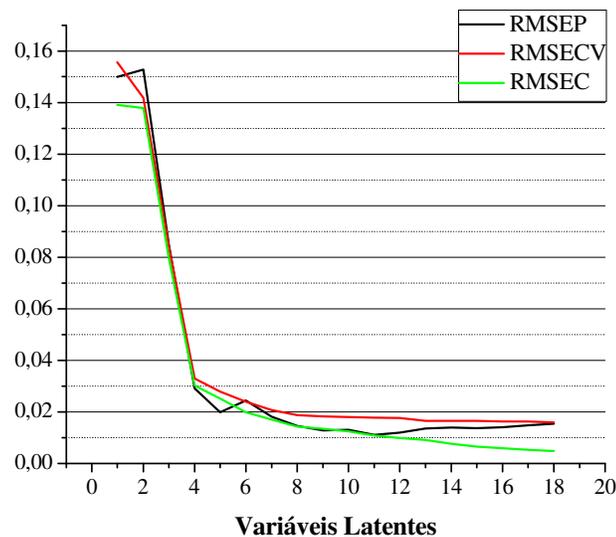
Observa-se na figura 5.4, que não há mais uma significativa diminuição do valor do RMSECV com a combinação de 2 intervalos em comparação ao valor de RMSECV da figura 5.3. Desta maneira, resolveu-se não prosseguir com as associações de intervalos, pois presumiu-se que a partir deste ponto poderia estar agregando ao modelo maior quantidade de dados prejudiciais do que benéficos a ele. Mesmo assim, para os estudos seguintes, o melhor resultado observado na figura 5.4 foi levado em consideração, que é associação dos intervalos 2 e 3.

#### *5.2.1.1.2 Definição dos melhores modelos obtidos por PLS na quantificação da fração A e predição de amostras externas*

Para a escolha do número de variáveis latentes, necessárias para que o modelo tenha boa robustez e capacidade de predição com segurança do valor de concentração para amostras externas, fez-se os gráficos combinando os valores de RMSECV, RMSEC e RMSEP dos modelos selecionados até o momento. A escolha foi feita para regiões onde os valores de RMSECV, RMSEC e RMSEP estivessem

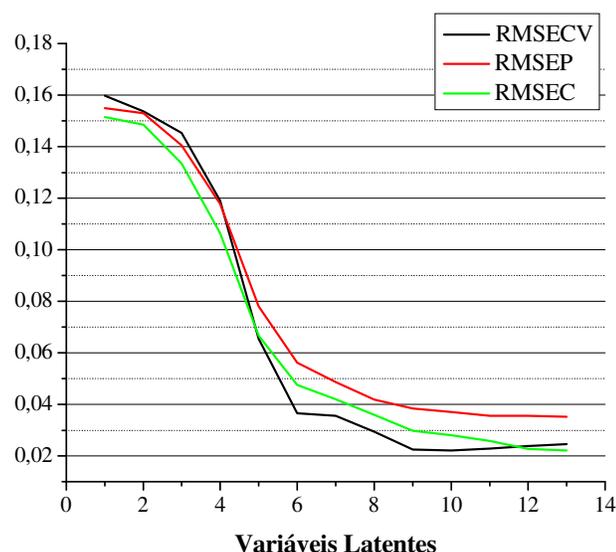
próximos e com os valores menores possíveis. Para essa seleção, foi realizada uma análise dos gráficos ponto a ponto e selecionados aqueles em que a diferença entre um erro e outro não fosse maior que 0,01. Nesta situação, não há risco de se estar supervalorizando os modelos de calibração (*overfitting*) ou adicionando-se ruído aos modelos.

Para os modelos construídos com a associação dos intervalos 2 e 3, o gráfico resultante está mostrado na figura 5.5. Observa-se que, com 8, 9 e 10 variáveis latentes, o modelo apresenta valores baixos e próximos de RMSECV, RMSEC e RMSEP.



**Figura 5.5** Perfil do RMSECV, RMSEC e RMSEP variando-se o número de variáveis latentes para a fração A na associação dos intervalos 2 e 3.

Depois de todas as combinações propostas acima, verificou-se quais seriam as variáveis latentes, necessárias para o melhor desempenho de cada modelo. A tabela 5.3 mostra os resultados destes gráficos construídos. Para o modelo construído para a região I, não foi encontrada nenhuma variável latente com o perfil esperado, pois os valores de RMSECV, RMSEC e RMSEP se encontravam próximos apenas para valores bastante elevados dos mesmos conforme figura 5.6.



**Figura 5.6** Perfil do RMSECV, RMSEC e RMSEP variando-se o número de variáveis latentes para a fração A na região I.

**Tabela 5.3** Seleção do número de variáveis latentes para cada modelo proposto para a quantificação da fração A

Regiões do espectro	Variáveis Latentes
Intervalo 2	7, 8
intervalo 2 + 3	8, 9, 10

Os modelos foram, então, construídos com o número de variáveis latentes para todas as possibilidades acima para a predição das amostras externas, a fim de se verificar a habilidade do modelo em determinar a concentração de amostras não presentes no modelo. A tabela 5.4 mostra os valores de  $R^2$  e RMSEP para esses modelos construídos.

**Tabela 5.4** Valores de  $R^2$  e RMSEP para os modelos construídos para a quantificação da fração A

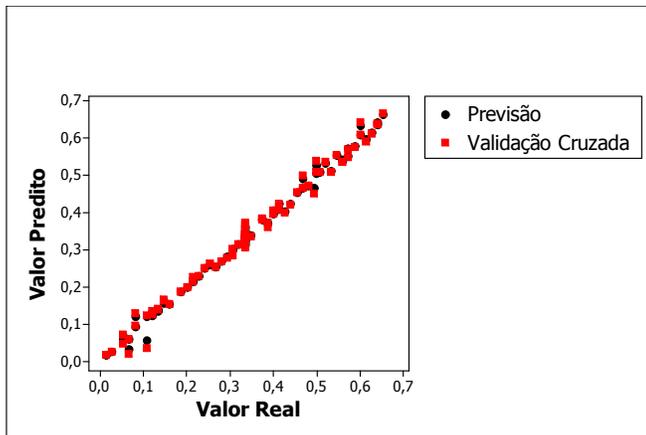
Regiões do espectro	VL	$R^2$	RMSEP
Intervalo 2	7	0,9902	0,0160
	8	0,9915	0,0131
Intervalos 2 + 3	8	0,9914	0,0146
	9	0,9924	0,0129
	10	0,9934	0,0131

Como o propósito deste trabalho é obter um modelo capaz de prever frações de petróleo em misturas das mesmas, a seleção dos modelos vistos na tabela 5.4 foi feita tomando-se por base o valor de RMSEP. Este parâmetro indica o erro padrão de predição para amostras externas ao modelo. Assim, selecionou-se para cada região o modelo com menor valor de RMSEP. A tabela 5.5 mostra os modelos selecionados por este critério.

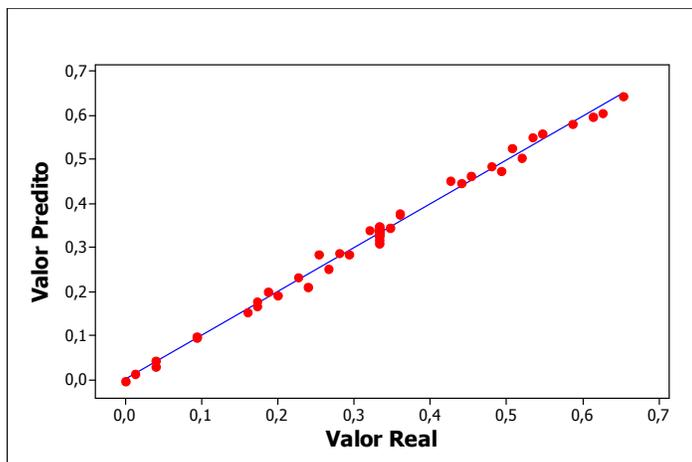
**Tabela 5.5** Modelos selecionados para a quantificação da fração A

Regiões do espectro	VL	R <sup>2</sup>	RMSEP
Intervalo 2	8	0,9915	0,0131
Intervalos 2 + 3	9	0,9924	0,0129

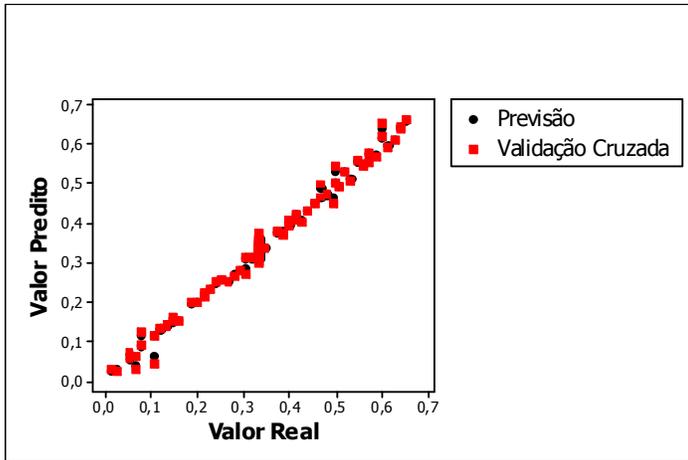
Verifica-se na tabela acima que ambas as regiões obtiveram valores satisfatórios de R<sup>2</sup> e RMSEP. Com o objetivo de se conferir o resultado de previsão dos modelos selecionados, foram construídos gráficos com os valores reais versus os valores preditos pelo modelo para amostras de calibração e de validação. As figuras 5.7 a 5.10 mostram os correspondentes gráficos para os modelos escolhidos.



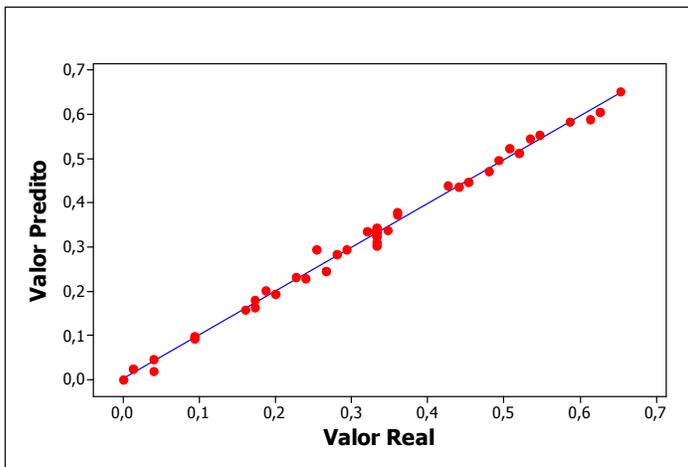
**Figura 5.7** Gráfico com o resultado de previsão da fração A em amostras de calibração no intervalo 2.



**Figura 5.8** Gráfico com o resultado de previsão da fração A em amostras de validação no intervalo 2.



**Figura 5.9** Gráfico com o resultado de predição da fração A em amostras de calibração na associação dos intervalos 2 e 3



**Figura 5.10** Gráfico com o resultado de predição da fração A em amostras de validação na associação dos intervalos 2 e 3.

Pode-se observar que as curvas obtidas com os valores reais *versus* valores preditos para as amostras de calibração e de validação mostram boa correlação. Observa-se, também, que os valores preditos para as amostras do conjunto de calibração seguem a tendência de aproximar-se da linha de tendência dos pontos de validação cruzada. Sabendo-se agora, que para os modelos testados tais características (tendência de aproximação da linha de regressão linear), foram encontradas, construiu-se a tabela 5.5, onde são apresentadas as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos.

**Tabela 5.6** Médias e Desvio Padrões para os modelos obtidos na quantificação da Fração A por PLS

<b>Regiões do espectro</b>	<b>Média Razão predito / real</b>	<b>Desvio Padrão</b>
Intervalos 2	0,9734	0,1537
Intervalos 2 + 3	0,9863	0,1998

Os resultados mostram que os modelos construídos nas duas regiões obtiveram um desempenho bastante semelhante, sendo o poder de previsão de ambos bastante satisfatórios. Com isso, conclui-se que para a quantificação da fração A por PLS não há necessidade de se associar intervalos, bastando a região correspondente ao intervalo 2 (771,47 a 941,19  $\text{cm}^{-1}$ ) para isso. Com esses resultados verifica-se também que embora a região I tenha se mostrado bastante interessante inicialmente, devido a ocorrência de bandas com absorbância significativa, ela não obteve o resultado esperado na previsão de amostras externas.

#### 5.2.1.2 Aplicação do método PCR para a quantificação da fração A

Foram construídos modelos por PCR tanto para a região I quanto nos intervalos e suas combinações selecionados para a construção dos modelos por mínimos quadrados parciais conforme tabela 5.3 (pág. 51).

Para a construção dos modelos em cada região, foram estabelecidos 4 valores mínimos de correlação entre as componentes principais e a fração A: 0,5, 0,2, 0,1 e 0,05. A partir desses valores foram selecionadas as componentes a serem utilizadas em cada modelo. A tabela 5.7 traz o número de PCs utilizadas na construção de cada modelo. Por exemplo, quando se usa três componentes principais no intervalo 2, o menor valor de correlação encontrado é 0,2, quando se usa 8 PCs, no mesmo intervalo, o valor mínimo de correlação entre a PC e as concentrações é de 0,05.

**Tabela 5.7** Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação da Fração A

Região	Valor Mínimo de Correlação	Nº de PCs utilizadas
Região I (1319,2 a 1546,7 e 2756 a 3062,7 cm <sup>-1</sup> )	0,5	1
	0,2	3
	0,1	7
	0,05	14
Intervalo 2 (771,47 a 941,19 cm <sup>-1</sup> )	0,5	1
	0,2	3
	0,1	5
	0,05	8
Intervalo 2 + 3 (771,47 a 941,19 e 943,12 a 1112,8 cm <sup>-1</sup> )	0,5	2
	0,2	3
	0,1	5
	0,05	9

Os modelos foram, então, construídos com o número de componentes principais para todas as possibilidades acima para a predição das amostras externas, a fim de se verificar a habilidade em se determinar a concentração de amostras não presentes no modelo.

#### *5.2.1.2.1 Definição dos melhores modelos obtidos por PCR na quantificação da fração A e predição de amostras externas*

Como primeiro critério de seleção dos modelos, adotou-se o valor de  $R^2$ , excluindo-se de uma análise mais minuciosa os modelos com  $R^2$  menor que 0,90. A tabela 4.8 apresenta os modelos que foram obtidos com  $R^2$  maiores que 0,90. Este valor de 0,90 foi estipulado para limitar a linearidade, pois quanto mais linear, melhor e mais confiável será o modelo.

**Tabela 5.8** Modelos, da fração A, que apresentaram  $R^2$  maior que 0,90

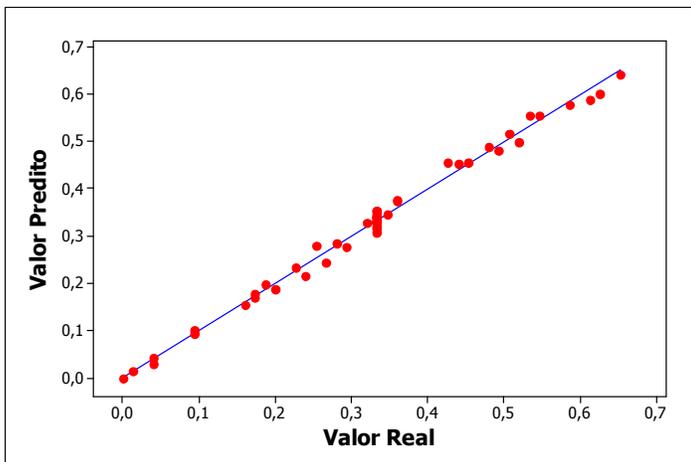
Região	Nº de PCs utilizadas	$R^2$
Região I	7	0,938
	14	0,970
Intervalo 2	3	0,931
	5	0,974
	8	0,990
Intervalos 2 + 3	3	0,947
	5	0,973
	9	0,988

Após a seleção dos modelos que apresentaram melhores valores para os coeficientes de correlação, escolheu-se o segundo critério de seleção dos melhores modelos. Como o propósito deste trabalho é obter um modelo capaz de prever frações de petróleo em misturas das mesmas, a seleção dos modelos vistos na tabela 5.8 foi feita tomando-se por base o valor de RMSEP. Assim, assumiu-se, como limite de seleção o valor de 0,020, ou seja, selecionou-se todos os modelos com valores inferiores a 0,020. O valor de 0,020 foi estipulado arbitrariamente, pois entendeu-se que, a partir deste valor, o erro seria muito elevado para as amostras externas ao modelo. A tabela 5.9 mostra os modelos selecionados por este critério.

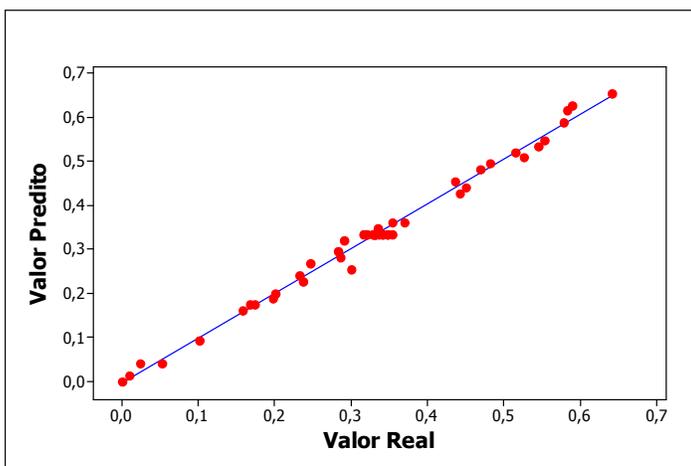
**Tabela 5.9** Modelos, da fração A, que apresentaram  $R^2$  maior que 0,90 e RMSEP menor que 0,020 por PCR

Região	Nº de PCs utilizadas	RMSEP
Intervalo 2	8	0,0145
Intervalos 2 + 3	9	0,0146

Os modelos representados na tabela 5.9 foram, então, os selecionados para expressar os resultados para a predição das amostras externas. Para se conferir esses resultados de previsão, as figuras 5.11 e 5.12 mostram os correspondentes gráficos relacionando os valores preditos e os valores reais para as amostras de validação nos modelos escolhidos.



**Figura 5.11** Gráfico com o resultado de predição da fração A em amostras de validação no intervalo 2 por PCR.



**Figura 5.12** Gráfico com o resultado de predição da fração A em amostras de validação na associação dos intervalos 2 e 3 por PCR.

Observa-se que para ambas as regiões as curvas obtidas com os valores reais *versus* valores preditos para as amostras de validação mostram boa correlação. A tabela 5.10 traz, então, as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos.

**Tabela 5.10** Média e desvio padrões para os modelos obtidos na quantificação da Fração A por PCR

Regiões do espectro	Média Razão predito / real	Desvio Padrão
Intervalo 2	0,9665	0,1587
Intervalos 2 + 3	0,9739	0,1719

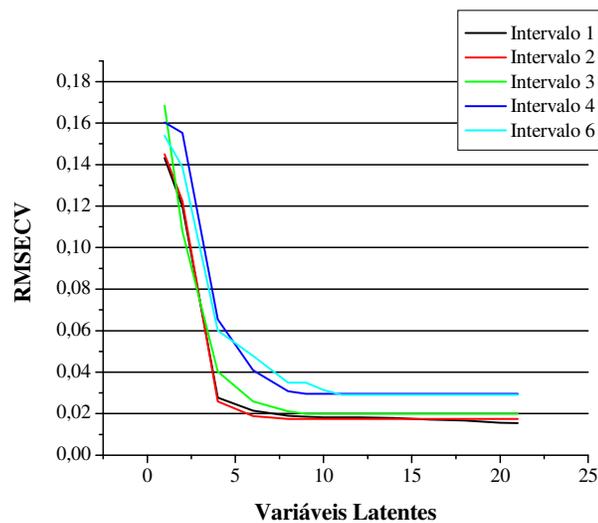
Com esses resultados observa-se que, assim como na construção dos modelos por PLS, os modelos obtidos por PCR nas duas regiões obtiveram um desempenho

bastante semelhante, sendo o poder de previsão de ambos bastante satisfatórios. Com isso, verifica-se que para a quantificação da fração A por PCR também não há necessidade de se associar intervalos, bastando a região correspondente ao intervalo 2 para isso. A região I, igualmente ao ocorrido na modelagem por PLS, não obteve resultados satisfatórios. Os modelos obtidos por PCR e PLS obtiveram desempenhos parecidos com valores de RMSEP e número de VLs e PCs bastante próximos.

## 5.2.2 Quantificação da fração B

### 5.2.2.1 Aplicação do método PLS para a quantificação da fração B

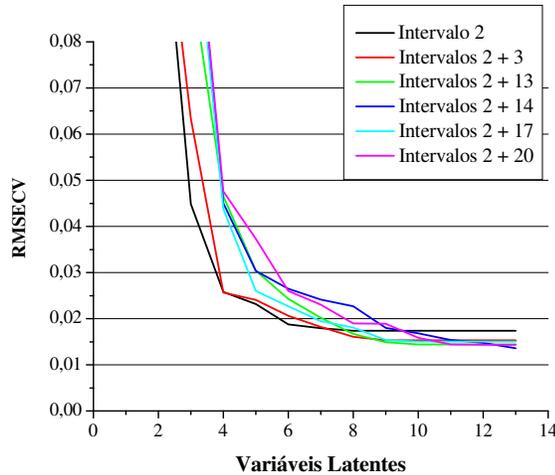
Procedeu-se, para a quantificação da fração B, da mesma forma como discutido para a fração A utilizando-se, também, os mesmos critérios. Portanto os perfis do RMSECV para os 5 intervalos que apresentaram o melhor desempenho quanto ao RMSECV na aplicação do PLS estão representados na figura 5.13.



**Figura 5.13** Perfil do RMSECV para a fração B com modelos construídos para as cinco melhores regiões.

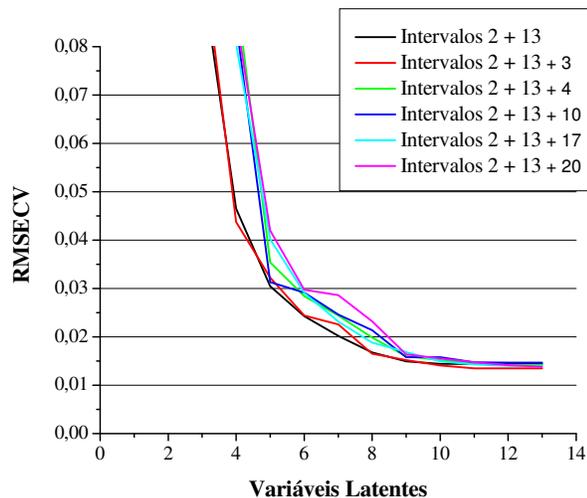
Como se pode observar, os dados contidos no intervalo 2 resultaram no melhor modelo. Com isso ele foi o selecionado para ser mantido fixo e os outros intervalos

foram adicionados a ele em separado. A figura 5.14 mostra os valores de RMSECV para os 5 melhores modelos obtidos combinando-se o intervalo 2 com os demais.



**Figura 5.14** Perfil do RMSECV da fração B para a associação de duas regiões para as 5 melhores combinações.

Verifica-se que a combinação do intervalo 2 com 13 levou a observação dos menores valor de RMSECV. Partiu-se, então, para uma terceira combinação, onde manteve-se as regiões 2 e 13 fixas, adicionando-se uma terceira região (figura 5.15).



**Figura 5.15** Perfil do RMSECV da fração B para a associação de três regiões para as 5 melhores combinações.

De acordo com o observado na figura 5.15, conclui-se que não há mais uma significativa diminuição do valor do RMSECV com a combinação de 3 intervalos em

comparação ao valor de RMSECV observado na figura 5.14 e, por isso, resolveu-se não prosseguir com as associações de intervalos, porém o melhor resultado observado na figura (associação dos intervalos 2, 3 e 13) foi levado em consideração para os estudos seguintes.

#### 5.2.2.1.1 Definição dos melhores modelos obtidos por PLS na quantificação da fração B e predição de amostras externas

Verificou-se, então, quais seriam as variáveis latentes necessárias para o melhor desempenho de cada modelo, através dos gráficos relacionando os valores de RMSECV, RMSEC e RMSEP. A tabela 5.11 mostra os resultados destes gráficos construídos. Para o modelo construído na região I, não foi encontrada nenhuma variável latente com o perfil esperado.

**Tabela 5.11** Seleção do número de variáveis latentes para cada modelo proposto para a quantificação da fração B

<b>Regiões do espectro</b>	<b>Variáveis Latentes</b>
Intervalo 2	6, 7, 8
Intervalos 2 + 13	8, 9
Intervalos 2 + 13 + 3	8, 9

Os modelos foram, então, construídos com o número de variáveis latentes para todas as possibilidades acima para a predição das amostras externas. A tabela 5.12 mostra os valores de  $R^2$  e RMSEP para os modelos selecionados até então.

**Tabela 5.12** Valores de  $R^2$  e RMSEP para os modelos construídos para a quantificação da fração B

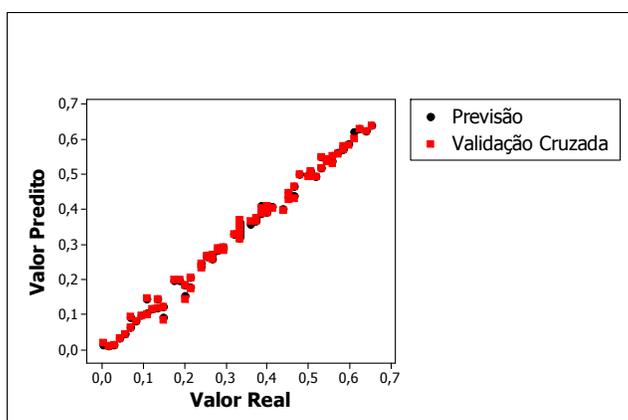
<b>Regiões do espectro</b>	<b>VL</b>	<b><math>R^2</math></b>	<b>RMSEP</b>
Intervalo 2	6	0,9883	0,0150
	7	0,9903	0,0150
	8	0,9918	0,0151
Intervalo 2 + 13	8	0,9918	0,0147
	9	0,9941	0,0131
Intervalo 2 + 13 + 3	8	0,9929	0,0133
	9	0,9940	0,0126

Em cada região foi selecionado, como para a quantificação da fração A, o menor valor de RMSEP conforme tabela 5.13.

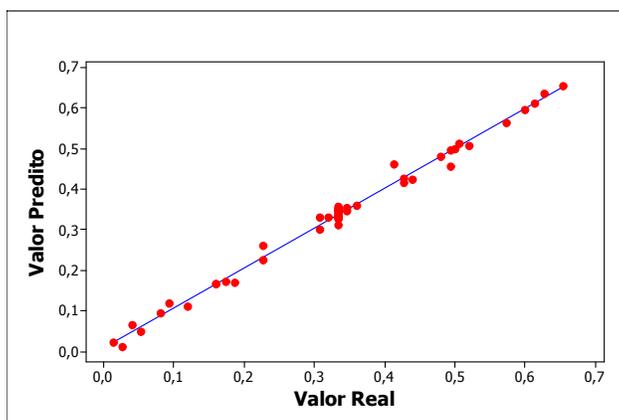
**Tabela 5.13** Modelos selecionados para a quantificação da fração B

Regiões do espectro	VL	R <sup>2</sup>	RMSEP
Intervalo 2	6	0,9883	0,01499
Intervalo 2 + 13	9	0,9941	0,01306
Intervalo 2 + 13 + 3	9	0,9940	0,01257

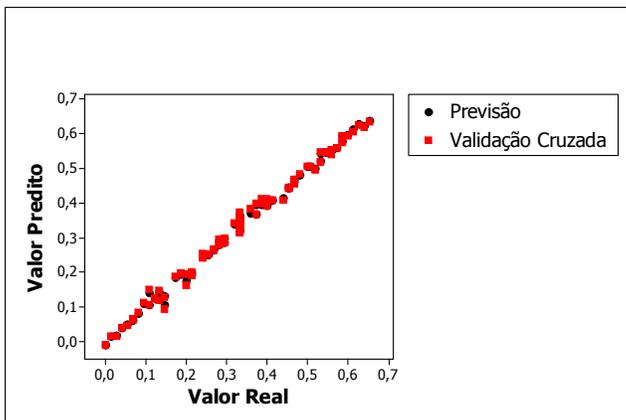
Para se conferir o resultado de previsão para amostras de calibração e previsão nos modelos selecionados plotou-se, então, os gráficos contendo os valores reais versus os valores preditos pelos modelos. As figuras 5.16 a 5.21 mostram os correspondentes gráficos para os modelos escolhidos.



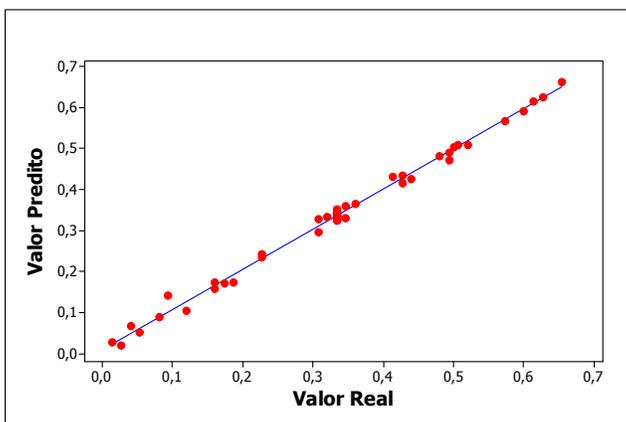
**Figura 5.16** Gráfico com o resultado de previsão da fração B em amostras de calibração no intervalo 2.



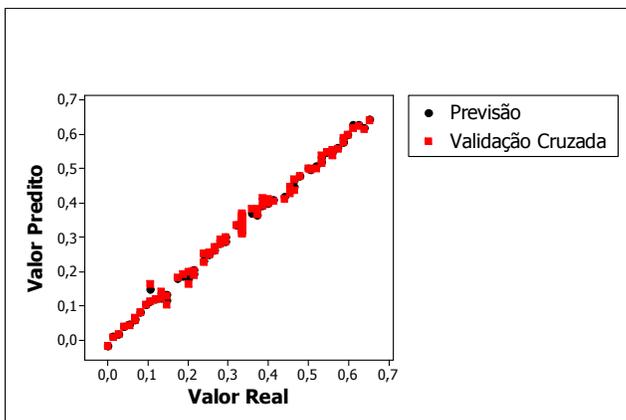
**Figura 5.17** Gráfico com o resultado de previsão da fração B em amostras de validação no intervalo 2.



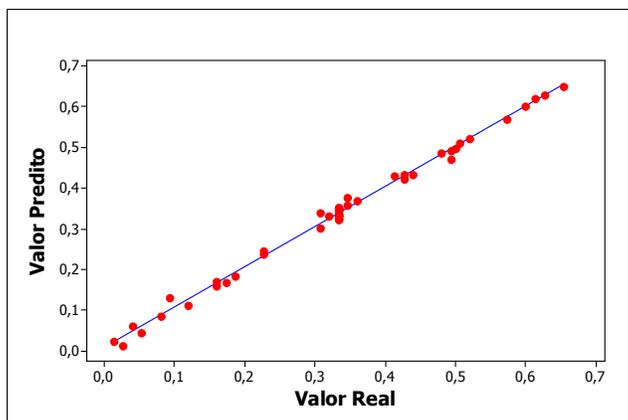
**Figura 5.18** Gráfico com o resultado de predição da fração B em amostras de calibração na associação dos intervalos 2 e 13.



**Figura 5.19** Gráfico com o resultado de predição da fração B em amostras de validacao na associação dos intervalos 2 e 13.



**Figura 5.20** Gráfico com o resultado de predição da fração B em amostras calibração na associação dos intervalos 2, 3 e 13.



**Figura 5.21** Gráfico com o resultado da predição da fração B em amostras validação na associação dos intervalos 2, 3 13.

Pode-se observar nas figuras anteriores uma boa correlação entre os valores reais *versus* valores preditos para as amostras de calibração e validação e que os valores preditos para as amostras do *set* de calibração seguem a tendência de aproximar-se da linha de tendência dos pontos de validação cruzada. Sabendo-se agora, que para os modelos testados obteve-se bons resultados, construiu-se a tabela 5.14, onde apresentam-se as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos.

**Tabela 5.14** Média e desvio padrões para os modelos obtidos na quantificação da Fração B por PLS

Regiões do espectro	Média Razão predito / real	Desvio Padrão
Intervalo 2	1,0300	0,1620
Intervalos 2 + 13	1,0420	0,2039
Intervalos 2 + 13 + 3	1,0270	0,1623

Os resultados permitem concluir que o modelo gerado pelo intervalo 2 mostrou um melhor desempenho, pois ele apresentou um menor número de variáveis latentes. É importante ressaltar que não houve variação significativa em relação ao poder de previsão nos três modelos. Verifica-se, portanto, que assim como para a quantificação da fração A, por PLS, não há necessidade de se associar intervalos para a fração B, bastando a região correspondente ao intervalo 2 (771,47 a 941,19  $\text{cm}^{-1}$ ) para isso. Da mesma maneira verifica-se, também que, para a região I não obteve-se o resultado esperado na previsão de amostras externas.

### 5.2.2.2 Aplicação do método PCR para a quantificação da fração B

Para a quantificação da fração B foram construídos modelos por PCR tanto para a região I, quanto para os intervalos e suas combinações selecionados para a construção dos modelos por mínimos quadrados parciais, conforme tabela 5.11 (pág. 60). Adotou-se os mesmos valores limites de correlação das componentes principais na quantificação da fração A para a quantificação da fração B. A tabela 5.15 traz o número de Componentes Principais correspondentes em cada região selecionada.

**Tabela 5.15** Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação da Fração B

Região	Valor Mínimo de Correlação	Nº de PCs utilizadas
Região I (1319,2 a 1546,7 $\text{cm}^{-1}$ e 2756 a 3062,7 $\text{cm}^{-1}$ )	0,5	2
	0,2	6
	0,1	8
	0,05	9
Intervalo 2 (771,47 a 941,19 $\text{cm}^{-1}$ )	0,5	2
	0,25	3
	0,1	4
	0,05	7
Intervalos 2 + 13 (771,47 a 941,19 e 2644,2 a 2812 $\text{cm}^{-1}$ )	0,5	2
	0,25	4
	0,1	5
	0,05	8
Intervalos 2 + 3 + 13 (771,47 a 941,19, 943,12 a 1112,8 e 2644,2 a 2812 $\text{cm}^{-1}$ )	0,5	2
	0,25	3
	0,1	5
	0,05	9

Os modelos foram, então, construídos com o número de componentes principais para todas as possibilidades acima para a predição das amostras externas.

5.2.2.2.1 Definição dos melhores modelos obtidos por PCR na quantificação da fração B e predição de amostras externas

Foram adotados para a quantificação da fração B os mesmos critérios adotados para a quantificação da fração A por PCR, portanto, inicialmente foi realizada a seleção dos modelos com  $R^2$  maior do que 0,90 os quais estão descritos na tabela 5.16.

**Tabela 5.16** Modelos, da fração B, que apresentaram  $R^2$  maior que 0,90

Região	Nº de PCs utilizadas	$R^2$
Região I	6	0,933
	8	0,966
	9	0,977
Intervalo 2	3	0,936
	4	0,974
	7	0,987
Intervalos 2 + 13	4	0,935
	5	0,972
	8	0,989
Intervalos 2 + 3 + 13	3	0,936
	5	0,968
	9	0,988

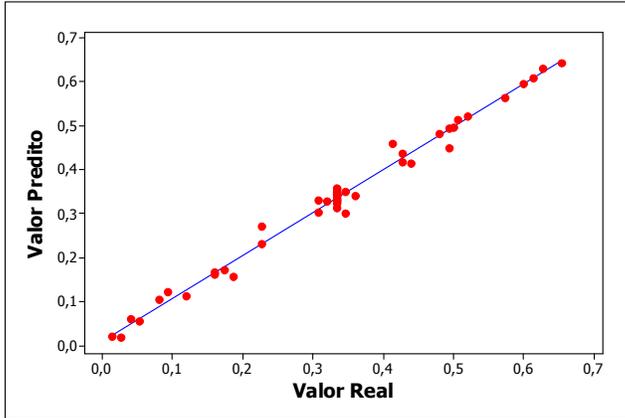
Selecionou-se, então, dentre os modelos descritos na tabela 4.16, aqueles com valores de RMSEP menores que 0,02 conforme tabela 5.17.

**Tabela 5.17** Modelos, da fração B, que apresentaram  $R^2$  maior que 0,90 e RMSEP menor que 0,020 por PCR

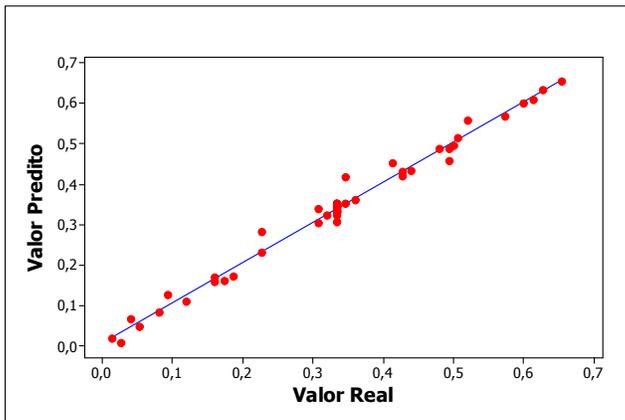
Região	Nº de PCs utilizadas	RMSEP
Intervalo 2	7	0,0181
Intervalos 2 + 13	8	0,0197
Intervalos 2 + 13 + 3	9	0,0166

Os modelos representados na tabela 5.17 foram, então, os selecionados para expressar os resultados para a predição de amostras. Para se conferir esses

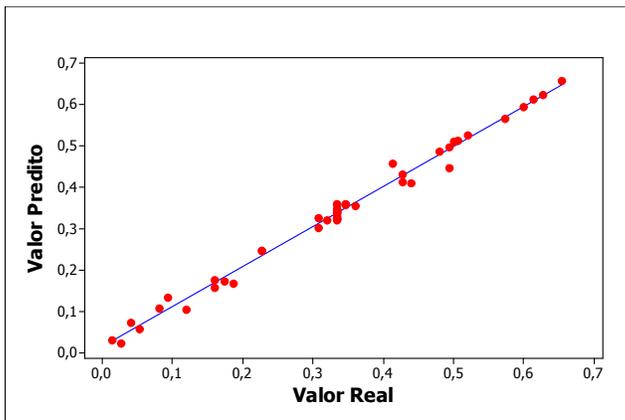
resultados de previsão, as figuras 5.22, 5.23 e 5.24 mostram os correspondentes gráficos relacionando os valores preditos e os valores reais para os modelos escolhidos.



**Figura 5.22** Gráfico com o resultado de previsão da fração B em amostras externas no intervalo 2 por PCR.



**Figura 5.23** Gráfico com o resultado de previsão da fração B em amostras externas na associação dos intervalos 2 e 13 por PCR.



**Figura 5.24** Gráfico com o resultado de previsão da fração B em amostras externas na associação dos intervalos 2, 3 e 13 por PCR.

Observa-se que para as 3 regiões selecionadas as curvas obtidas com os valores reais *versus* valores preditos para as amostras de validação mostram boa

correlação. A tabela 5.18 traz, então, as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos.

**Tabela 5.18** Média e desvio padrões para os modelos obtidos na quantificação da Fração B por PCR

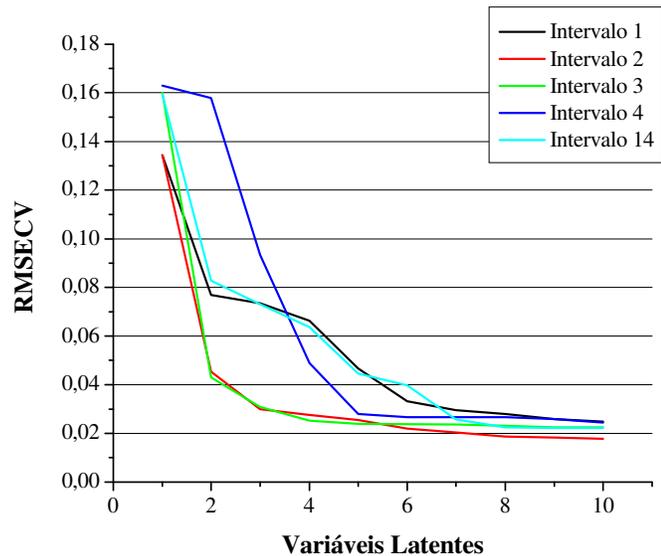
<b>Regiões do espectro</b>	<b>Média Razão predito / real</b>	<b>Desvio Padrão</b>
Intervalo 2	1,0352	0,1459
Intervalos 2 + 13	1,0317	0,1669
Intervalos 2 + 13 + 3	1,0609	0,2267

Com esses resultados observa-se que, assim como na construção dos modelos por PLS, os modelos obtidos por PCR nas três regiões obtiveram um desempenho bastante semelhante com poder de previsão satisfatórios. Com isso, verifica-se que para a quantificação da fração B por PCR também não há necessidade de se associar intervalos, bastando a região correspondente ao intervalo 2 para isso. A região I, igualmente ao ocorrido na modelagem por PLS, não mostrou resultados satisfatórios. Os modelos obtidos por PLS apresentaram desempenhos ligeiramente melhores que os obtidos por PCR.

### **5.2.3 Quantificação da fração C**

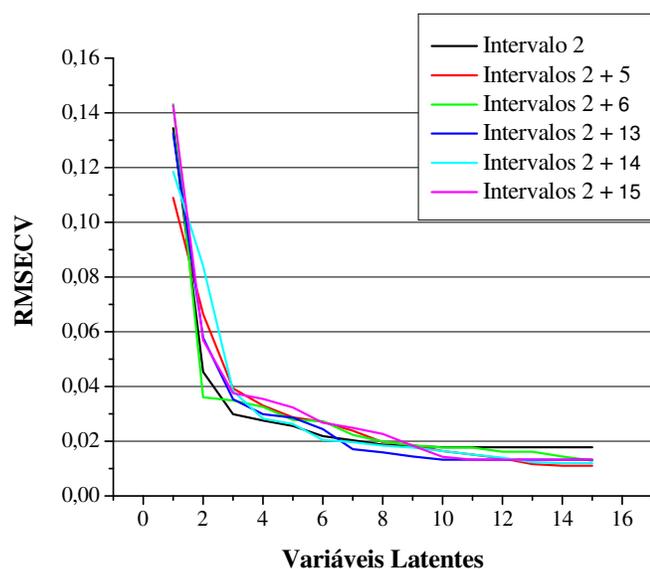
#### **5.2.3.1 Aplicação do método PLS para a quantificação da fração C**

Os mesmos procedimentos e parâmetros adotados para a quantificação das frações A e B descritos anteriormente foram utilizados na quantificação da fração C sendo assim, a figura 4.25 traz os perfis do RMSECV para os 5 intervalos que apresentaram o melhor desempenho quanto ao RMSECV na aplicação do método PLS.



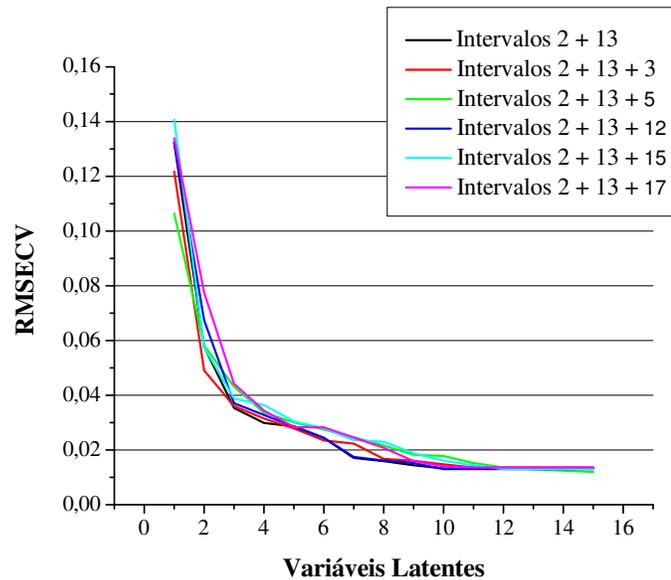
**Figura 5.25** Perfil do RMSECV para a fração C com modelos construídos para as cinco melhores regiões.

Como se pode observar, assim como para as frações A e B, o intervalo 2 resultou no melhor modelo e por isso, da mesma forma como para as outras frações, foi mantido fixo e adicionado a ele os outros intervalos em separado. A figura 5.26 mostra os valores de RMSECV para os 5 melhores modelos obtidos combinando-se o intervalo 2 com os demais.



**Figura 5.26** Perfil do RMSECV da fração C para a associação de duas regiões para as 5 melhores combinações

Pela figura observa-se que a combinação do intervalo 2 com o intervalo 13 obteve o menor valor de RMSECV e, por isso, decidiu-se manter esses dois intervalos fixos e adicionar a eles uma terceira região.



**Figura 5.27** Perfil do RMSECV da fração C para a associação de três regiões para as 5 melhores combinações

De acordo com o observado na figura 5.27, conclui-se que não houve nenhuma melhora com a combinação de 3 intervalos. Por isso, não prosseguiu-se mais com as associações e, nesse caso, esse resultado não foi levado em consideração.

#### 5.2.3.1.1 Definição dos melhores modelos obtidos por PLS na quantificação da fração C e predição de amostras externas

Foram, então, plotados gráficos relacionando os valores de RMSECV, RMSEC e RMSEP para as regiões selecionadas. A tabela 5.19 mostra os resultados desses gráficos, ou seja, mostra quais são as variáveis latentes necessárias para o melhor desempenho de cada modelo.

**Tabela 5.19** Seleção do número de variáveis latentes para cada modelo proposto para a quantificação da fração C

<b>Regiões do espectro</b>	<b>Variáveis Latentes</b>
Intervalo 2	7
	10
Intervalo 2 + 13	10

Os modelos foram, então, construídos com o número de variáveis latentes para todas as possibilidades acima para a predição das amostras externas. A tabela 5.20 traz os valores de  $R^2$  e RMSEP para esses modelos construídos.

**Tabela 5.20** Valores de  $R^2$  e RMSEP para os modelos construídos para a quantificação da fração C

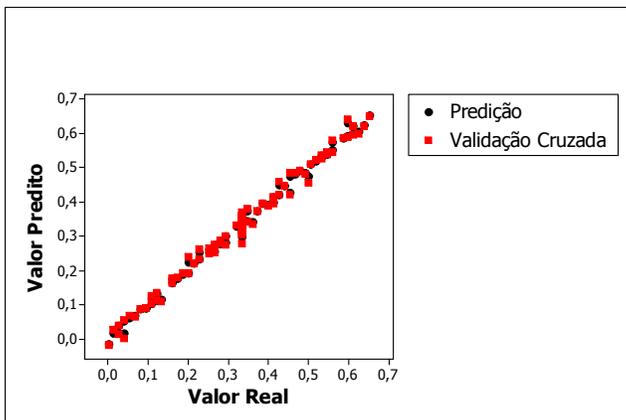
<b>Regiões do espectro</b>	<b>VL</b>	<b><math>R^2</math></b>	<b>RMSEP</b>
Intervalo 2	7	0,9899	0,0184
	10	0,9936	0,0159
Intervalo 2 + 13	10	0,9966	0,0138

Selecionou-se então, dentre os modelos restantes para o intervalo 2, aquele com menor valor de RMSEP. A tabela 5.21 traz, portanto, os modelos selecionados para a quantificação da fração C por PLS.

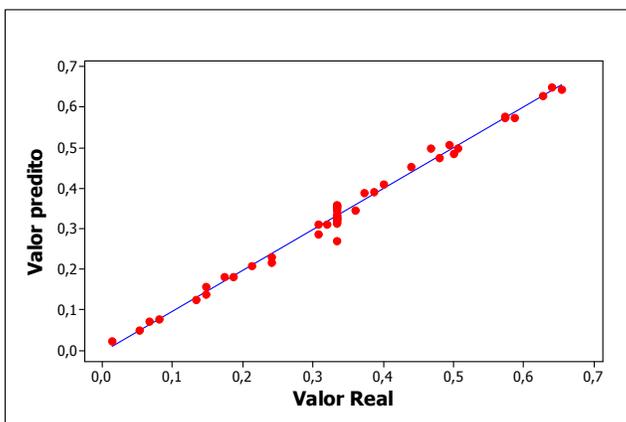
**Tabela 5.21** Modelos selecionados para a quantificação da fração C por PLS

<b>Regiões do espectro</b>	<b>VL</b>	<b><math>R^2</math></b>	<b>RMSEP</b>
Intervalo 2	10	0,9936	0,0159
Intervalo 2 + 13	10	0,9966	0,0138

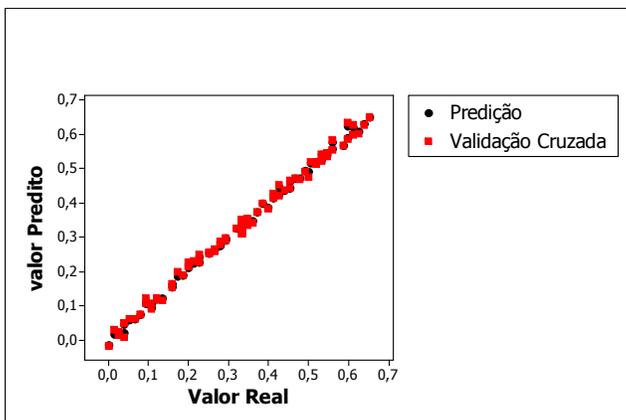
Os gráficos contendo os valores reais versus o valor predito pelos modelos foram desenhados para se conferir o resultado de previsão para amostras de calibração e previsão nos modelos selecionados. As figuras 5.28 a 5.31 mostram os correspondentes gráficos para os modelos escolhidos.



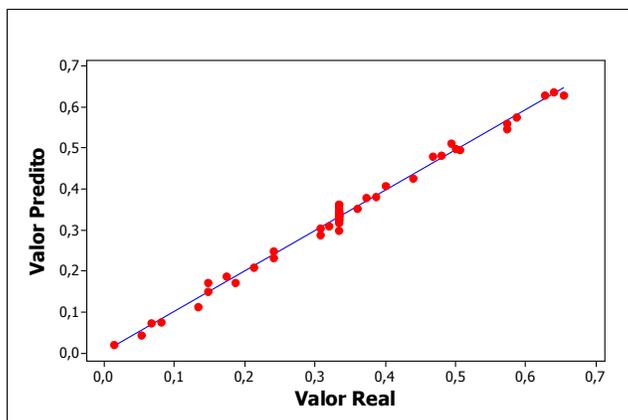
**Figura 5.28** Gráfico com o resultado de predição da fração C em amostras de calibração no intervalo 2.



**Figura 5.29** Gráfico com o resultado de predição da fração C em amostras de validação no intervalo 2.



**Figura 5.30** Gráfico com o resultado de predição da fração C em amostras de calibração na associação dos intervalos 2 e 13.



**Figura 5.31** Gráfico com o resultado de predição da fração C em amostras de validação na associação dos intervalos 2 e 13.

Pode-se observar que as curvas obtidas com os valores reais *versus* valores preditos para as amostras de calibração e validação mostram boa correlação. Observa-se, também, que os valores preditos para as amostras do *set* de validação seguem a tendência de aproximar-se da linha de tendência dos pontos de validação cruzada. A tabela 5.22 apresenta as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos.

**Tabela 5.22** Média e desvio padrões para os modelos obtidos na quantificação da Fração C por PLS

Regiões do espectro	Média Razão predito / real	Desvio Padrão
Intervalo 2	1,0041	0,1093
Intervalos 2 + 13	1,0028	0,0923

Os resultados mostram que os modelos construídos nas duas regiões obtiveram um desempenho bastante semelhante, sendo o poder de previsão de ambos bastante satisfatórios. Com isso, conclui-se que para a quantificação da fração C por PLS também não há necessidade de se associar intervalos, bastando a região correspondente ao intervalo 2 para isso. Com esses resultados verifica-se também que mais uma vez a região I não proporcionou o resultado esperado na previsão de amostras externas.

### 5.2.3.2 Aplicação do método PCR para a quantificação da fração C

Para a quantificação da fração C por PCR foram construídos modelos tanto para a região I quanto para os intervalos e suas combinações selecionados para a construção dos modelos por mínimos quadrados parciais conforme tabela 5.19 (pág. 70).

Utilizou-se, para a fração C, os mesmo valores limites de correlação das componentes principais utilizados na quantificação das frações A e B. A tabela 5.23 traz o número de Componentes Principais correspondentes em cada região selecionada.

**Tabela 5.23** Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação da Fração C

Região	Valor Mínimo de Correlação	Nº de PCs utilizadas
Região I (1319,2 a 1546,7 $\text{cm}^{-1}$ e 2756 a 3062,7 $\text{cm}^{-1}$ )	0,5	1
	0,2	3
	0,1	7
	0,05	9
Intervalo 2 (771,47 a 941,19 $\text{cm}^{-1}$ )	0,5	1
	0,2	3
	0,05	8
Intervalos 2 + 13 (771,47 a 941,19 $\text{cm}^{-1}$ e 2644,2 a 2812 $\text{cm}^{-1}$ )	0,5	1
	0,2	3
	0,1	4
	0,05	9

Os modelos foram, então, construídos com o número de componentes principais para todas as possibilidades acima para a predição das amostras de validação.

#### *5.2.3.2.1 Definição dos melhores modelos obtidos por PCR na quantificação da fração C e predição de amostras externas*

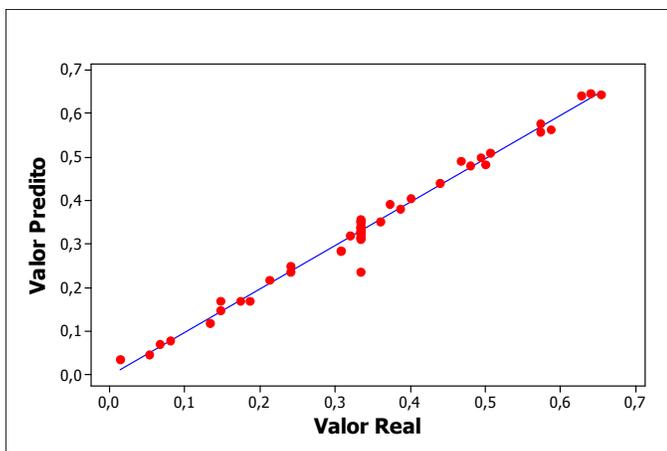
Para a definição dos melhores modelos para a predição de amostras externas foram adotados para a quantificação da fração C os mesmo critérios adotados para a

quantificação das frações A e B por PCR, portanto, inicialmente foi realizada a seleção dos modelos com  $R^2$  maior do que 0,90 os quais estão descritos na tabela 5.24.

**Tabela 5.24** Modelos, da fração C, que apresentaram  $R^2$  maior que 0,90

Região	Nº de PCs utilizadas	$R^2$
Região I	7	0,975
	9	0,986
Intervalo 2	3	0,972
	8	0,986
Intervalos 2 + 13	3	0,946
	4	0,966
	9	0,992

Posteriormente foram selecionados, dentre os modelos descritos na tabela 5.24, aqueles com valores de RMSEP menor que 0,02, porém apenas o intervalo 2, utilizando 8 PCs, apresentou um RMSEP de 0,0194. Todos os outros modelos apresentaram valores de RMSEP maior que 0,02 e por isso foram descartados. Para se conferir o resultado de previsão deste modelo selecionado, a figura 5.32 mostra o correspondente gráfico relacionando os valores preditos e os valores reais para o modelo escolhido e a tabela 5.25 apresenta a média e o desvio padrão obtido da razão dos valores preditos/reais para este modelo.



**Figura 5.32** Gráfico com o resultado de previsão da fração C em amostras externas no intervalo 2 por PCR.

**Tabela 5.25** Média e desvio padrões para o modelo obtido na quantificação da Fração C por PCR

<b>Regiões do espectro</b>	<b>Média Razão predito / real</b>	<b>Desvio Padrão</b>
Intervalo 2	1,0609	0,2267

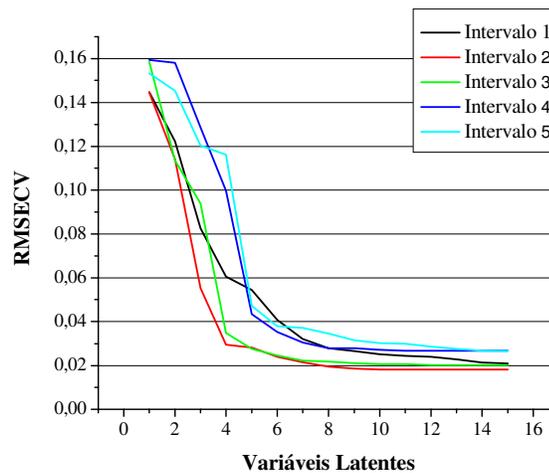
A partir destes resultados observa-se que, diferentemente da modelagem por PLS, a associação de intervalos não levou a resultados satisfatórios nem mesmo próximos aos obtidos pelo intervalo 2. A região I, igualmente ao ocorrido na modelagem por PLS, não proporcionou bons resultados. Os modelos obtidos por PLS apresentaram desempenhos ligeiramente melhores que os obtidos por PCR.

#### **5.2.4 Quantificação das frações A, B e C simultaneamente**

##### 5.2.4.1 Aplicação do método PLS para a quantificação das frações A, B e C

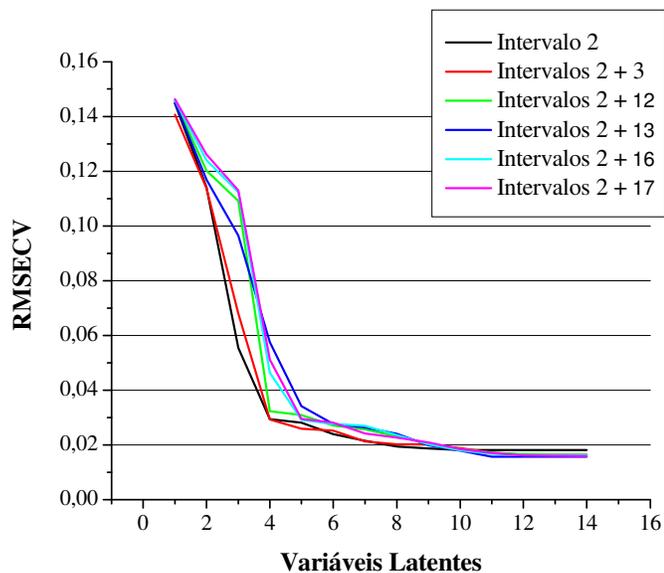
Após ter sido realizada a quantificação das frações A, B e C individualmente, com resultados bastante satisfatórios, partiu-se para a construção de um modelo global, ou seja, um modelo para se quantificar as 3 frações ao mesmo tempo. Adotou-se para isso o mesmo procedimento utilizado na quantificação das frações individualmente. Porém, devido aos resultados insatisfatórios da região I (1319,2 a 1546,7 e 2756 a 3062,7  $\text{cm}^{-1}$ ) observados na quantificação das 3 frações separadamente, ela não foi mais testada nem para quantificação das frações A, B e C simultaneamente nem para a determinação da densidade e da viscosidade que serão vistas mais adiante.

Portanto, para a quantificação das frações A, B e C, inicialmente foram calculados os valores de RMSECV e VL para cada um dos 20 intervalos em que o espectro foi dividido (tabela 4.5, pág. 41). A figura 5.33 apresenta o perfil do RMSECV para as 5 melhores regiões.



**Figura 5.33** Perfil do RMSECV para as frações A, B e C com modelos construídos para as cinco melhores regiões.

Verifica-se na figura 5.33 que assim como para a quantificação das frações A, B e C separadamente, o intervalo 2 mostrou o melhor desempenho e, por isso, ele foi mantido fixo e adicionado a ele cada um dos outros 19 intervalos.



**Figura 5.34** Perfil do RMSECV para as frações A, B e C com modelos construídos para a associação de 2 regiões para as cinco melhores regiões.

De acordo com o observado na figura 5.34, conclui-se que não houve melhora alguma do valor de RMSECV com a associação de 2 intervalos, apenas com grande

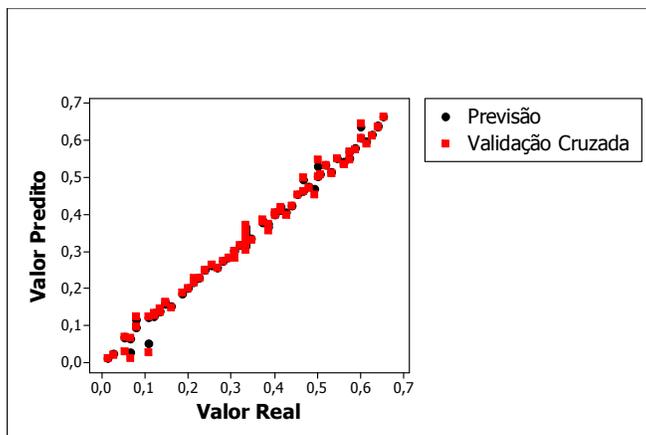
quantidade de variáveis latentes houve uma ligeira melhora. Desta maneira, esse resultado não foi considerado na construção dos modelos.

#### 5.2.4.1.1 Definição dos melhores modelos obtidos por PLS na quantificação das frações A, B e C e predição de amostras externas

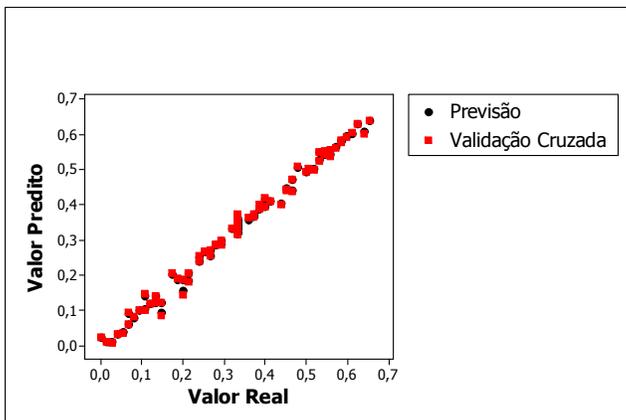
Foi, então, feito o gráfico relacionando os valores de RMSECV, RMSEC e RMSEP para o intervalo 2 (região selecionada) e a partir dele verificou-se que o melhor desempenho, com valores de RMSECV, RMSEC e RMSEP baixos e próximos, ocorreu com 8 VLs apenas. A tabela 5.26 mostra os valores de  $R^2$  e RMSEP para este modelo e as figuras 5.35 a 5.40, os gráficos contendo os valores reais versus o valor predito plotados, para cada uma das frações.

**Tabela 5.26** Valores de  $R^2$  e RMSEP para o modelo obtido para a quantificação das frações A, B e C por PLS

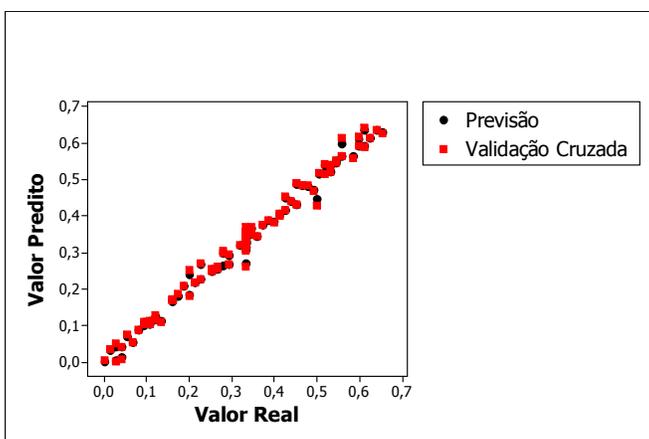
Região	VL	$R^2$			RMSEP		
		A	B	C	A	B	C
Intervalo 2	8	0,9911	0,9895	0,9891	0,0147	0,0148	0,0183



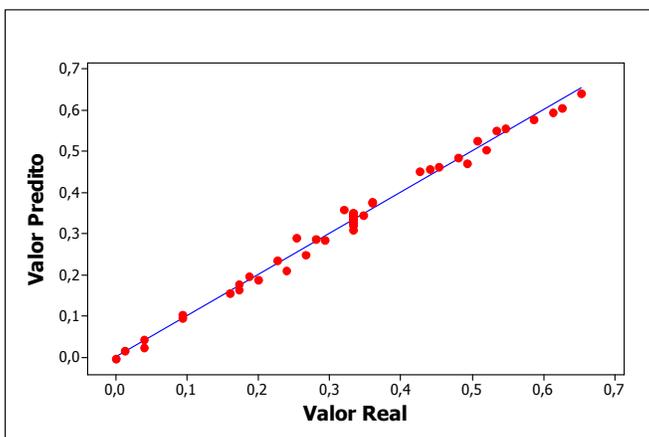
**Figura 5.35** Gráfico com o resultado de predição da fração A, para o modelo global, em amostras de calibração no intervalo 2.



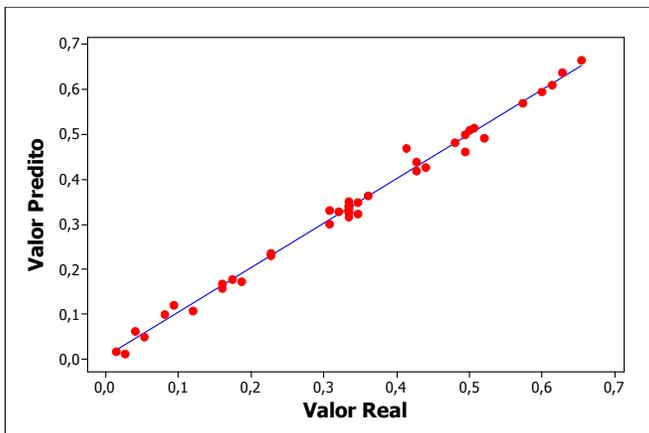
**Figura 5.36** Gráfico com o resultado de predição da fração B, para o modelo global, em amostras de calibração no intervalo 2.



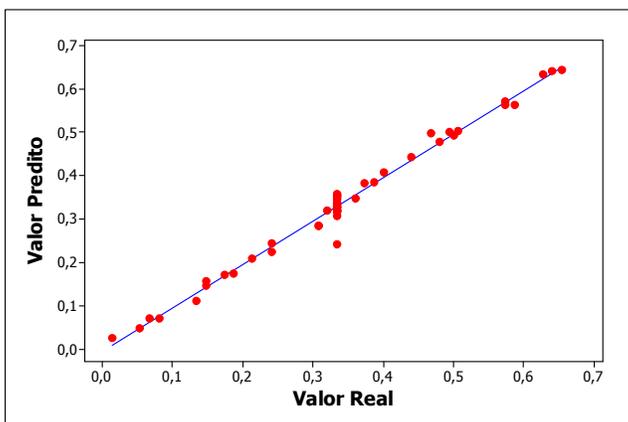
**Figura 5.37** Gráfico com o resultado de predição da fração C, para o modelo global, em amostras de calibração no intervalo 2.



**Figura 5.38** Gráfico com o resultado de predição da fração A, para o modelo global, em amostras de validação no intervalo 2.



**Figura 5.39** Gráfico com o resultado de predição da fração B, para o modelo global, em amostras de validação no intervalo 2.



**Figura 5.40** Gráfico com o resultado de predição da fração C, para o modelo global, em amostras de validação no intervalo 2.

Pode-se observar que as curvas obtidas com os valores reais *versus* valores preditos para as amostras de calibração e validação mostram boa correlação sendo assim, construiu-se a tabela 5.27, onde são apresentadas as médias e os desvios padrões obtidos da razão dos valores preditos/reais para o modelo selecionado.

**Tabela 5.27** Média e desvio padrões para o modelo obtido na quantificação das frações A, B e C por PLS

Região	Média Razão predito / real			Desvio Padrão		
	A	B	C	A	B	C
Intervalo 2	0,9790	1,0208	1,0053	0,1620	0,1422	0,1537

Com os resultados apresentados até esse ponto, observa-se que o intervalo 2 é a região de escolha para a quantificação das frações de diesel por PLS nas misturas tanto individual quanto simultaneamente. Para o modelo global a associação de

intervalos não forneceu resultados de previsão satisfatórios. Quanto ao poder de previsão do modelo global comparado aos modelos individuais, esses últimos apresentaram um desempenho ligeiramente melhor, porém bastante parecidos com o modelo global. Sendo assim, concluiu-se que é possível determinar as quantidades de frações de diesel individual e simultaneamente por PLS.

#### 5.2.4.2 Aplicação do método PCR para a quantificação das frações A, B e C

Para a construção de modelos por PCR na quantificação simultânea das frações A, B e C foi utilizado apenas o intervalo 2, visto ser essa única região selecionada na modelagem por PLS. Nessa modelagem utilizou-se os mesmos resultados obtidos na quantificação das frações individualmente, porém foram calculados os valores médios de RMSEP. Para a seleção das componentes principais utilizou-se os mesmos valores limites de correlação utilizados para a quantificação das frações A, B e C individualmente. A tabela 5.28 traz o número de PCs correspondentes em cada região selecionada e os valores mínimos de correlação entre as componentes principais e as concentrações de A, B e C.

**Tabela 5.28** Número de PCs utilizadas na construção dos modelos, por PCR, na quantificação das frações A, B e C

Região	Valor Mínimo de Correlação	Nº de PCs utilizadas
Intervalo 2 (771,47 a 941,19 cm <sup>-1</sup> )	0,5	2
	0,2	4
	0,1	5
	0,05	8

Os modelos foram, então, construídos com o número de componentes principais para todas as possibilidades acima visando a predição das amostras de validação.

4.2.4.2.1 Definição dos melhores modelos obtidos por PCR na quantificação simultânea das frações A, B e C e predição de amostras externas

Para a definição dos melhores modelos e a predição de amostras externas, adotou-se os mesmos critérios adotados para a quantificação das frações A, B e C individualmente. Portanto, dentre os modelos gerados, foram selecionados primeiramente aqueles com valores de  $R^2$  maior que 0,90 conforme tabela 5.29.

**Tabela 5.29** Modelos, das frações A B e C, que apresentaram  $R^2$  maior que 0,90 por PCR

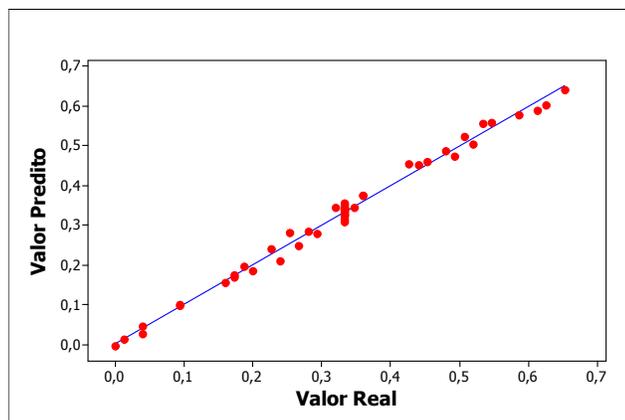
Região	Nº de PCs utilizadas	$R^2$		
		A	B	C
Intervalo 2	4	0,955	0,974	0,972
	5	0,974	0,978	0,977
	8	0,991	0,989	0,989

A partir dos modelos descritos na tabela acima, foi feita uma nova seleção: excluiu-se aqueles com valores de RMSEP maior que 0,020. Conforme pode ser observado na tabela 5.30, apenas o modelo construído no intervalo 2 utilizando 8 PCs resultou em um modelo com RMSEP menor que 0,020. Este modelo foi, então, o selecionado para a quantificação das frações A, B e C por PCR.

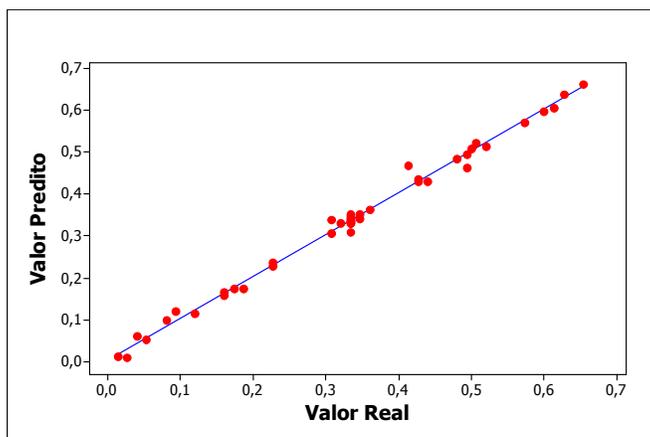
**Tabela 5.30** Modelos, das frações A, B e C, que apresentaram  $R^2$  maior que 0,90 e RMSEP menor que 0,02 por PCR

Região	Nº de PCs utilizadas	RMSEP		
		A	B	C
Intervalo 2	8	0,0147	0,0140	0,0185

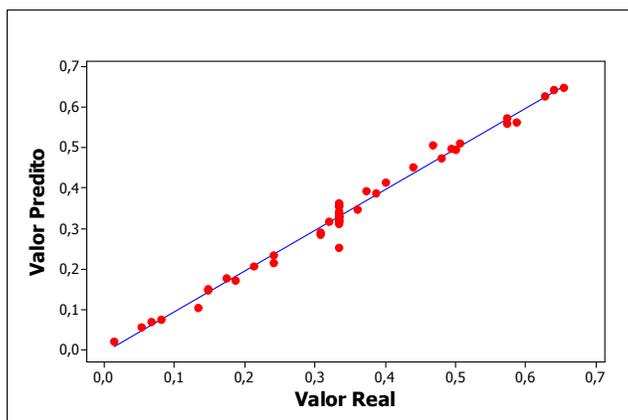
Para uma melhor visualização dos resultados de predição obtidos pelo modelo selecionado, as figuras 5.41 a 5.43 mostram a relação entre os valores preditos e os valores reais para as frações A, B e C respectivamente.



**Figura 5.41** Gráfico com o resultado de predição da fração A, para o modelo global, em amostras de validação no intervalo 2 por PCR.



**Figura 5.42** Gráfico com o resultado de predição da fração B, para o modelo global, em amostras de validação no intervalo 2 por PCR.



**Figura 5.43** Gráfico com o resultado de predição da fração C, para o modelo global, em amostras de validação no intervalo 2 por PCR.

Pode-se observar nas figuras acima, uma boa relação entre valores preditos e valores reais para as três frações. Posteriormente foram calculados as médias e os

desvios padrões obtidos da razão dos valores preditos/reais para os modelos, conforme tabela 5.31.

**Tabela 5.31** Média e desvio padrões para o modelo obtidos na quantificação das Frações A, B e C por PCR

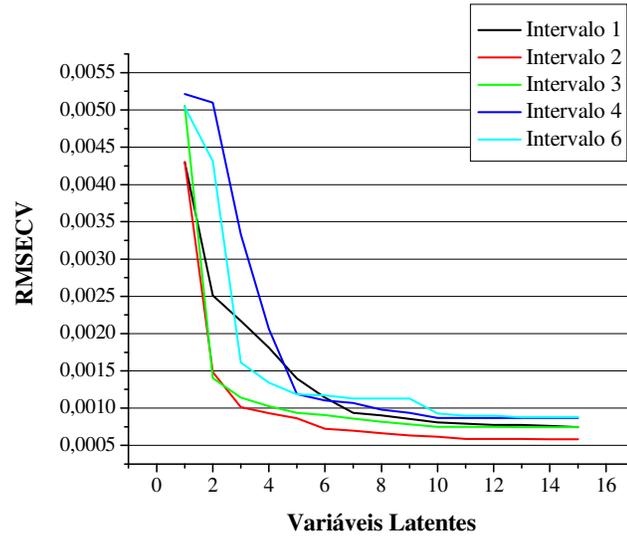
Região	Média Razão predito / real			Desvio Padrão		
	A	B	C	A	B	C
Intervalo 2	0,9775	1,0088	0,9973	0,1587	0,1426	0,1049

A partir da análise dos resultados observa-se que o modelo obtido por PCR na quantificação simultânea das frações A, B e C apresentou um desempenho praticamente igual ao obtido por PLS. Conclui-se, portanto, que o intervalo 2 é a região de escolha para a quantificação das frações de diesel não só por PLS quanto por PCR individual e simultaneamente. Ambos os métodos apresentaram desempenhos bastante parecidos, sendo que no geral o PLS mostrou-se ligeiramente melhor.

## 5.2.5 Predição da densidade

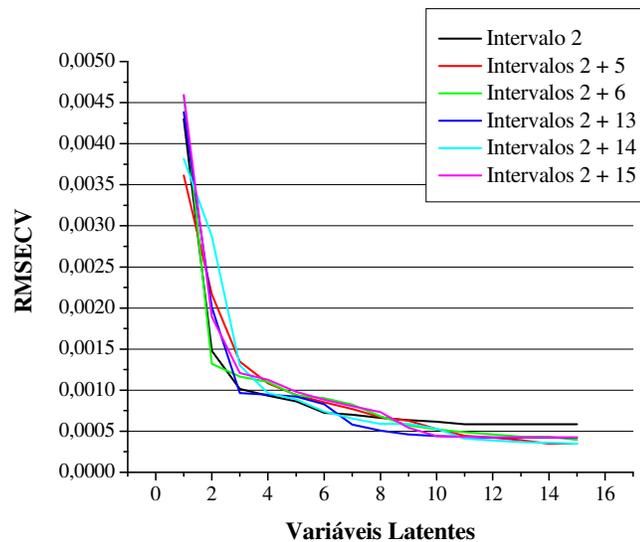
### 5.2.5.1 Aplicação do método PLS para a predição da densidade

Após ter sido realizada a quantificação das frações A, B e C individual e simultaneamente por PLS e PCR, partiu-se para a construção de modelos para se estimar a densidade e da viscosidade das amostras utilizando-se, também, os mesmos critérios. Sendo assim, os perfis do RMSECV para os 5 intervalos que indicaram o melhor desempenho quanto ao RMSECV na aplicação do PLS estão representados na figura 5.44.



**Figura 5.44** Perfil do RMSECV para a densidade com modelos construídos para as cinco melhores regiões

Observa-se que o intervalo 2, assim como para a quantificação das frações A, B e C individual e simultaneamente, indicou o melhor desempenho e, por isso, este foi mantido fixo e adicionado a ele cada intervalo em separado. A figura 5.45 mostra as 5 melhores combinações.



**Figura 5.45** Perfil do RMSECV para a densidade para a associação de duas regiões para as 5 melhores combinações

O melhor desempenho observado na figura 5.45 foi com a associação dos intervalos 2 e 13, porém resolveu-se não prosseguir com a associação dos intervalos visto que a diminuição do RMSECV não foi tão significativa. Mesmo assim, para os estudos seguintes, esse resultado foi levado em consideração.

#### 5.2.5.1.1 Definição dos melhores modelos obtidos por PLS na determinação da densidade e predição de amostras externas

Depois de todas as combinações propostas acima, verificou-se quais seriam as variáveis latentes necessárias para o melhor desempenho de cada modelo a partir dos gráficos relacionando os valores de RMSECV, RMSEC e RMSEP. A tabela 5.32 mostra os resultados destes gráficos construídos.

**Tabela 5.32** Seleção do número de variáveis latentes para cada modelo proposto para a predição da densidade

Regiões do espectro	Variáveis Latentes
Intervalo 2	3
	7
Intervalo 2 + 13	5
	6

Os modelos foram, então, construídos com o número de variáveis latentes para todas as possibilidades acima a fim de se verificar a habilidade na predição da densidade de amostras não presentes no modelo. A tabela 5.33 mostra os valores de  $R^2$  e RMSEP para esses modelos construídos.

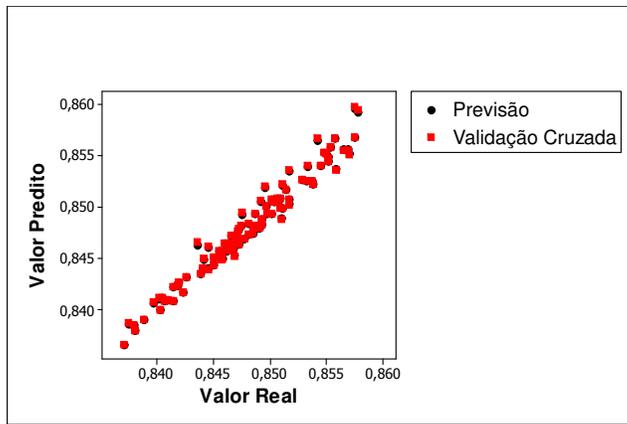
**Tabela 5.33** Valores de  $R^2$  e RMSEP para os modelos construídos para a predição da densidade

Regiões do espectro	VL	$R^2$	RMSEP
Intervalo 2	3	0,9652	0,00099
	7	0,9877	0,00052
Intervalo 2 + 13	5	0,9721	0,00078
	6	0,9787	0,00069

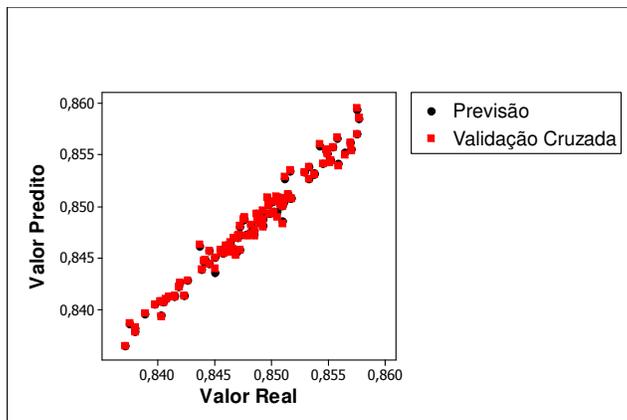
Diferentemente dos modelos construídos para a quantificação das frações, a seleção dentre os modelos para a densidade não foi feita pelo menor valor de RMSEP e sim pelo menor número de VL em cada região, visto que os valores de RMSEP já se encontravam bastante baixos (menores que 0,001). A tabela 5.34 mostra, portanto, os modelos selecionados para se estimar a densidade por PLS e as figuras 5.46 a 5.49 os correspondentes gráficos relacionando os valores preditos e os valores reais para as amostras de calibração e validação.

**Tabela 5.34** Modelos selecionados para a predição da densidade por PLS

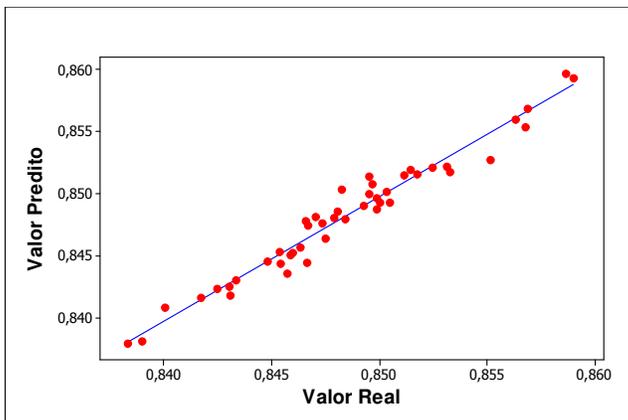
Regiões do espectro	VL	R <sup>2</sup>	RMSEP
Intervalo 2	3	0,9652	0,00099
Intervalo 2 + 13	5	0,9721	0,00078



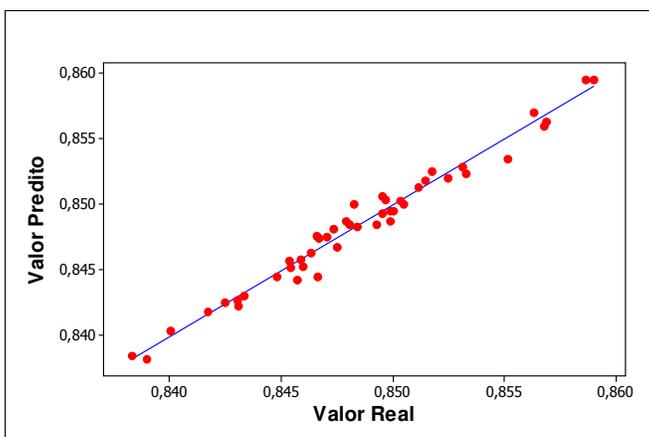
**Figura 5.46** Gráfico com o resultado de predição da densidade em amostras de calibração no intervalo 2.



**Figura 5.47** Gráfico com o resultado de predição da densidade em amostras de calibração na associação dos intervalos 2 e 13.



**Figura 5.48** Gráfico com o resultado de predição da densidade em amostras de validação no intervalo 2.



**Figura 5.49** Gráfico com o resultado de predição da densidade em amostras de validação na associação dos intervalos 2 e 13.

As figuras mostram que a relação linear entre as medidas reais e preditas para as amostras de calibração e validação são satisfatórias. A tabela 5.35 mostra as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos.

**Tabela 5.35** Média e desvio padrões para os modelos obtidos na predição da densidade por PLS

Regiões do espectro	Média Razão predito / real	Desvio Padrão
Intervalo 2	0,9997	0,0012
Intervalos 2 + 13	0,9998	0,0009

Diante dos resultados obtidos pode-se concluir que o modelo iPLS pode ser utilizado para prever valores de densidade de frações de diesel de maneira bastante satisfatória não necessitando da aplicação do siPLS, pois ambos forneceram resultados bastante semelhantes.

### 5.2.5.2 Aplicação do método PCR para a predição da densidade

Os mesmos critérios adotados para a quantificação das frações por PCR foram utilizados para se estimar a densidade. Portanto, a quantidade de PCs utilizadas na construção de cada modelo de acordo com os valores mínimos de correlação entre a densidade e as componentes principais estão descritos na tabela 5.36.

**Tabela 5.36** Número de PCs utilizadas na construção dos modelos, por PCR, na predição da densidade

Região	Valor Mínimo de Correlação	Nº de PCs utilizadas
Intervalo 2 (771,47 a 941,19 cm <sup>-1</sup> )	0,5	1
	0,2	2
	0,1	4
	0,05	7
Intervalos 2 + 13 (771,47 a 941,19 e 2644,2 a 2812 cm <sup>-1</sup> )	0,5	1
	0,2	4
	0,05	7

Os modelos foram, então, construídos com o número de componentes principais para todas as possibilidades acima para a predição das amostras externas.

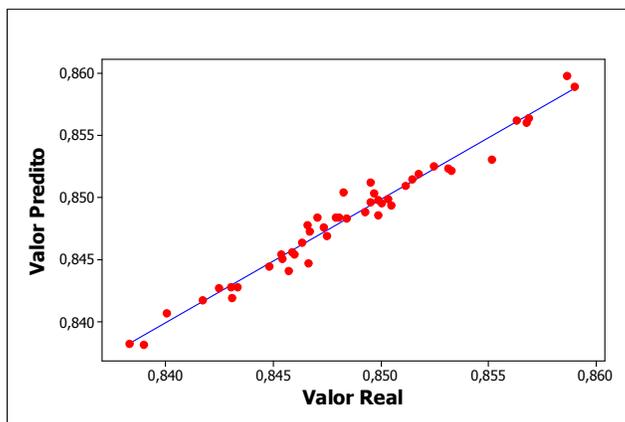
#### *5.2.5.2.1 Definição dos melhores modelos obtidos por PCR na determinação da densidade e predição de amostras externas*

Dentre os modelos construídos, primeiramente foram selecionados aqueles com valor de R<sup>2</sup> maior que 0,90 conforme tabela 5.37.

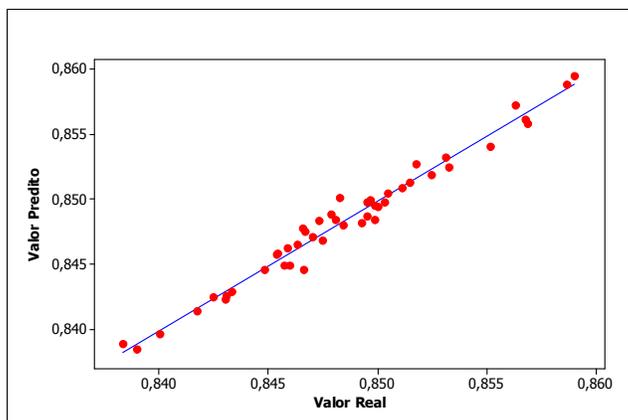
**Tabela 5.37** Modelos que apresentaram  $R^2$  maior que 0,90, por PCR, na predição da densidade

Região	Nº de PCs utilizadas	$R^2$	RMSEP
Intervalo 2	2	0,907	0,00135
	4	0,970	0,00086
	7	0,981	0,00089
Intervalos 2 + 13	4	0,967	0,00077
	7	0,979	0,00073

Verifica-se nessa tabela que os valores de RMSEP foram bastante abaixo de 0,02. Portanto, selecionou-se para as duas regiões os modelos construídos com 4 PCs, por apresentarem valores de RMSEP menor que 0,001 e possuírem diferença quase insignificante para os modelos construídos com 7 PCs. As figuras 5.50 e 5.51 mostram os correspondentes gráficos para esses dois modelos relacionando os valores preditos e reais da densidade.



**Figura 5.50** Gráfico com o resultado de predição da densidade em amostras de validação no intervalo 2 por PCR.



**Figura 5.51** Gráfico com o resultado de predição da densidade em amostras de validação na associação dos intervalos 2 e 13 por PCR.

A tabela 5.38 mostra, portanto, as médias e os desvios padrões obtidos da razão dos valores preditos/reais para os modelos selecionados.

**Tabela 5.38** Média e desvio padrões para o modelo obtido na predição densidade por PCR

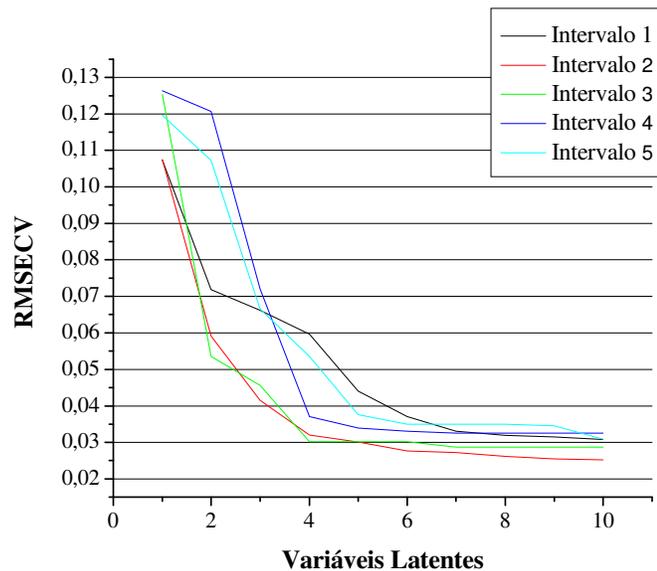
<b>Regiões do espectro</b>	<b>Média Razão predito / real</b>	<b>Desvio Padrão</b>
Intervalo 2	0,9999	0,0010
Intervalos 2 + 13	0,9998	0,0009

Diante dos resultados obtidos pode-se concluir que o modelo PCR pode ser utilizado para estimar valores de densidade de frações de diesel de maneira bastante satisfatória na região do intervalo 2 ( $771,47$  a  $941,19 \text{ cm}^{-1}$ ) não necessitando da associação de intervalos, pois ambas obtiveram resultados bastante parecidos.

## 5.2.6 Predição da viscosidade

### 5.2.6.1 Aplicação do método PLS para a determinação da viscosidade

Assim como foi feito na construção dos modelos por mínimos quadrados parciais descritos anteriormente, para se estimar a viscosidade inicialmente foram calculados os valores de RMSECV para os 20 intervalos em que o espectro foi dividido. Sendo assim, os perfis do RMSECV para os 5 intervalos que obtiveram o melhor desempenho estão representados na figura 5.52.

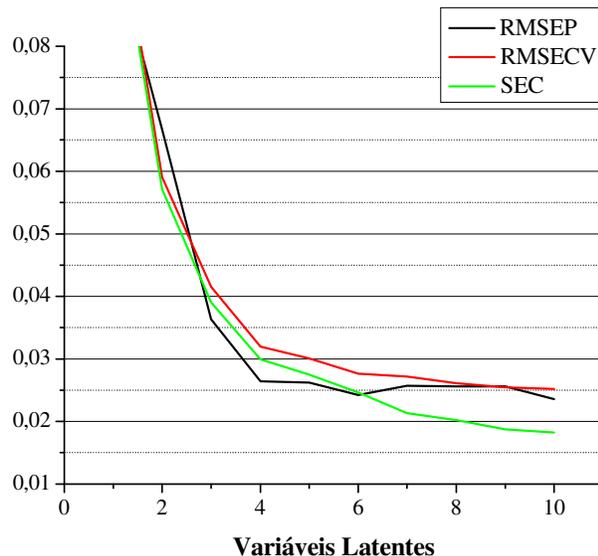


**Figura 5.52** Perfil do RMSECV para a viscosidade com modelos construídos para as cinco melhores regiões.

Como se pode observar na figura acima, o intervalo 2 mais uma vez apresentou o melhor desempenho e por isso ele foi mantido fixo e os outros intervalos foram adicionados a ele. Essas associações foram analisadas quanto aos valores de RMSECV, porém nenhuma delas forneceu um resultado melhor do que o intervalo 2 sozinho e portanto, foram descartadas.

#### *5.2.6.1.1 Definição dos melhores modelos obtidos por PLS na determinação da densidade e predição de amostras externas*

Foi, então, feito o gráfico relacionando os valores de RMSECV, RMSEC e RMSEP para o intervalo 2 para se estimar a viscosidade.

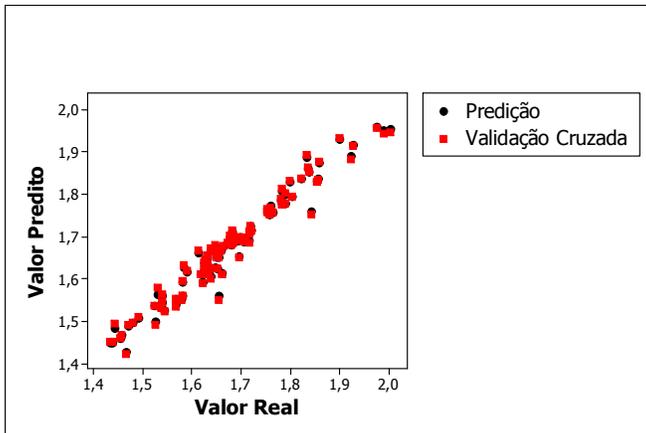


**Figura 5.53** Perfil do RMSECV, RMSEC e RMSEP variando-se o número de variáveis latentes para a viscosidade no intervalo 2.

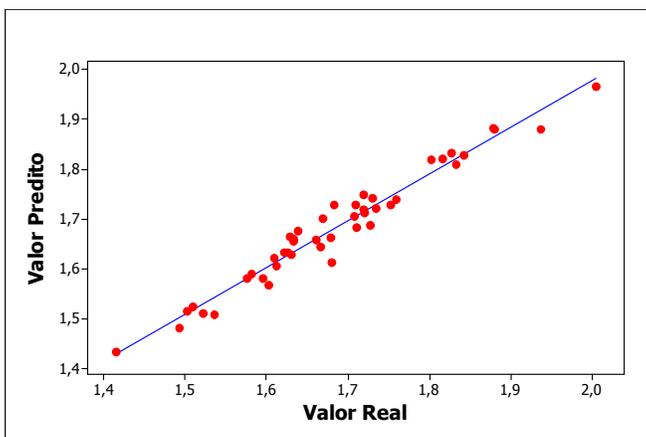
Analisando-se a figura 5.53 verifica-se que todos os pontos possuem valores altos e relativamente distantes para RMSECV, RMSEC e RMSEP, sendo assim, selecionou-se apenas o modelo gerado utilizando 6 VLs. A tabela 5.39 mostra os valores de  $R^2$ , RMSEP, as médias e os desvios padrões obtidos da razão dos valores preditos/reais para esse modelo e as figuras 5.54 e 5.55, os gráficos relacionando valores preditos e reais para as amostras de calibração e validação respectivamente.

**Tabela 5.39** Valores de  $R^2$  e RMSEP para o modelo construído para a determinação da viscosidade

Regiões do espectro	VL	$R^2$	RMSEP	Média Razão predito / real	Desvio Padrão
Intervalo 2	6	0,9609	0,0242	0,9999	0,0144



**Figura 5.54** Gráfico com o resultado de predição da viscosidade em amostras de calibração na associação no intervalo 2.



**Figura 5.55** Gráfico com o resultado de predição da viscosidade em amostras de validação no intervalo 2.

Com isso, pode-se concluir que o modelo gerado pelo método iPLS não obteve um desempenho tão bom quanto dos modelos anteriores (para a determinação das frações e da densidade) com RMSEP menor que 0,02, porém seus resultados são, ainda assim, satisfatórios sendo, portanto, o modelo perfeitamente aceitável para a quantificação da viscosidade.

#### 5.2.6.2 Aplicação do método PCR para a predição da viscosidade

Os mesmos critérios adotados para a determinação das frações e da densidade por PCR foram utilizados para se estimar a viscosidade. Portanto, a quantidade de PCs utilizadas na construção de cada modelo de acordo com os valores mínimos de correlação entre a viscosidade e as componentes principais estão descritos na tabela 5.40.

**Tabela 5.40** Número de PCs utilizadas na construção dos modelos, por PCR, na predição da viscosidade

Região	Valor Mínimo de Correlação	Nº de PCs utilizadas
	0,5	1
Intervalo 2 (771,47 a 941,19 cm <sup>-1</sup> )	0,2	4
	0,1	5
	0,05	10

Os modelos foram, então, construídos com o número de componentes principais para todas as possibilidades acima para a predição das amostras de validação.

#### 5.2.6.2.1 Definição dos melhores modelos obtidos por PCR na determinação da viscosidade e predição de amostras externas

Dentre os modelos construídos, primeiramente foram selecionados aqueles com valor de R<sup>2</sup> maior que 0,90 conforme tabela 5.41.

**Tabela 5.41** Modelos que apresentaram R<sup>2</sup> maior que 0,90, por PCR, na predição da viscosidade

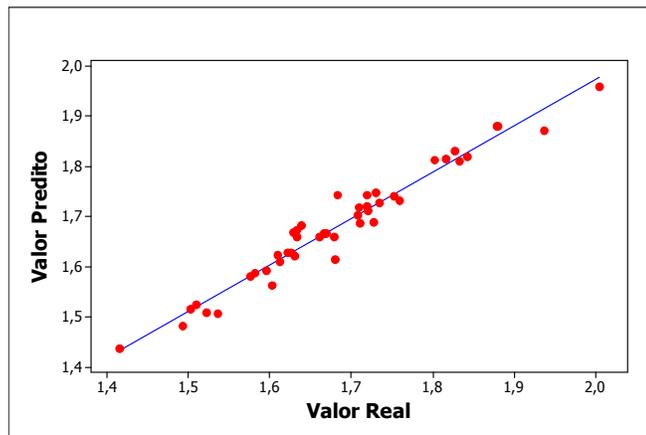
Região	Nº de PCs utilizadas	R <sup>2</sup>	RMSEP
Intervalo 2	4	0,939	0,0272
	5	0,954	0,0256
	10	0,972	0,0244

Verifica-se nessa tabela que todos os valores de RMSEP foram acima de 0,02. Portanto, selecionou-se o modelo construído com 5 PCs. Embora o modelo utilizando 10 PCs tenha resultado em um valor menor de RMSEP, o mesmo foi bastante próximo do obtido com 5 PCs não compensando, portanto, o uso do dobro de componentes principais. A tabela 5.42 mostra as médias e os desvios padrões obtidos da razão dos valores preditos/reais para o modelo selecionado e a figura

5.56 o correspondente gráfico relacionando os valores de viscosidade preditos e reais.

**Tabela 5.42** Média e Desvio Padrão para o modelo selecionado para a predição da viscosidade por PCR

Região	Nº de PCs utilizadas	R <sup>2</sup>	RMSEP	Média Razão predito / real	Desvio Padrão
Intervalo 2	5	0,9540	0,0256	0,999376959	0,015



**Figura 5.56** Gráfico com o resultado de predição da viscosidade em amostras de validação no intervalo 2 por PCR.

Diante dos resultados obtidos pode-se concluir que a predição da viscosidade de frações de diesel por PCR obteve um desempenho bastante parecido ao gerado por iPLS e, da mesma forma, levou a um modelo perfeitamente aceitável para se estimar a viscosidade, porém com resultados não tão bons quanto dos modelos anteriores (gerados para a determinação das frações e da densidade) com RMSEP menor que 0,02.

## **6 CONCLUSÕES**

Os modelos desenvolvidos neste trabalho obtiveram resultados bastante satisfatórios demonstrando, assim, que a técnica espectrométrica FTIR-ATR combinada a ferramentas quimiométricas tais como PLS e PCR podem ser usadas para a quantificação do petróleo de origem (% v/v) em misturas de frações de diesel, bem como para a predição de sua densidade e viscosidade.

A análise por componentes principais (PCA) permitiu, através de inspeção visual rápida, a identificação do petróleo de origem predominante em cada mistura.

O modelo para a quantificação das frações simultaneamente forneceu valores de RMSEP e  $R^2$  próximos aos obtidos pelos modelos gerados para a quantificação das frações individualmente demonstrando uma boa correlação entre eles e possibilitando, assim, se gerar um banco de dados como forma de acompanhar e reconhecer o petróleo através de suas frações com maior confiabilidade dos resultados.

Todos os modelos gerados a partir do método siPLS obtiveram valores de RMSEP e número de variáveis latentes (VLs) utilizados bem próximos aos gerados pelos respectivos modelos iPLS não justificando, portanto, o seu uso em nenhum caso.

Quanto aos modelos gerados por iPLS e PCR, concluímos que ambos podem ser utilizados para a quantificação do petróleo de origens distintas a partir de suas frações, bem como para se estimar as densidades e viscosidades dessas misturas, satisfatoriamente com resultados semelhantes (em alguns casos o método iPLS mostrou-se apenas ligeiramente melhor que o PCR) não havendo, portanto, diferenças significativas que levem a escolha de um ou outro modelo.

## 6.1 SUGESTÕES PARA TRABALHOS FUTUROS

Neste trabalho foram desenvolvidas metodologias inovadoras de caracterização de frações de petróleo empregando-se quimiometria e espectroscopia no infravermelho. A principal frente para continuação dessa pesquisa visaria a aplicação futura dessas metodologias na caracterização (qualificação e quantificação) de *blending* de petróleo bruto, implicando na aplicação de um protocolo rápido e simples de

identificação dos petróleos presentes a partir de um banco de dados previamente existente. Sugere-se, também, o estudo de metodologias para, a partir da quantificação dos petróleos de origem em misturas de petróleo, a previsão de outros parâmetros dessas frações como, por exemplo, BTEX, entre outros.

O modelo desenvolvido pode ser também aplicado no estudo de misturas de biodiesel:diesel visando a identificação da matriz e da quantidade de biodiesel (BX) utilizado na composição com o diesel, bem como, ser um instrumento seguro, simples e rápido na identificação de adulterações das mesmas.

# **REFERÊNCIAS BIBLIOGRÁFICAS**

- 1 – SPEIGHT, J. G.; **Handbook of Petroleum Product analysis**, Wiley-Interscience. USA, 2002.
  
- 2 – ANP - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis - **Petróleo e derivados**. Disponível em: <[http://www.anp.gov.br/petro/refino\\_editorial.asp](http://www.anp.gov.br/petro/refino_editorial.asp)>. Acesso em 03 de outubro de 2007.
  
- 3 – RIAZI, M. R.; **Characterization and properties of petroleum fractions**. ASTM Stock Number: MNL50, First Edition, USA, Philadelphia, PA, 2005.
  
- 4 – BRAUN, S.; APPEL, L. G.; SCHMAL, M.; A poluição gerada por máquinas de combustão interna movidas a diesel – a questão dos particulados. Estratégias atuais para a redução e controle das emissões e tendências futuras. **Quim. Nova**, v.27, 472-482, 2003.
  
- 5 – SKOOG, D A.; LEARY, J. J.; **Principles of Instrumental Analysis**, 4<sup>th</sup> ed., Saunders: London, 1992.
  
- 6 – HARRIS, C. D.; **Análise Química Quantitativa**, 6<sup>a</sup>. Ed., Rio de Janeiro: LTC Editora, 2008.
  
- 7 – PARISOTTO, G.;. **Determinação do número de acidez total em resíduo de destilação atmosférica e de vácuo do petróleo empregando a espectroscopia no infravermelho (ATR-FTIR) e calibração multivariada**. Dissertação de mestrado, Universidade Federal de Santa Maria, RS, 2007.
  
- 8 – OLINGER, J. M.; GRIFFITHS, P. R.; Quantitative effects of an absorbing matrix on near-infrared diffuse reflectance spectra, **Anal. Chem.**, v.60, 2427-2435, 1988.
  
- 9 – HARRICK, N. J., Multiple reflection cells for internal reflection spectrometry, **Anal. Chem.**, v.36, 188 – 191, 1964.

10 – FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L.O.; Quimiometria I: calibração multivariada, um tutorial, **Quím. Nova**, v.22, 724-731, 1999.

11 – OLIVEIRA, J. S.; **Avaliação da qualidade de biodiesel por espectroscopia FTIR e FTNIR associadas à quimiometria**. Dissertação de Mestrado, Universidade de Brasília, DF, 2007

12 – ZENI, D.; **Determinação de cloridrato de propranolol em medicamentos por espectroscopia no infravermelho com calibração multivariada (PLS)**. Dissertação de Mestrado, Universidade Federal de Santa Maria, RS, 2005.

13 – HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C.; **Multivariate Data Analysis**, 5<sup>th</sup> ed., Prentice-Hall: London, 1998.

14 – ARAUJO, T. P.; **Emprego de espectroscopia no infravermelho e métodos quimiométricos para a análise direta de tetracilinas em leite bovino**. Dissertação de Mestrado, Universidade Estadual de Campinas, SP, 2007.

15 – CORREIA, P. R. M.; FERREIRA, M. M. C.; Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos, **Quím. Nova**, v.30, 2007.

16 – BRERETON, R. G.; **Chemometrics: Data Analysis for the laboratory and Chemical Plant**. John Wiley & Sons, USA, 2003.

17 - JOHNSON, R. A., WICHERN, D. W; **Applied Multivariate Statistical Analysis**. 2. ed. New Jersey: Prentice Hall, 1988.

18 – MOTA, M. F. B.; **Implantacao de um sistema de destilacao atmosferica de petroleos no labpetro-UFES e estudos quimiométricos de frações**. Dissertação de Mestrado, Universidade Federal do Espírito Santo, ES, 2008.

19 - THOMAS, E. V.; A Primer on Multivariate Calibration, **Anal. Chem.**, v.66, n.15, 795A - 803A, 1994.

20 – SOTELLO, F. F.; **Aplicacao da espectroscopia de infravermelho próximo na caracterizacao de petroleo. Simulação de uma unidade de destilação atmosférica.** Tese de Doutorado, Escola Politécnica da Universidade de São Paulo, SP, 2006.

21 - CENTNER, V.; MASSART, D. Elimination of uninformative variables for multivariate calibration, **Anal. Chem.**, v.68, n.21, 3851-3858, 1996.

22 – EINAX, J. W.; ZWANZIGER, H. W.; GEIB, S.; **Chemometrics in environmental analysis**, Wiley-VCH 1997, USA , 1997

23 - SEKULIC, S.; SEASHOLTZ, M. B.; WANG, Z.; KOWALSKI, B. R.; Nonlinear multivariate calibration methods in analytical chemistry, **Anal. Chem.**, v.65, n.19, A835-A845, 1993.

24 – PEDRO, A. M. K.; **Determinação simultânea e não-destrutiva de sólidos totais e solúveis, licopeno e beta-caroteno em produtos de tomate por espectroscopia no infravermelho próximo utilizando calibração multivariada.** Dissertação de Mestrado, Universidade Estadual de Campinas, SP, 2004.

25 – PINHEIRO, G. R. V.; **Redes neurais aplicadas na inferência de propriedades de derivados de petróleo.** Dissertação de Mestrado, Pontifícia Universidade Católica do Rio de Janeiro, RJ, 1996.

26 – GELADI, P.; KOWALSKI, B. R.; **Anal. Chim. Acta**, v.185, 1- 17, 1986.

27 – VANDEGINSTE, B. G. M.; MASSART, D. L.; BUYDENS, L. M. C.; JING, S.; LEWI, P. J.; Smeyers-Verbeke, J. **Handbook of Chemometrics and Qualimetrics: Part B.** Amsterdam: Elsevier, 1998.

28 – BEEBE, K. R.; KOWALSKI, B. R. An Introduction to Multivariate Calibration and Analysis, **Anal. Chem.**, v.59, n.17, 1007A-1017A, 1987.

29 – NORGAARD, L.; SAUDLAND, A.; WAGNER, J.; NIELSEN, J. P.; MUNCK, L.; ENGELSEN, S.B.; Interval partial least-square regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy, **Appl. Spectrosc.**, v.54, n.3, 413-419, 2000.

30 – NETO, W.B.; **Parâmetros de qualidade de lubrificantes e óleo de oliva através de espectroscopia vibracional, calibração multivariada e seleção de variáveis.** Tese de Doutorado, Universidade Estadual de Campinas, SP, 2005.

31 – OLIVEIRA, F. C. C.; SOUZA, A. T. P. C.; DIAS, J. A.; DIAS, S. C. L.; RUBIM, J. C.; A escolha da faixa espectral no uso combinado de métodos espectroscópicos e quimiométricos, **Quím. Nova**, v.27, 218-225, 2004.

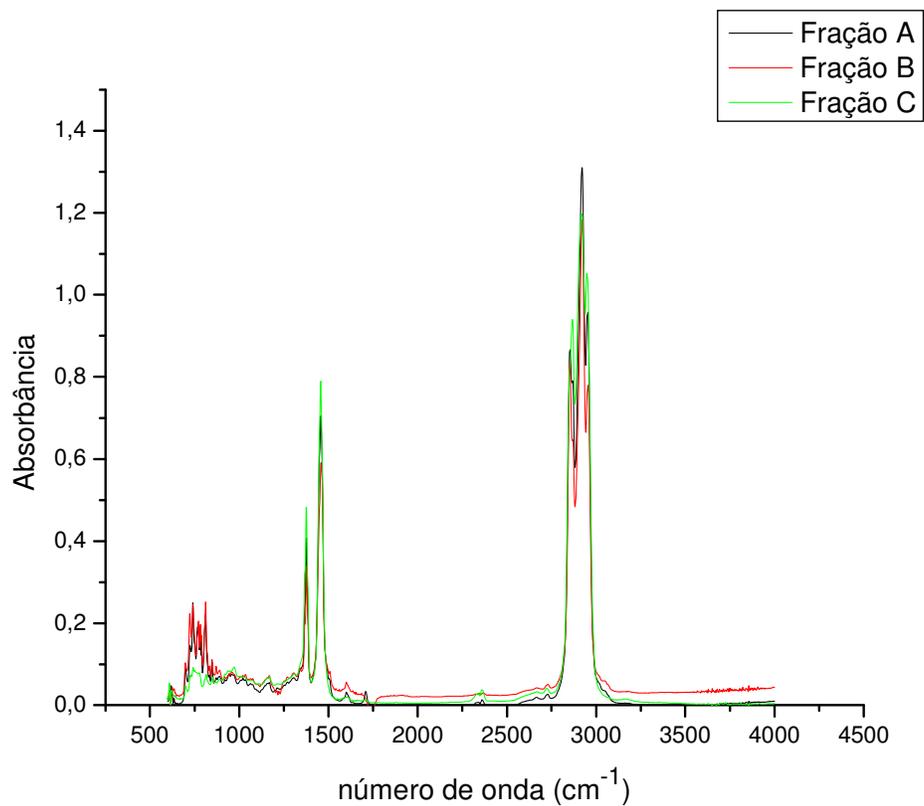
32 – BRERETON, R. G.; Introduction to multivariate calibration in analytical chemistry, **Analyst**, v.125, 2125-2154, 2000.

33 – COSTA FILHO, P.A.; POPPI, R.J.; Algoritmo genético em química, **Quím. Nova**, v.22, n.3, 405-411, 1999.

34 – SOARES, I.P.; REZENDE, T.F.; SILVA, R.C.; CASTRO, E.V.R.; FORTES, I.C.P.; Multivariate Calibration by Variable Selection for Blends of Raw Soybean Oil/Biodiesel from Different Sources Using Fourier Transform Infrared Spectroscopy (FTIR) Spectra Data, **Energy & Fuels**, v22, 2079-2083, 2008.

## ANEXO A – ESPECTROS DE INFRAVERMELHO

Os espectros de infravermelho, obtidos para as frações A, B e C estão representados na figura A1.1.



**Figura A1.1** Espectros obtidos das frações A, B e C