MARCELO APARECIDO MENDONÇA

CLASSIFICAÇÃO DE GASOLINAS COMERCIAIS ATRAVÉS DE MÉTODOS ESTATÍSTICOS MULTIVARIÁVEIS

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Engenharia.

São Paulo

2005

MARCELO APARECIDO MENDONÇA

CLASSIFICAÇÃO DE GASOLINAS COMERCIAIS ATRAVÉS DE MÉTODOS ESTATÍSTICOS MULTIVARIÁVEIS

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Mestre em Engenharia.

Área de Concentração:
Engenharia Química

Orientador:
Galo A. Carrillo Le Roux

São Paulo

2005

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência do seu orientador.

São Paulo, 29 de Abril de 2005.

Assinatura do autor:

Assinatura do orientador:

Mendonça, Marcelo Aparecido

Classificação de gasolinas comerciais através de métodos estatísticos multivariáveis. São Paulo, 2005.

98 p.

Dissertação (Mestrado) – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Química.

1. Classificação 2. Métodos Multivariáveis 3. Análise Discriminante 4. PCA 5. PLS 6. Gasolina

I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Química II. t

"Tenha vergonha de morrer até que tenhas conseguido realizar algum grande feito para a humanidade".

Horace Hann

À minha esposa Shirley e meu filho Julian.

AGRADECIMENTOS

Ao Prof. Dr. Galo A. Carrillo Le Roux, pela amizade, paciência e dedicada orientação.

Ao Prof. Dr. Roberto Guardani e ao Prof. Dr. Cláudio A. Oller do Nascimento, pelas sugestões importantes para aperfeiçoamento desta dissertação.

Ao Instituto de Pesquisas Tecnológicas do Estado de São Paulo.

À Agência Nacional do Petróleo.

À minha chefe Heloisa, pelo apoio, amizade e compreensão.

À minha amiga Alessandra, pelo apoio, paciência e contribuições.

Aos meus colegas do Laboratório de Combustíveis e Lubrificantes do IPT, pelo apoio nos ensaios e amizade.

Aos meus colegas do Laboratório de Análises Químicas Orgânicas do IPT, pelo apoio nos ensaios.

Aos meus pais Waldo e Adelaide, pelo apoio e carinho com o Julian.

A minha sogra Isabel, pelo apoio, paciência e carinho com o Julian.

Aos colegas do LSCP-EPUSP, em especial ao Reinaldo Ap. Teixeira, Francisco F. Sotelo e Gustavo, pelas contribuições e amizade nestes últimos anos.

E em especial à Deus, por mais esta realização em minha vida.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANP	Agência Nacional do Petróleo
PMQC	Programa de Monitoramento de Qualidade dos Combustíveis
AEAC	Álcool Etílico Anidro Combustível
LCL	Laboratório de Combustíveis e Lubrificantes
IPT	Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A
NIR	Infravermelho próximo
RON	Número de Octano Pesquisa
MON	Número de Octano Motor
IAD	Índice antidetonante
LDA	Análise Discriminante Linear
QDA	Análise Discriminante Quadrática
PCA	Análise do Componente Principal
PLS	Mínimos Quadrados Parciais
RDA	Análise Discriminante Regularizada
SIMCA	*Soft Independent Modeling by Class Analogy*
k-NN	*k-Nearest Neighbor*
LGO	Gasóleo Leve
LSR	Resíduo Lateral Leve e nafta
SNV	*Standard Normal Variate transform*
PC	Componente Principal
QSAR	*Quantitative Structure-Activity Relationship*
ANNs	Redes Neurais Artificiais
W-NN	Método combinado de NN com *clusters* compactos
GAs	Algoritmos Genéticos
GA-NN	Método combinado de Algoritmos Genéticos com NN
FR	Reconhecimento Facial
SSS	*Small Size Sample*
LAQO	Laboratório de Análises Químicas Orgânicas
LSCP	Laboratório de Simulação e Controle de Processos
EPUSP	Escola Politécnica da Universidade de São Paulo

PFE	Ponto final de ebulição
PE	Procedimento de Ensaio
RBC	Rede Brasileira de Calibração
PIE	Ponto Inicial de Ebulição
MIR	Infravermelho médio
CFR	*Cooperative Fuel Research Committee*
NMR	Ressonância Magnética Nuclear
FT-IR	Transformada de Fourier
PMC	Produtos de Marcação Compulsória
MDA	Análise Discriminante Múltipla
LV	Variável latente
Dadosfq	Dados físicos químicos brutos
Dadosfqsd	Dados físicos químicos sem massa específica
Dadosfqmar	Dados físicos químicos brutos com incerteza de medição
Dadosfqmarcsd	Dados físicos químicos brutos com incerteza de medição sem massa específica

LISTA DE SIMBOLOS

ε	absorvidade molar
λ	comprimento de onda
ν	freqüência
ρ	massa específica (g/cm^3)
$\Phi(x_i)$	função discriminante linear em x_i
$\Sigma_{combinada}$	matriz de covariância combinada
λ_i	autovalor
Σ_k	matriz de covariância de classe k
μ_k	média dos vetores de classe k
π_k	probabilidade da classe k
A	Absorbância
b	coeficiente de regressão
c	velocidade da luz
d	caminho óptico
da	massa específica do ar à temperatura de ensaio
dw	massa específica da água à temperatura de ensaio
E	energia
E	matriz residual da matriz X
F	matriz residual da matriz Y
h	constante de Planck
I	intensidade remanescente após a interação com a amostra
I_0	intensidade espectral da radiação emitida pelo espectrômetro
K	tipo de classe
n	total de objetos do conjunto de treinamento (calibração)
n_k	número de objetos dentro da classe k
P	período de oscilação
P	matriz *loading* da matriz X
p_i	vetor *loading* da matriz X
Q	matriz *loading* da matriz Y

q	vetor *loading* da matriz Y
R^2	coeficiente de correlação
T	matriz *scores* da matriz X
Ta	período observado da oscilação do tubo contendo ar
t_i	vetor *scores* da matriz X
Tw	período observado da oscilação do tubo contendo água
U	matriz *scores* da matriz Y
u	vetor *scores* da matriz Y
w	pesos de X
X	matriz de dados genérica
x_i	amostra i
Y	concentração molar
Z	eixo dos *scores* discriminantes

RESUMO

Neste trabalho estuda-se a aplicação de métodos estatísticos multivariáveis para a classificação de gasolinas comerciais em conformidade à legislação vigente. Atualmente, a ANP baseia a classificação em limites máximos e mínimos para uma série de diferentes propriedades físico-químicas. O objetivo do trabalho é propor uma metodologia para fazer uma triagem das amostras coletadas durante o Programa de Monitoramento da Qualidade dos Combustíveis através de um método de classificação. Ela utiliza a espectroscopia NIR, que é uma técnica rápida e não destrutiva, como método analítico. Com isto será possível reduzir o número de ensaios físico-químicos que não necessariamente seriam realizados sistematicamente em todas as amostras, reduzindo-se os custos e aumentando-se a quantidade de postos monitorados.

As análises NIR produzem grandes quantidades de dados, o que leva à utilização de técnicas estatísticas multivariáveis para estabelecer as metodologias de classificação. Neste trabalho utilizam-se técnicas já consagradas, como a PCA e a PLS para a compressão dos dados e a LDA e QDA para a classificação das amostras. Os dados analisados correspondem às propriedades físico-químicas e aos espectros NIR de um conjunto de 216 amostras de gasolinas comerciais, utilizado para a concepção dos modelos de classificação, e de outro de 50 amostras, utilizado para a validação dos modelos. Os modelos testados no trabalho foram as combinações da PCA-LDA, PCA-QDA, PLS-LDA, PLS-QDA, PLS (regressão) e a análise dos gráficos de *scores* (*biplot*). Os melhores desempenhos foram obtidos pelos gráficos dos *scores*, em seguida pela regressão PLS, PLS-QDA, PCA-QDA e PLS-QDA. Existem ainda algumas etapas a serem alcançadas para tornar prática a utilização da classificação de gasolinas comerciais através de NIR, no entanto, a contribuição deste estudo é importante pois permitiu demonstrar a sua viabilidade técnica.

ABSTRACT

In this work, the application of multivariable statistical methods for the classification of commercial gasoline in accordance to applicable laws in Brazil is studied. In the present, the ANP bases the classification of gasoline on lower and upper bounds defined for a number of physico-chemical properties. The objective of this work is to propose an alternative analysis methodology, that is adequate for making a pre-sorting of the samples collected by the Fuel Quality Monitoring Program through a classification method. This method is based on NIR spectroscopy, that is a fast and non-destructive technique, as the analytical method. In this way, it would be possible to reduce the number of physico-chemical analyses, as it would be possible not to perform them on every sample, reducing costs and increasing the quantity and frequency of gas stations that could be monitored.

NIR analyses produce a great quantity of data, that makes the use of multivariable statistical techniques necessary in order to set up classification methodologies. In this work the well-known PCA and PLS techniques are used for data compression, and LDA and QDA analyses for sample classification. The data studied correspond to the physico-chemical properties and NIR spectra of a total of 216 commercial gasoline samples, used for model design, and of a 50 samples, used for validation. The classification methods that are tested are combinations of PCA-LDA, PCA-QDA, PLS-LDA, PLS-QDA, PLS (regression) and data compression scores graphical analysis (biplot). Best performance was obtained with compression scores graphical analysis, followed by PLS regression, PLS-QDA, PCA-QDA and PLS-QDA. There are still some steps to be fulfilled before the usage of commercial gasoline classification through NIR could be practical. However, this study has shown that this methodology is technically feasible.

CAPÍTULO 1

1 INTRODUÇÃO

Atualmente são comercializadas nos postos de combustíveis do Brasil gasolinas provenientes do petróleo de diversas regiões do território nacional e do mundo. Há uma grande gama de refinarias, formuladores e distribuidoras de gasolinas, fazendo com que a composição deste produto varie bastante. O controle da comercialização destes combustíveis é feito atualmente pela Agência Nacional do Petróleo (ANP) através de fiscalizações em postos, distribuidoras e refinarias e pelo Programa de Monitoramento da Qualidade dos Combustíveis (PMQC) dos postos de abastecimento em todo o Brasil. Este programa tem o objetivo de avaliar permanentemente a qualidade dos combustíveis comercializados no País, desde o produtor até o consumidor final. Ele prevê a realização de diversos ensaios (testes físico-químicos) nas amostras de gasolinas para verificar se a mesma atende às especificações das portarias da ANP. As coletas destas amostras são feitas de forma aleatória nos postos de combustíveis.

Através do programa de monitoramento é possível mapear os problemas de não-conformidade e direcioná-los para tomada de ações pela fiscalização da ANP. As principais causas de não-conformidades da gasolina, para diferentes regiões do Brasil, em julho de 2004 foram identificadas pelos ensaios de destilação (44%), teor de álcool (37%), octanagem (16%) e outros (3%). Na região do estado de São Paulo e no mesmo período os ensaios foram: destilação (56%), teor de álcool (27%), octanagem (14%) e outros (3%). O ensaio para verificação de marcador de solventes não é utilizado comumente no monitoramento, em alguns períodos a ANP estabelece a realização deste ensaio em no máximo 10% de toda coleta realizada. Outros dados interessantes de se conhecer é a porcentagem de não conformidades e número de amostras coletadas através do monitoramento. No período de janeiro a junho de 2004 foram coletadas em todo o Brasil 44.934 amostras de gasolinas, sendo que 2.124 (4,7%) eram amostras não-conformes. No estado de São Paulo foram coletadas

14.086 e destas 1.440 (10,2%) eram não conformes (www.anp.gov.br). Na região da grande São Paulo (objeto deste trabalho), foram coletadas 2.699 amostras de gasolinas, no período de maio a junho de 2004, sendo que 302 (11,2%) eram amostras não-conformes. A região da grande São Paulo engloba um total de 85 municípios, correspondendo a aproximadamente 3.600 postos monitorados e cerca de 1.000 amostras ensaiadas por mês.

A maioria dos métodos utilizados nos ensaios para o monitoramento da qualidade requer um longo tempo de análise e altos investimentos em equipamentos analíticos. Com a intenção de reduzir estes custos, muitas empresas lançaram no mercado analisadores portáteis pelo princípio do infravermelho, que determinam várias propriedades da gasolina. Porém, a maioria destes equipamentos possui bancos de dados e modelos matemáticos ajustados para a realidade da gasolina de origem do país fabricante (ex.: Europa, Estados Unidos, etc). Estes países possuem gasolinas com composição diferente da gasolina nacional, principalmente no quesito adição de oxigenados. Atualmente adiciona-se à gasolina nacional do tipo A, 25 % (± 1%) de álcool etílico anidro combustível (AEAC), formando a gasolina do tipo C.

Como algumas propriedades não têm uma predição muito boa devido aos fatores citados acima, muitos pesquisadores têm desenvolvido trabalhos correlacionando os resultados obtidos por análises em espectroscópio no infravermelho com as análises físico-químicas através de técnicas estatísticas multivariáveis.

O objetivo deste trabalho é aplicar ferramentas estatísticas multivariáveis à classificação de gasolinas comerciais. Através destas visa-se construir modelos que permitam classificar da forma mais rápida e confiável possível, amostras de gasolinas comum comercializadas nos postos de combustíveis. A idéia é fazer uma triagem das amostras coletadas durante o Programa de Monitoramento da Qualidade dos Combustíveis (PMQC), realizado pelo Laboratório de Combustíveis e Lubrificantes (LCL) do Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A, que é conveniado à Agência Nacional do Petróleo. As amostras coletadas,

aleatoriamente nos postos da grande São Paulo e região, serão classificadas através das técnicas estatísticas multivariáveis em “conforme” e “não conforme”. Evita-se assim, a necessidade de ensaiar amostras, que já se sabe que são de boa qualidade. Para comparação das técnicas utilizadas no trabalho será utilizada a porcentagem de classificação errada (% erros) e/ou a porcentagem de classificação correta (% acertos).

Um dos objetivos deste trabalho é minimizar o número de análises físico-químicas a serem realizadas, substituindo estas por análises baseadas em NIR. Para tal, é interessante implementar uma metodologia baseada nesta análise que permita classificar as amostras em conformes e não conformes.

Nesta metodologia seria desejável que se evite ao máximo erros do tipo 2, ou seja, erros em que se considerem conformes amostras não conformes. Isto porque, a ocorrência de erros do tipo 1 levaria apenas à realização de análises físico-químicas inutilmente, enquanto que a ocorrência de erros do tipo 2 levaria à aceitação de amostras não conformes que não seriam submetidas a análises físico-químicas. A classificação proposta pela ANP é baseada em limites máximos e mínimos para uma série de diferentes propriedades físico-químicas.

Apesar da aparente redução no número de ensaios que um laboratório prestador de serviços poderia sofrer, na prática este número poderia até aumentar. O contrato junto à ANP pode ser mantido e a quantidade de amostras ensaiadas também, porém com um diferencial: que a maioria das amostras ensaiadas (de acordo com o nível de confiança obtido pelo modelo) seria praticamente de amostras não conformes. O ganho pode ser entendido de duas formas: (1) para o laboratório que além de realizar os mesmos testes físico-químicos faria os testes preliminares para a seleção das amostras e (2) para a ANP, e consequentemente a sociedade, teríamos um universo maior de postos monitorados em relação ao que é realizado hoje em dia, e com isso um mapeamento mais denso dos postos irregulares facilitando a atuação da fiscalização da ANP.

Além desta introdução, o presente trabalho está dividido em 4 capítulos, como descrito a seguir:

- Capítulo 2, revisão bibliográfica, onde são apresentados métodos de discriminação e algumas aplicações relatadas na literatura.

- Capítulo 3, onde são apresentados as metodologias, materiais e equipamentos utilizados neste trabalho para as seguintes análises e técnicas:
  - Análises físico-químicas, segundo Portaria n° 309 da ANP: Densidade, destilação (T10%, T50%, T90% e PF), teor de álcool e pelo analisador portátil: teor de álcool, benzeno e octanagem (RON, MON e IAD).
  - Análise da presença de marcador, segundo a Portaria n° 274 da ANP.
  - Análises das amostras através de espectrometria por infravermelho próximo (NIR).
  - Técnicas estatísticas multivariáveis para classificação das amostras de gasolina, conforme descrito a seguir:
    - LDA (Análise Discriminante Linear), aplicado nas análises físico-químicas e marcador,
    - QDA (Análise Discriminante Quadrática), aplicado nas análises físico-químicas e marcador,
    - PCA-LDA (PCA: Análise de Componentes Principais), aplicado nos espectros NIR,
    - PCA-QDA, aplicado nos espectros NIR,
    - PLS-LDA (PLS: Mínimos Quadrados Parciais), aplicado nos espectros NIR,
    - PLS-QDA (PLS: Mínimos Quadrados Parciais), aplicado nos espectros NIR,
    - Regressão (PLS) para a predição da conformidade,
    - *Biplot* utilizando PCA e PLS.

- Capítulo 4, onde é apresentada a aplicação de técnicas de discriminação a amostras de combustível coletadas nos postos revendedores da grande São Paulo.

- No capítulo 5 são apresentadas as conclusões deste trabalho e as propostas para a sua continuação.

CAPÍTULO 2

2 REVISÃO BIBLIOGRÁFICA

2.1 Métodos de Classificação

Segundo Johnson et al. (1998) discriminação e classificação são técnicas multivariáveis com objetivos distintos: a análise discriminante tem o objetivo de descrever algebricamente e graficamente características diferenciais de objetos (observações) de muitas coleções conhecidas (populações), ela é conhecida também como "separação". Já a técnica de classificação classifica os objetos (observações) dentro de duas ou mais classes. A ênfase é na derivação de uma regra que pode ser usada para otimizar a classificação de novos objetos dentro das classes, conhecida também por "alocação". Porém, na prática estas duas definições freqüentemente se confundem tornando difícil suas distinções.

Em 1938, Fisher, R. A., introduziu o termo discriminação e a análise discriminante no primeiro tratamento de problemas de separação (Johnson, et al.,1998 e Otto, 1999). Dentre as técnicas mais tradicionais de discriminação encontram-se a LDA (*Linear Discrimant Analysis*), a QDA (*Quadratic Discrimant Analysis*) e a RDA (*Regularized Discriminant Analysis*) (Wu, et al., 1996).

Atualmente, a disponibilidade de técnicas analíticas que resultam em grandes massas de dados, tais como as técnicas espectroscópicas, em particular o NIR (infravermelho próximo), têm levado à utilização de técnicas estatísticas multivariáveis aliadas a metodologias de classificação. A aplicação da LDA, QDA e RDA a grandes massas de dados tem sérias limitações, já que o número de variáveis de entrada tem que ser menor do que o número de amostras (objetos). Conseqüentemente, a utilização da LDA em espectros NIR requer a pré-seleção ou a transformação dos dados, para as quais são utilizadas as técnicas estatíticas multivariável. Dentre as técnicas de transformação de dados mais utilizadas estão a

PCA (*Principal Component Analysis*) e o PLS (*Partial Least Squares*), técnicas hoje em dia bem estabelecidas e difundidas (Johnson et al., 1998).

A escolha de um método de classificação depende do conjunto de dados. A performance da LDA e da QDA depende do tamanho do conjunto de treinamento, de se a classificação de população segue uma distribuição normal multivariável e de quanto as matrizes de covariância podem ser consideradas iguais entre si, ou não, nas classificações.

Em alguns casos, a análise discriminante tradicional (LDA, QDA e RDA) não tem um desempenho adequado. Desenvolvimentos recentes têm levado à utilização de técnicas tais como a SIMCA (*Soft Independent Modeling by Class Analogy*) e a k-NN (*k-Nearest Neighbor*) que são baseadas em medidas de similaridade dentro de uma classe (Tominaga, 1999).

2.2 Aplicações em literatura

A espectroscopia no infravermelho próximo (NIR) tem sido freqüentemente aplicada como um método analítico que fornece informação suficiente para a determinação de moléculas orgânicas (como por exemplo: proteínas e gorduras) e parâmetros qualitativos de produtos da agricultura e da indústria de alimentos. Aplicações mais recentes têm sido realizadas nas indústrias do petróleo, têxtil, carvão, cosméticos, polímeros, química, tintas e farmacêutica. Segundo McClure (2003) historicamente este método pode ser organizado cronologicamente, de acordo com as aplicações tecnológicas do NIR, em seis períodos: (1) descoberta (1800-1939), (2) os anos da definição (1940-1959), (3) os anos da agricultura (1960-1979), (4) os anos da quimiometria (1980-1989), (5) os anos da industria (1990-1999) e (6) os anos da imagem (2000-).

Em seu trabalho McClure (2003) apresenta a distribuição anual dos últimos 53 anos (1950-2003) de publicações sobre o NIR. De 1800 a 1950 a CNIRS (Base de

dados de NIR mantido pelo Comitê bibliográfico do conselho para espectroscopia no infravermelho próximo, do qual o autor faz parte) cita somente 91 artigos. Em 1970 atingiu-se a marca de 100 publicações por ano e em 1980 com o surgimento do computador pessoal e aumento da produção de instrumentos NIR, o número de publicações cresceu exponencialmente até 1997 quando atingiu a marca de 3000 publicações por ano. Após esta data houve um declínio na produção por ano, porém esta queda observada pode ser enganosa, pois muitos pesquisadores tratam o NIR como um método infravermelho apenas. Entre 1997-2003 (McClure, 2003) as publicações específicas sobre infravermelho próximo excederam a respeitável marca de 500 publicações por ano.

Com respeito a trabalhos em que foram utilizados o NIRS (*Near-Infrared Reflectance Spectroscopy*) e o FT-NIR (*Fourier Transform Near-Infrared*) como métodos analíticos juntamente com métodos de classificação, foram encontradas 40 aplicações em revistas indexadas entre 2002-2005. Destas, 11 tratam de aplicações em alimentos, 11 em agricultura, 5 em farmácia, 3 em reconhecimento de padrões, 3 em química, 2 em medicina e em biologia, petroquímica, astronomia, física e geologia foi registrada 1 em cada área. Quanto aos métodos de classificação mais utilizados: em 10 publicações foi o SIMCA, em 7 a LDA, em 3 a KNN, em 3 a CDA (*Canonical Discriminant Analysis*), em 3 a MDA (*Multiple Discriminat Analysis*), entre outros.

No campo do petróleo, há uma ampla gama de aplicações relacionadas a produtos destilados, tais como gasolina, diesel, nafta e querosene. Muitas das aplicações reportadas estão associadas à previsão da qualidade para produtos destilados médios, incluindo nafta, gasolina e diesel, cobrindo propriedades tais como composição de hidrocarbonetos, destilação, número de octanagem, pressão de vapor Reid e índice de cetano (Honigs, 1985; Kelly, 1990; Litani-Barzilai, 1997; Chung, 1999; Kim, 2000; etc.). Poucas aplicações existem nesta área em classificação ou discriminação.

Conforme foi apresentado acima, muitos trabalhos sobre análise discriminante e classificação são encontrados na literatura, nas mais variadas áreas. No entanto, para este trabalho, a aplicação em si pode não ser muito importante, mas sim a metodologia empregada, pois esta pode ser utilizada neste estudo.

Na área de derivados de petróleo, em um estudo recente, Kim et al. (2000) utilizaram a espectroscopia NIR para classificar em linha seis derivados diferentes (diesel, gasolina, querosene, LGO - Gasóleo Leve, LSR – Resíduo Lateral Leve e nafta) provenientes de uma coluna de destilação. Eles analisaram os espectros de 273 amostras e construíram gráficos dos espectros dos seis produtos de petróleo. Duas faixas espectrais foram excluídas da análise, baseado em informação espectroscópica a priori das amostras. Os dados foram pré-tratados pelo método SNV (*Standard Normal Variate transform)*, e a PCA foi utilizada para compressão e extração das informações relevantes, e depois um modelo de classificação para cada classe foi obtido com estas informações. As 700 informações espectrais foram reduzidas para 9 dimensões de PCs (componentes principais), que representavam 99,97% da variância dos dados originais. A análise do gráfico biplot dos dois primeiros componentes principais (PC1xPC2, representando 98,06% da variância dos dados originais) permitiu visualizar uma boa classificação da gasolina, querosene, LSR e nafta. Já o diesel e LGO misturam-se no espaço bi-dimensional, no entanto, eles se separaram no espaço de 9 dimensões. A função discriminante utilizada é baseada na QDA. A porcentagem de erros tipo 1 do classificador obtido foi inferior a 6%, desta forma ele foi considerado pelos autores como suficientemente preciso.

Dos grupos mais ativos na área de classificação podemos destacar o da ChemAC Pharmaceutical Institute de Bruxelas, do qual fazem parte Wu e Massart. Wu et al. (1996) utilizaram 7 conjuntos de dados de NIR para comparar classificadores. Três dos conjuntos foram especialmente projetados e os demais foram obtidos da industria. Os conjuntos de dados correspondiam a problemas de classificação de amostras de classes diferentes, tais como: Duas classes, uma de para-xileno puro e a outra de para-xileno contaminado com orto-xileno; três classes de mistura de diferentes produtos (*celulose, manitol, sucrose, sal sódico de sacarina e*

ácido cítrico); quatro classes de produtos de polímeros; duas classes, uma de butanol puro, e a outra de butanol com diferentes concentrações de água; classes de drogas com diferentes dosagens e três tipos de placebos; treze classes de polímeros; e duas classes de solventes, um puro e o outro não.

Foram obtidos os espectros NIR dos 7 conjuntos de dados. Em seguida eles aplicaram a transformada SNV como método de pré-tratamento. Aplicando os classificadores discriminantes (LDA, QDA e RDA) os autores concluíram que a QDA não produziu resultados satisfatórios porque não se estava trabalhando com classes de amostras de grande dimensão em comparação com o número de variáveis apresentado no problema. Ele só apresentaria vantagens em comparação ao LDA, caso as matrizes de covariância das classes fossem bastante diferentes. A LDA produz melhores resultados quando os tamanhos das amostras são pequenos. A RDA sempre fornece resultados equivalentes ou melhores que a LDA e a QDA. Em muitos casos ela se reduz automaticamente à LDA e, algumas vezes à QDA. Por fim, o ganho que se pode obter aplicando RDA em vez de LDA e QDA é relativamente pequeno na prática, e às vezes não é suficiente para se obter uma boa classificação.

Em outro estudo, Wu et al. (1997) concluíram que a LDA, junto com a PCA como método de redução de dimensão, obteve um excelente desempenho na classificação dos dados de três conjuntos de espectros NIR. O primeiro conjunto consistia de 60 espectros NIR (1376-2398 nm, 512 comprimentos de ondas) de três bateladas de excipientes (compostos da formulação de uma droga sem o princípio ativo) o qual era formado por uma mistura de diferentes produtos (*celulose, manitol, sucrose, sal sódico de sacarina e ácido cítrico*), o segundo conjunto era formado por 83 espectros NIR (1330-2352 nm, 512 comprimentos de ondas) de 4 classes de produtos de polímeros e o terceiro continha 135 espectros NIR (110-2500 nm, 700 comprimentos de ondas) de tabletes contendo diferentes dosagens de ingrediente ativo experimental, placebo e um comparador clínico. O objetivo deste estudo foi o teste de diferentes algoritmos de PCA com o intuito de comparar principalmente o tempo de processamento, que era elevado.

Uma aplicação importante das técnicas de classificação tem sido a QSAR (*Quantitative Structure-Activity Relationship*) muito utilizada na área farmacêutica. Tominaga (1999) apresentou um estudo em que foram comparados sete métodos de classificação na análise de três tipos de agentes quimioterapêuticos (antibacteriais, antineoplásticos e antifungais) baseada em 156 descritores. Os métodos foram: PCA-LDA, SIMCA, PLS2, Redes Neurais Artificiais (ANNs), NN, método combinado de NN com *clusters* compactos (W-NN) e um método combinado de Algoritmos Genéticos (GAs) com NN (GA-NN). O total de amostras utilizadas para treinamento (calibração) dos métodos foi de 12242 entre os três agentes.

Segundo Lu et al. (2003) o reconhecimento facial (FR) é outro domínio em que a discriminação tem um grande potencial de aplicação e para o qual tem sido muito estudada. Nas últimas duas décadas numerosos algoritmos de FR foram propostos. Dentre as metodologias FR mais utilizadas, aquelas baseadas em LDA têm apresentado um bom desempenho. O problema do reconhecimento facial se caracteriza pelo que se denomina de "amostra de tamanho pequeno" (SSS – *Small Size Sample*): os dados têm dimensões elevadas, sendo que o número de dados por objeto é elevado, mas o número de amostras de treinamento é normalmente muito pequeno em comparação com a dimensão do espaço amostral.

CAPÍTULO 3

3 MATERIAIS E MÉTODOS

Neste capítulo são apresentados os materiais e métodos utilizados para a realização dos ensaios físico-químicos para a classificação de gasolinas. São apresentados também os métodos estatísticos multivariáveis propostos para a análise de discriminação e classificação. Uma breve descrição do uso e significância de cada método é abordada durante o capítulo.

Os métodos utilizados neste trabalho foram divididos em dois grupos:
- Métodos analíticos (parte experimental).
- Métodos estatísticos multivariáveis (parte computacional).

3.1 Métodos Analíticos

Os ensaios analíticos de destilação, teor de álcool, massa específica, teor de benzeno e octanagem foram realizados no Laboratório de Combustíveis e Lubrificantes (LCL) do Instituto de Pesquisas Tecnológicas de São Paulo (IPT) segundo a Portaria ANP N° 309 de 27/12/2001. O ensaio de marcador de solventes foi realizado no Laboratório de Análises Químicas Orgânicas (LAQO) do IPT segundo a Portaria ANP N° 274 de 01/11/2001. O ensaio de espectroscopia pelo princípio no infravermelho próximo (NIR) foi realizado no Laboratório de Simulação e Controle de Processos (LSCP) da Escola Politécnica da Universidade de São Paulo (EPUSP), Departamento de Engenharia Química.

A tabela 3.1 mostra o resumo dos ensaios realizados neste trabalho, o método utilizado e a especificação ou limites estabelecidos pela Agência Nacional do Petróleo (ANP) quando aplicável.

Os ensaios de octanagem (MON e IAD) foram realizados conforme procedimento de ensaio laboratorial para analisadores portáteis de gasolina, a especificação da portaria se refere ao ensaio realizado em motor monocilíndrico. Estes ensaios estão mais detalhados no capítulo sobre octanagem.

Tabela 3.1 - Métodos utilizados nos ensaios deste trabalho

Característica	Unidade	Especificação Gasolina Comum – Tipo C	Método
Álcool Etílico Anidro Combustível - AEAC	%vol	25 ± 1	NBR 13992* e PE*
Massa Específica a 20°C	Kg/m^3	anotar	NBR 14065* ASTM D4052*
Destilação (10% evaporado) - max.	°C	65,0	NBR 9619* ASTM D86*
Destilação (50% evaporado) - max.	°C	80,0	
Destilação (90% evaporado) - max.	°C	145,0-190,0	
Destilação (PFE) - max.	%vol	220,0	
N° de Octano Motor – MON – min.	-	82,0	MB 457, ASTM D2700 e PE*
Índice antidetonante – IAD – min.	-	87,0	ASTM D 2699, D 2700 e PE*
Benzeno, max.	%vol	1,0	ASTM D 6277* PE*
Marcador	ppb	-	PE*
Espectroscopia NIR	-	-	Manual do Fabricante*

PE = Procedimento de Ensaio
* Método utilizado no trabalho

Devido ao alto custo dos ensaios, naqueles realizados no LCL-IPT foi feita somente uma leitura (análise) de cada amostra, porém foi realizado um controle estatístico do processo ou gráfico de controle. O gráfico de controle é construído plotando as médias das leituras diárias e a amplitude entre as leituras. A média geral é obtida entre todos os pontos plotados, já os limites de controle (superior e inferior) são obtidos pelo produto entre a amplitude média das leituras e o valor estatístico tabelado.

Resumidamente, para os ensaios de destilação, teor de álcool (proveta) e massa específica foram mostrados as médias gerais e os desvios padrões obtidos neste controle, sempre comparando com a repetitividade e reprodutibilidade do método utilizado. Como parâmetro de avaliação foi adotada a reprodutibilidade do método já que em alguns casos o equipamento era o mesmo, porém o técnico era trocado semanalmente, ou como na destilação os equipamentos eram diferentes assim como o técnico. Para os ensaios no analisador portátil de gasolina foi apresentada somente a curva de calibração com os valores obtidos na regressão.

Nos ensaios de marcador de solventes realizado no LAQO-IPT também foi feita somente uma leitura de cada amostra, e o controle foi realizado utilizando padrões e gráfico de controle. Já os ensaios para obtenção dos espectros NIR, realizados no LSCP-USP, foram feitos em duplicata e a sua média foi utilizada.

3.1.1 Destilação

O ensaio de destilação propicia uma medida em termos de volatilidade, das proporções relativas de todos os hidrocarbonetos componentes de uma gasolina (Campos, 1990). Esta característica (volatilidade) tem um efeito sempre importante na sua segurança e desempenho. A volatilidade é o determinante principal da tendência de um hidrocarboneto de, potencialmente, produzir vapores explosivos (ASTM D86 e NBR 9619).

O gráfico das porcentagens do destilado e suas temperaturas correspondentes formam a chamada curva de destilação, de grande utilidade para prever o desempenho da gasolina no motor (figura 3.1) e para a detecção de adulterações com produtos de características diferentes. A temperatura de evaporação dos 10% indica a quantidade de frações leves presentes na gasolina e deve ser baixa o suficiente para promover uma partida fácil e rápida do motor sob condições normais de temperatura ambiente. As características de aquecimento e aceleração do motor dependem das frações intermediárias controladas pelo ponto 50%. A fração pesada da gasolina é

verificada pelos pontos 90%, ponto final de ebulição e porcentagem do resíduo. Estes componentes contribuem para uma boa economia de combustível, porém uma adulteração que provoque alteração na temperatura de 90% evaporados pode provocar danos ao motor, como por exemplo, depósitos excessivos na câmara de combustão, formação de vernizes e borras (Campos, 1990).

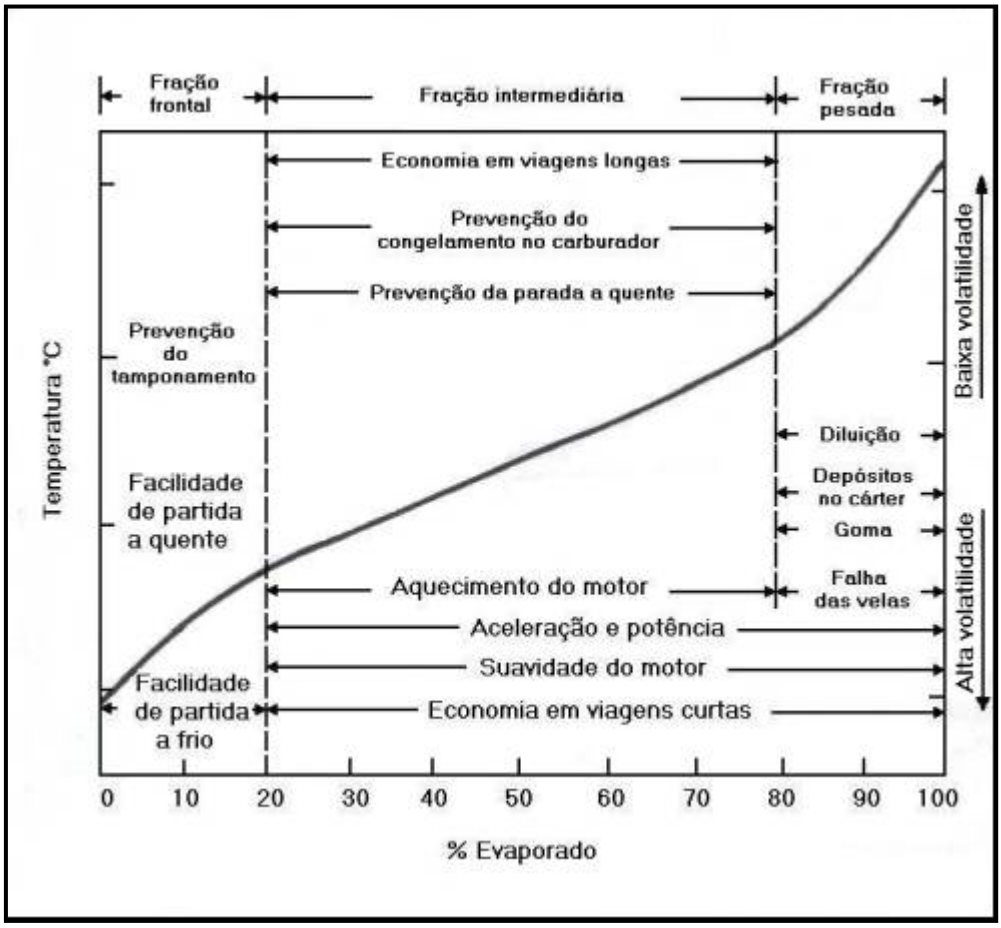

Figura 3.1 - Curva de destilação da gasolina com desempenho do motor
(Fonte: Campos, 1990)

Para os ensaios de destilação foram utilizados 5 destiladores automáticos, sendo que 2 destiladores eram da marca Normalab, modelo 440 NDI e três destiladores da marca ISL modelo AD 86 5G. Todos os destiladores foram calibrados eletronicamente de acordo com as especificações dos fabricantes, as termoresistências utilizadas para medição da temperatura de evaporação foram calibradas pelo Laboratório de Metrologia do IPT que pertence à Rede Brasileira de

Calibração (RBC). O sensor de nível e as provetas utilizadas para medição do volume de condensado foram calibradas de acordo com recomendações do manual do fabricante.

Como os destiladores são automáticos um programa específico para gasolina foi desenvolvido e gravado no próprio destilador padronizando as condições de ensaio, conforme especificações da norma e do fabricante.

Os ensaios foram realizados conforme as normas ASTM D86 e NBR 9619. Resumidamente o ensaio consiste em destilar 100 ml de gasolina sob condições específicas (tabela 3.2), condensar o destilado em proveta calibrada de 100 ml e registrar as temperaturas do recuperado porcento (condensado) correspondentes às porcentagens de destilado especificadas pela Portaria da ANP (10%, 50%, 90%, PFE e resíduo). Após o PFE, o equipamento automaticamente resfria o balão. O resíduo da destilação é medido e registrado no programa do destilador. O cálculo de recuperado porcento para evaporado porcento é realizado pelo programa, e impresso.

O recuperado por cento é definido como sendo o volume, em mililitros, de condensado, observado na proveta graduada, em conexão com uma leitura simultânea do termômetro. O evaporado porcento é a soma do recuperado porcento com a perda porcento, que é cem menos a recuperação total porcento (NBR 9619).

O Ponto Final de Ebulição (PFE) é a máxima leitura do termômetro obtida durante o ensaio. Isto ocorre usualmente após a evaporação de todo o líquido do fundo do balão. O Ponto Inicial de Ebulição (PIE) é a leitura do termômetro que é observada no instante em que a primeira gota de condensado cai da extremidade inferior do tubo do condensador (NBR 9619).

O controle estatístico de processo dos cinco destiladores utilizados neste trabalho é apresentado na tabela 3.3. O valor da temperatura da porcentagem do evaporado, para cada destilador, corresponde à média aritmética de 39 medidas em duplicata. Estas medidas foram feitas em dias diferentes e operadores diferentes. O

maior desvio entre os destiladores ocorreu na temperatura dos 10% evaporado e foi de 3,48°C, comparando com a reprodutibilidade do método o valor é aceitável. As demais temperaturas também estão abaixo da reprodutibilidade do método.

Graficamente pode-se ter uma boa visualização do controle estatístico dos ensaios de destilação. A figura 3.2 apresenta o gráfico de controle do destilador número 1 para a temperatura dos 10% evaporados. Como mencionado anteriormente cada ponto foi medido duas vezes e sua média foi plotada no gráfico. A linha central (reta contínua) é a média geral de todos os pontos medidos, a linha contínua acima e abaixo da média geral corresponde ao limite de controle superior e inferior respectivamente. As linhas pontilhadas representam a repetitividade e as linhas tracejadas a reprodutibilidade segundo a Norma ASTM D86.

Tabela 3.2 – Especificações para destilação

Preparação do destilador		**Durante execução do ensaio**	
Condições	**Grupo1 (Gasolina)**	**Condições**	**Grupo1 (Gasolina)**
Balão, mL	125	Temperatura do banho de refrigeração, °C	0 a 1
Termômetro ASTM	7C	Temperatura do banho em torno da proveta, °C	13 a 18
Diâmetro de orifício, mm	38	Tempo decorrido entre o início do aquecimento e o PIE, min.	5 a 10
Temperatura no início do ensaio: Balão e termômetro, °C	13 a 18	Tempo decorrido entre o PIE e os 5% recuperados, s	60 a 100
Proveta e 100 mL de amostra, °C	13 a 18	Taxa média de condensação de 5% recuperados até 5 mL de resíduo no balão, mL/min.	4 a 5
		Tempo decorrido entre os 5 mL residuais e o PFE, min. (max.)	5

Fonte: ASTM D86/NBR 9619

Tabela 3.3 – Controle estatístico de processo da destilação

Destilador	Dados do gráfico de controle			
	T10% (°C)	T50% (°C)	T90% (°C)	Ponto Final (°C)
1	52,99	71,84	171,10	208,96
2	52,85	71,77	169,93	208,94
3	52,30	72,02	170,34	209,11
4	55,78	72,37	172,88	209,63
5	55,19	72,02	173,02	208,66
Média	53,82	72,00	171,45	209,06
Desvio Padrão	1,55	0,23	1,43	0,36
Repetitividade (ASTM)	2,62	2,62	2,18	3,50
Reprodutibilidade (ASTM)	5,64	6,87	4,53	10,50

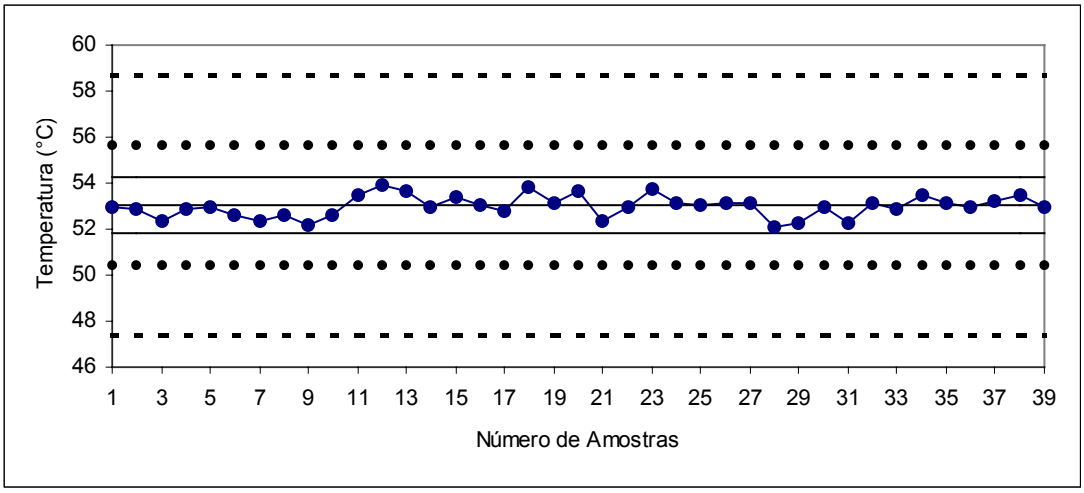

Figura 3.2 – Representação gráfica do controle estatístico de processo para a temperatura dos 10% evaporados do destilador 1.

3.1.2 Teor de Álcool

No Brasil, a adição de álcool etílico anidro combustível (AEAC) à gasolina é feita há muitos anos. Atualmente, adiciona-se 25% ± 1% de álcool anidro na mistura com gasolina. A adição de oxigenados à gasolina deveria ter somente a finalidade de reduzir as emissões tóxicas, porém esta quantidade pode ser alterada pelo Ministério

da Agricultura devido a vários fatores, como por exemplo, a baixa ou alta produção de álcool pelas usinas açucareiras.

A adição de álcool na gasolina já foi muito discutida há alguns anos atrás e existem duas fortes correntes: uma favorável e outra desfavorável. Acredita-se que, em proporções até a faixa de 15% a 20% de álcool em volume, a utilização de misturas de gasolina-álcool traz vantagens apreciáveis. Embora, o problema do uso do álcool anidro é mais uma questão de preço (Campos, 1990).

Uma das vantagens de se misturar álcool à gasolina é que o álcool proporciona o aumento de sua octanagem. E esse aumento depende da composição da gasolina e, principalmente, do nível de octanagem dela. Nos baixos níveis de octanagem, o álcool tem excelente valor de mistura. O álcool tem elevado calor latente de vaporização (205 cal/g) comparado com a gasolina (80 cal/g), o que aumenta o esfriamento no motor, influenciando favoravelmente a resistência à detonação. Outra vantagem é o aproveitamento do excesso de produção de álcool-anidro produzido pelas usinas açucareiras e ainda a redução de poluição ambiental provocada pelas emissões do escape dos veículos (Campos, 1990).

As desvantagens desta adição é que com um teor muito elevado de álcool a eficiência do motor diminui. E se a relação ar/combustível necessária para combustão completa do álcool é de 9:1, e da gasolina é de 15:1, com porcentagens elevadas de álcool na mistura, faz-se necessário ajustar e regular o carburador (ou bico injetor) para operar com a mistura. Até um teor de 15% de álcool não há necessidade de nova regulagem (Campos, 1990).

Nos ensaios para determinação do teor de álcool foram utilizados dois métodos distintos, um de acordo com a norma NBR 13992 (portaria da ANP) que utiliza provetas de vidro graduadas, e o outro que utiliza o analisador portátil de gasolina pelo princípio de espectrofotometria no infravermelho (IV), marca Petrospec, modelo GS1000.

3.1.2.1 Método Proveta (NBR 13992)

As provetas foram calibradas pelo Laboratório de Metrologia do IPT e um fator de correção foi calculado levando em conta o erro do volume da proveta e o coeficiente de expansão volumétrica da gasolina e da solução de cloreto de sódio a 10%. O ensaio consiste em colocar 50 ml da amostra de gasolina na proveta de vidro, adicionar solução aquosa de NaCl (10% p/v) até completar o volume de 100 ml, tampar a proveta e inverter cuidadosamente por dez vezes. Deixar em repouso e aguardar por 15 minutos até a separação completa das duas camadas. Anotar o volume final (ml) da camada aquosa. O teor de álcool presente na gasolina é calculado conforme a equação 3.1:

$$\%\text{álcool} = \left[(A - 50)x2\right] + 1 \qquad (3.1)$$

onde:
A é o volume final da camada aquosa.

O controle estatístico do processo para o ensaio de teor de álcool (tabela 3.4) foi realizado com amostra de gasolina com aproximadamente 25% de álcool. A comparação é realizada com a especificação da portaria da ANP, pois a norma NBR atualmente não possui cálculo de reprodutibilidade e repetitividade.

Tabela 3.4 – Controle estatístico de processo do teor de álcool (proveta)

Item	Dados do gráfico de controle
	Método da proveta-Teor de álcool (%vol)
Média Geral	24.5
Desvio Padrão	0.5
Reprodutibilidade	±1 (portaria)
Repetitividade	-

3.1.2.2 Analisador Portátil

O analisador portátil GS1000 utiliza 17 filtros que selecionam as bandas espectrais de interesse, correspondentes ao MIR (Infravermelho médio), podendo quantificá-las pela quantidade de luz absorvida por cada componente. A determinação das propriedades é realizada comparando os espectros contidos na memória eletrônica do GS1000, e através de modelos matemáticos, prevendo o valor da propriedade de uma outra amostra (Petrospec, 1999).

As propriedades da gasolina que este analisador pode determinar são: Teor de álcool (etanol, metanol e outros oxigenados), composição química, benzeno, MON e RON e destilação (T10% e T50%).

O modelo matemático do analisador portátil (IV) possui um banco de dados original de fábrica. A gasolina nacional se diferencia da gasolina dos Estados Unidos (fabricante do equipamento) devido à origem do petróleo para refino e principalmente pela adição de álcool etílico anidro. Outros fatores, como alterações nas legislações brasileiras, processo de refino do petróleo e abertura do mercado para gasolinas importadas fazem com que este modelo tenha necessidade de constante atualização. Por isso o modelo foi ajustado com amostras de gasolina do programa de monitoramento da qualidade de combustíveis (ANP), e com amostras preparadas pela equipe do Laboratório de Combustíveis e Lubrificantes do IPT (figura 3.3).

O ensaio é rápido e simples, a amostra de gasolina é colocada em frasco de vidro que é rosqueado no mangote do equipamento, o frasco é pressurizado e a amostra é enviada (empurrada) para a célula de amostragem através de um tubo, o equipamento faz a leitura e o resultado é reportado em porcentagem de álcool etílico. O tempo utilizado para limpeza da célula entre uma leitura e outra foi de 120 segundos. O equipamento automaticamente passa a nova amostra a ser analisada pela célula e este tempo é estimado para garantir que nenhum resíduo da amostra anterior interfira no resultado. O tempo total gasto para cada leitura foi de 5 minutos, entre homogeneização da amostra, limpeza, estabilização, leitura e retirada da amostra.

O total de amostras utilizadas no ajuste do modelo foi de 494. O coeficiente de correlação dos pontos (R^2) obtido foi de 0,9969. Pelo gráfico pode-se notar que a concentração maior de pontos está entre os pontos 20% e 26%, que é a variação de porcentagem de álcool normalmente alterada pelo governo . Segundo o fabricante, este equipamento tem um limite de detecção ou calibração que vai até 27%. Apesar do modelo estimar valores acima deste limite, seus resultados não são precisos.

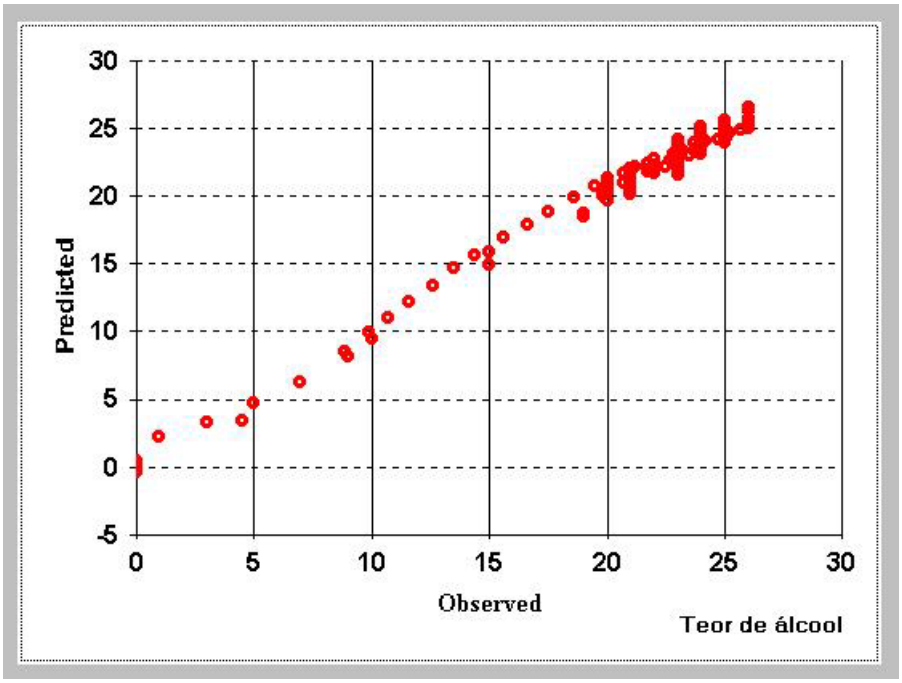

Figura 3.3 - Curva de ajuste do analisador portátil para o teor de álcool (Fonte: IPT/SP)

3.1.3 Massa específica

A determinação de massa específica pode ser feita por diversas técnicas e instrumentos/equipamentos. Dentre os instrumentos/equipamentos mais utilizados para medição desta propriedade, existem os picnômetros, os densímetros vibracionais e os densímetros de imersão.

As normas NBR 7148 e NBR 14065 determinam a análise de massa específica em produtos de petróleo. O método utilizado neste trabalho segue a norma

NBR 14065. Esta norma prescreve a determinação da massa específica e da densidade relativa de destilados de petróleo e óleos viscosos, que podem ser manuseados normalmente como líquidos a temperaturas de ensaio entre 15°C e 35°C.

A massa específica é o quociente da massa de um corpo pelo volume ocupado por esse corpo, a uma dada temperatura e densidade relativa é a relação entre a massa específica de uma substância a uma dada temperatura e a massa específica da água a uma dada temperatura. A massa específica é uma propriedade física fundamental que pode ser usada em conjunto com outras propriedades para caracterizar tanto frações leves, quanto frações pesadas de petróleo e produtos de petróleo (NBR 14065).

Os ensaios foram realizados em dois densímetros digitais, marca Anton Par, modelos DMA 48 e DMA 4500. Os equipamentos foram calibrados com água pura (certificado RBC) na temperatura de ensaio (20°C). O ensaio consiste basicamente em injetar a amostra (gasolina) no tubo em U do densímetro, por meio de uma seringa de plástico (3 mL), tomando-se o cuidado para não formar bolhas no tubo. Após a injeção da gasolina deve-se aguardar a estabilização do equipamento e o resultado será mostrado no mostrador do equipamento. Os resultados são expressos em g/cm^3 e convertidos para kg/m^3 conforme portaria da ANP.

Estes densímetros utilizam o método do tubo de oscilação em formato de U (densimetria vibracional) para fazer medidas de densidade em uma grande faixa de viscosidade e de temperatura. Um único oscilador de referência, somado ao tubo de oscilação em U, dá estabilidade e faz com que ajustes a outras temperaturas além de 20 °C sejam virtualmente desnecessários. Medindo a atenuação causada pela viscosidade da amostra no tudo de oscilação em U, o DMA 4500 corrige automaticamente os erros relativos à influência da viscosidade na medição da densidade. Dois termômetros integrados, Pt 100 de platina, provêm uma alta precisão no controle de temperatura, e são reconhecidos por padrões internacionais (DMA4500, 2003).

O densímetro vibracional é constituído basicamente por um tubo em forma de U e por um sistema de excitação eletrônica. Este sistema provoca um impulso à amostra para que esta oscile no tubo por um certo número de vezes. A quantidade de vezes que a amostra oscila é registrada por um medidor de freqüência. Esta freqüência de oscilação irá variar conforme varia a massa da amostra. Quanto maior for a massa da amostra menor será sua oscilação, logo esta será mais densa. O período de oscilação será convertido em massa específica. A relação entre massa específica e o período de oscilação é estabelecida pelo modelo de *Spring*, conforme mostra a equação 3.2:

$$\rho = (A \times P) - B \qquad (3.2)$$

onde:
P é o período de oscilação,
A e B são as constantes do equipamento.

As constantes A e B, são calculadas a partir de períodos de oscilação observados. Os cálculos das constantes são obtidos quando o tubo em U oscila contendo apenas ar e logo em seguida contendo apenas água bidestilada (figura 3.4 e 3.5 respectivamente). Este procedimento recebe o nome de calibração.

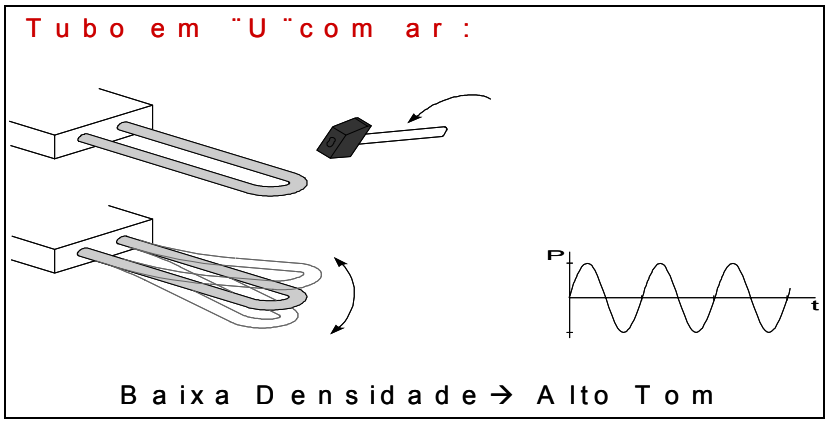

Figura 3.4 – Tubo em U com ar

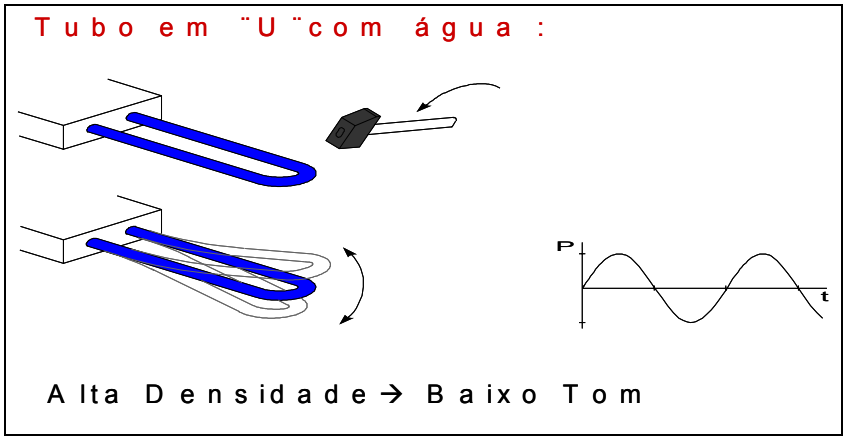

Figura 3.5 – Tubo em U com água

Os valores das constantes são dados pelas seguintes equações conforme a norma NBR 14065:

$$A = \frac{\left[Tw^2 - Ta^2\right]}{dw - da} \tag{3.3}$$

$$B = Ta^2 - \left(A \times da\right) \tag{3.4}$$

Onde:

Tw= período observado da oscilação do tubo contendo água;

Ta= período observado da oscilação do tubo contendo ar;

dw= massa específica da água à temperatura de ensaio;

da= massa específica do ar à temperatura de ensaio.

Os dados do controle estatístico de processo para o ensaio de massa específica são mostrados na tabela 3.5. Segundo a portaria da ANP não existe um valor especificado para este ensaio. A média geral obtida para a amostra utilizada durante o controle, ficou muito próxima entre os dois densímetros. A diferença entre eles foi de 0,1 kg/m^3 o que já satisfaz a repetitividade exigida pelo método, portanto com certeza o valor obtido está dentro da reprodutibilidade do método.

Tabela 3.5 – Controle estatístico de processo (massa específica)

Dados do gráfico de controle (Massa específica)		
	kg/m3	
	DMA 48	DMA 4500
Média Geral	755,7	755,8
Desvio Padrão	0,2	0,3
Reprodutibilidade (ASTM)	0,5	
Repetitividade (ASTM)	0,1	

3.1.4 Benzeno

O teor de benzeno é controlado devido aos riscos à saúde que ele pode causar quando é liberado na atmosfera durante a queima do combustível pelo motor.

Neste ensaio utilizou-se o mesmo analisador portátil GS1000 e o mesmo procedimento de análise citado no capítulo sobre teor de álcool. O ajuste do modelo matemático (figura 3.6) foi realizado com amostras de gasolina do monitoramento da qualidade e matriz de amostras descritas pela norma ASTM D 6277, porém os teores de benzeno das amostras que foram inseridas no equipamento foram obtidos por cromatografia gasosa.

O total de amostras utilizadas no ajuste do modelo foi de 161. O coeficiente de correlação dos pontos (R^2) obtido foi de 0,9822. A maior concentração de pontos no gráfico está entre os pontos 0% e 1%, porque a maioria das gasolinas comercializada possui o teor de benzeno nesta faixa. O limite de detecção do equipamento para esta propriedade é de 5%. Os pontos de 1% a 4% foram adicionados ao modelo utilizando gasolina comum do tipo A fornecida pela Petrobrás. Através de análise cromatográfica o teor de benzeno da gasolina A era encontrado, e a partir deste, quantidades conhecidas de benzeno (P.A.) era adicionada na amostra. Acima de 4 % a curva não apresentou um bom ajuste e esses pontos não foram utilizados.

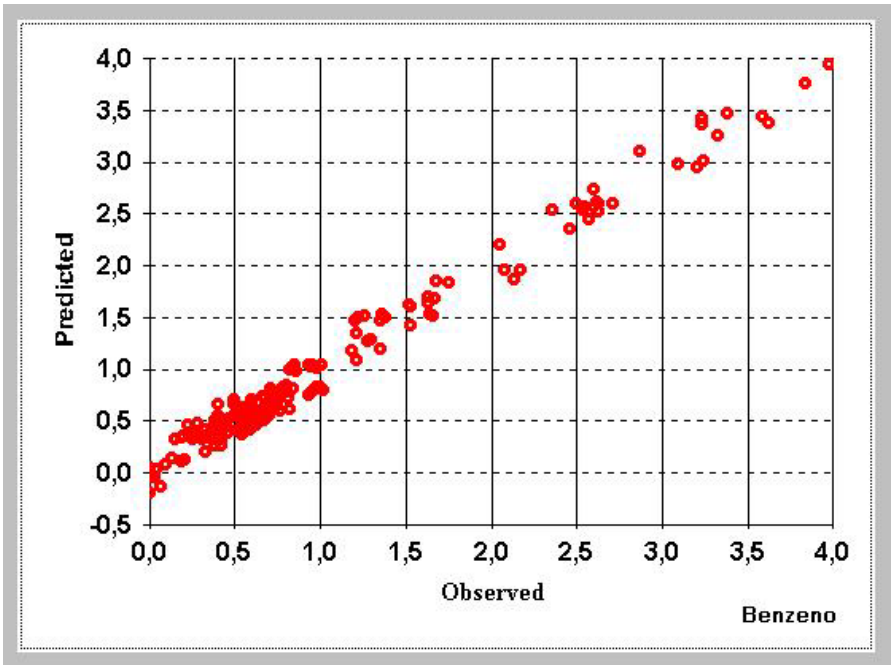

Figura 3.6 - Curva de ajuste do analisador portátil GS1000 para o benzeno (Fonte: IPT/SP)

3.1.5 Octanagem (MON, RON e IAD)

A octanagem da gasolina é uma característica muito importante, porque é uma medida de sua qualidade antidetonante ou capacidade de resistir à detonação. Na combustão de uma gasolina fora de especificação com baixo índice antidetonante, por exemplo, não ocorrerá queima total do combustível e o resíduo (gás) sofrerá aquecimento e compressão pelo pistão do motor, causando auto-ignição e detonação violenta. Esta detonação é percebida no dia a dia como um ruído metálico forte, mais popularmente chamado de "batida de pino". Além de produzir um som indesejável e desperdiçar energia do combustível, a detonação prolongada superaquece as válvulas, velas e pistões encurtando a vida útil do motor.

Para se obter o Índice Antidetonante (IAD) especificado pela portaria da ANP utiliza-se a média aritmética entre o número de octano motor (MON – *motor octane number*), com o número de octano pesquisa (RON – *Research Octane Number*).

O MON e o RON são obtidos em um motor monocilíndrico padrão, com taxa de compressão variável, que foi desenvolvido pelo *Cooperative Fuel Research Committee* (CFR). O RON é determinado em condições relativamente suaves, isto é, temperatura baixa da mistura e velocidade baixa do motor. Já o MON é determinado em condições mais severas, ou seja, temperatura alta da mistura e velocidades relativamente altas.

Estes ensaios são demorados e utilizam padrões de combustíveis caros para regulagem (calibração) do motor, então como alternativa para este trabalho, foi utilizado o analisador portátil GS1000 para obtenção do MON e IAD. Como no álcool e benzeno, foram adicionadas ao modelo matemático amostras de gasolinas com valores conhecidos. Estes valores foram obtidos no motor CFR do Laboratório de Motores do IPT. As figuras 3.7 e 3.8 apresentam os ajustes obtidos para os ensaios de RON e MON respectivamente.

O total de amostras utilizadas no ajuste do modelo foi de 334. O coeficiente de correlação dos pontos (R^2) obtido foi de 0,9667. Apesar de não estar especificado na portaria da ANP o valor mínimo do RON para uma gasolina comum é de 92, ele pode ser estimado utilizando os valores mínimos do MON e IAD. A preparação de amostras com valores conhecidos, para o RON e o MON, não é uma tarefa tão fácil quanto a preparação de amostras contendo álcool e benzeno, por isso a dificuldade em obter pontos na curva abaixo do valor 91.

O total de amostras utilizadas no ajuste do modelo foi de 366. O coeficiente de correlação dos pontos (R^2) obtido foi de 0,9360. O valor do MON especificado na portaria da ANP para o ensaio em motor é de 82. A dificuldade neste caso é obter pontos acima de 86 e abaixo de 80,5. A curva apresenta bastantes pontos entre 80,5 e 82 (mínimo especificado) devido à própria característica das amostras comercializadas.

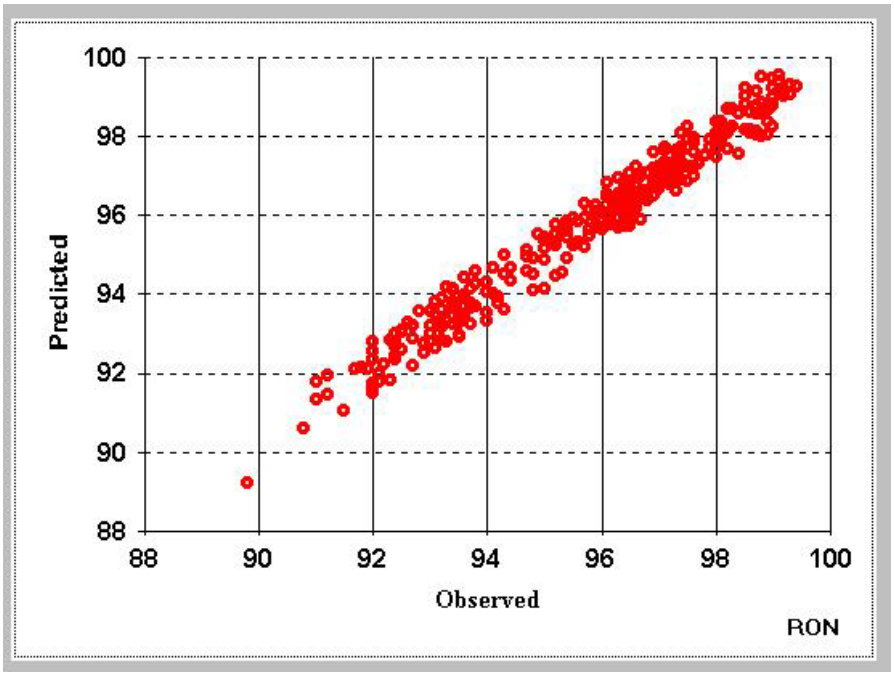

Figura 3.7 - Curva de ajuste do analisador portátil para o RON (Fonte: IPT/SP)

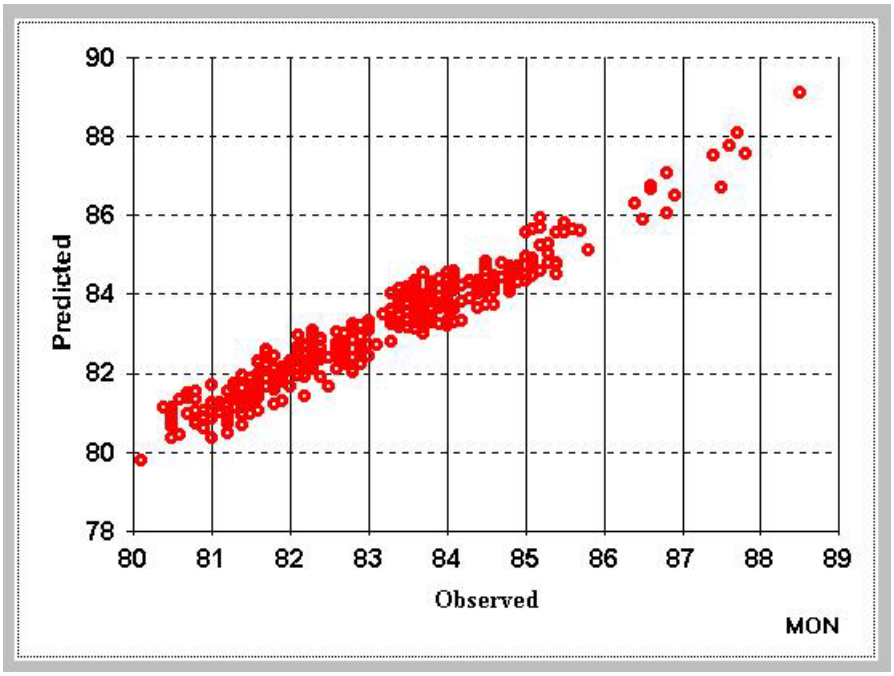

Figura 3.8 - Curva de ajuste do analisador portátil para o MON (Fonte: IPT/SP)

Tanto para o RON como para o MON pode-se observar que o valor do coeficiente de correlação dos pontos da curva não apresentaram um valor de ajuste

tão bom quanto às demais propriedades analisadas pelo GS1000. Os pontos distribuídos na reta possuem uma larga amplitude, isto se deve ao fato de não ser possível a preparação de amostras padrões para serem analisadas no motor CFR e até pela própria variabilidade deste ensaio.

Para o índice antidetonante não é necessário fazer uma curva de calibração, o seu resultado é obtido analogamente ao ensaio do motor CFR.

3.1.6 Espectroscopia por infravermelho próximo - NIR

O infravermelho próximo (NIR) tem-se tornado uma técnica muita popular para uma grande variedade de indústrias nos mais variados tipos de análises sobretudo a partir de meados da década de 90 (Siesler, et al., 2002). A espectroscopia por infravermelho é um método de análise rápido e não destrutivo, ou seja, a amostra não sofre nenhum dano e pode ser reaproveitada.

O infravermelho próximo (NIR) recebe este nome devido a sua proximidade do espectro visível, pois ele corresponde ao intervalo do comprimento de onda de aproximadamente 800 a 2500 nm, ou em números de ondas de 12500 a 4000 cm^{-1} (Siesler, et al., 2002). A figura 3.9 apresenta o espectro eletromagnético dividido em regiões, e como pode ser visto, o infravermelho fica próximo à luz visível.

No infravermelho, a luz interage com a molécula, desequilibrando as suas cargas. O campo elétrico da onda interage com um campo dipolo causado pela distribuição excedente das cargas na molécula. Isto ajuda a explicar porque moléculas simétricas como nitrogênio não podem interagir com a luz.

Como moléculas diferentes na presença de luz do infravermelho têm freqüências diferentes, elas podem ser caracterizadas pela determinação das

freqüências que tenham sido absorvidas. A quantidade da luz absorvida é proporcional à concentração.

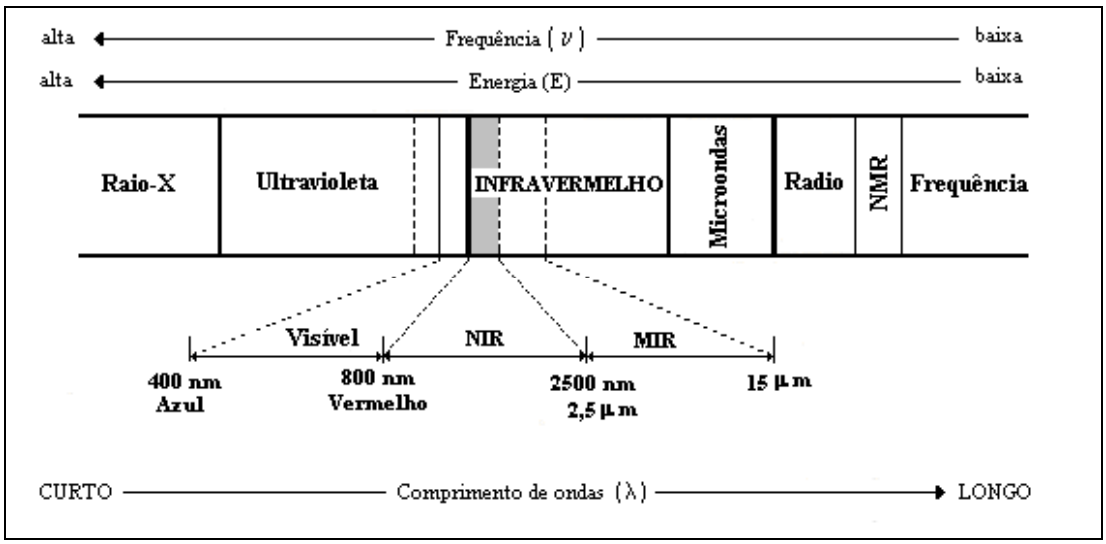

Figura 3.9 – O espectro eletromagnético dividido em regiões

Fonte: Pavia, 1979 e Siesler, 2002

(NIR: Infravermelho próximo, MIR: Infravermelho médio e NMR: Ressonância magnética nuclear)

Nota-se pela figura 3.9 que o comprimento de onda (λ) é inversamente proporcional à freqüência (ν) conforme a relação:

$$\nu = \frac{c}{\lambda} \qquad (3.5)$$

onde:
c é a velocidade da luz.

A equação (3.6) mostra que a energia é diretamente proporcional à freqüência:

$$E = h \cdot \nu \qquad (3.6)$$

onde:

h é a constante de Planck.

Em geral, os dados NIR podem ser interpretados segundo a lei de Beer, a qual assume que existe uma relação linear entre a concentração molar (*Y*) e a absorbância $A = -log(I/I_0)$ desta substância. O termo I_0 indica a intensidade espectral da radiação emitida pelo espectrômetro e *I* é a intensidade remanescente após a interação com a amostra. A equação (3.7) apresenta a relação entre a concentração e a absorbância e com a absorvidade molar (ε) e o caminho óptico (*d*).

$$A = \varepsilon \times d \times Y \tag{3.7}$$

Como o caminho ótico *d* é dado para um determinado sensor, se a absortividade molar ε pudesse ser encontrada na literatura, a concentração poderia ser obtida diretamente (equação 3.8) assumindo que somente o composto a ser quantificado absorve radiação na mistura.

$$Y = \frac{A}{\varepsilon \times d} \tag{3.8}$$

Na tabela 3.6, estão resumidas as regiões do espectro e os tipos de energia de transição. Várias dessas regiões, incluindo o infravermelho, oferecem informações vitais sobre as estruturas de moléculas orgânicas (Pavia,1979).

Como dito anteriormente as bandas do NIR possuem relativamente baixa absortividade molar (baixa sensibilidade) e são largas, em decorrência disto possuem características de sobretons e bandas de combinação. Os sobretons e as bandas de combinação são o coração da espectroscopia NIR, e é a não-harmonicidade que determina a ocorrência e as propriedades espectrais como, freqüência e intensidade, das bandas do NIR (Siesler, et al. 2002).

Tabela 3.6 - Tipos de Transições de Energia e cada Região do Espectro Eletromagnético.

Região do Espectro	Energia de Transição
Raio-X	Quebrando ligação química
Ultravioleta/Visível	Eletrônica
Infravermelho	Vibracional
Microondas	Rotacional
Radiofreqüência	Rotação Nuclear (Ressonância Magnética Nuclear)
	Rotação do Elétron (Ressonância de Rotação do Elétron)

(Fonte: Pavia, 1979)

As ligações químicas com elevada não-harmonicidade são aquelas envolvendo átomos leves, como o hidrogênio. Estas ligações vibram em alta energia e com uma larga amplitude quando sofrem movimentos de estiramento, absorvendo assim, a maior parte da intensidade. Deste modo, a região espectral NIR é dominada pela absorção associada com grupos funcionais XHn. Sendo que essas absorções são resultados dos sobretons de combinações de bandas fundamentais envolvendo o modo de estiramento e o dobramento de vibração de tais grupos. Alguns sobretons e combinações de bandas ocorrem com suficiente regularidade para caracterizar grupos moleculares, da mesma forma como é caracterizada a banda fundamental da região do infravermelho médio-MIR (Siesler, et al. 2002 e Pavia, 1979).

Em hidrocarbonetos alifáticos, o primeiro conjunto de combinações de bandas ocorre entre 2000 e 2400 nm, os primeiros sobretons entre 1600 e 1800 nm e os segundos sobretons entre 1000 e 1200 nm. Grupos olefínicos dão origem a bandas específicas de absorção NIR em 1620 e 2100 nm para o grupo vinil e 1180, 1680, 2150, 2190 nm para cis-olefinas. Ligações CH aromáticas têm o primeiro e o segundo sobretons localizados em 1685 e 1143 nm respectivamente. Os primeiros e segundos sobretons das vibrações do estiramento O-H em álcoois estão localizados aproximadamente em 1400 e 1000 nm respectivamente.

A interpretação dos espectros NIR é realizada de maneira análoga à dos espectros do MIR, porém devido à baixa absortividade molar dos componentes, que muitas vezes se sobrepõem, os espectros NIR não são facilmente interpretáveis, havendo a necessidade da aplicação de métodos de calibração multivariada para extrair a informação desejada contida neles (Carrillo Le Roux e Sotelo, 2004). Já as bandas espectrais no MIR são consideradas bandas fundamentais, pois seus picos são específicos, nítidos e sensíveis. Embora os dois métodos espectroscópicos forneçam informação vibracional, cada um tem suas próprias vantagens e desvantagens que devem ser consideradas na análise quantitativa (Chung et al, 1999).

3.1.6.1 Operação do espectrômetro de FT-IR

Neste trabalho foi utilizado o espectrômetro marca ABB Bomem, modelo FTLA2000-160, os espectros foram coletados inicialmente em toda a faixa spectral NIR (14000 a 4000 cm^{-1}), em intervalos de 4 cm^{-1} com 46 varreduras. A base deste espectrômetro é um interferômetro FT-IR (Transformada de Fourier). O interferômetro FT-IR consiste basicamente de um espelho fixo, um espelho em movimento e um *beamsplitter* (divisor ótico), conforme esquema da figura 3.10. O interferograma armazena as informações do espectro e este é interpretado usando a transformada de Fourier. Aplicando a transformada de Fourier ao interferograma obtêm-se espectros "crus" (ver figura 3.10). O espectro "cru" é um gráfico de intensidade de luz sobre o detector em função da freqüência ótica. Este tipo de espectro contém informações não apenas da amostra presente no compartimento ou acessório de amostragem, mas também sobre todo o instrumento, inclusive da fonte de todos os componentes óticos, do ar ambiente e de qualquer contaminação que estiver presente no caminho ótico (Manual FTLA2000, 2002).

A figura 3.10 mostra que a luz da fonte do infravermelho entra no interferômetro e se divide entre dois raios iguais pelo *beamsplitter*. Um raio é refletido no espelho fixo, que reflete na parte inferior do *beamsplitter*. O outro raio é

transmitido para o espelho em movimento que também reflete na parte inferior do *beamsplitter*.

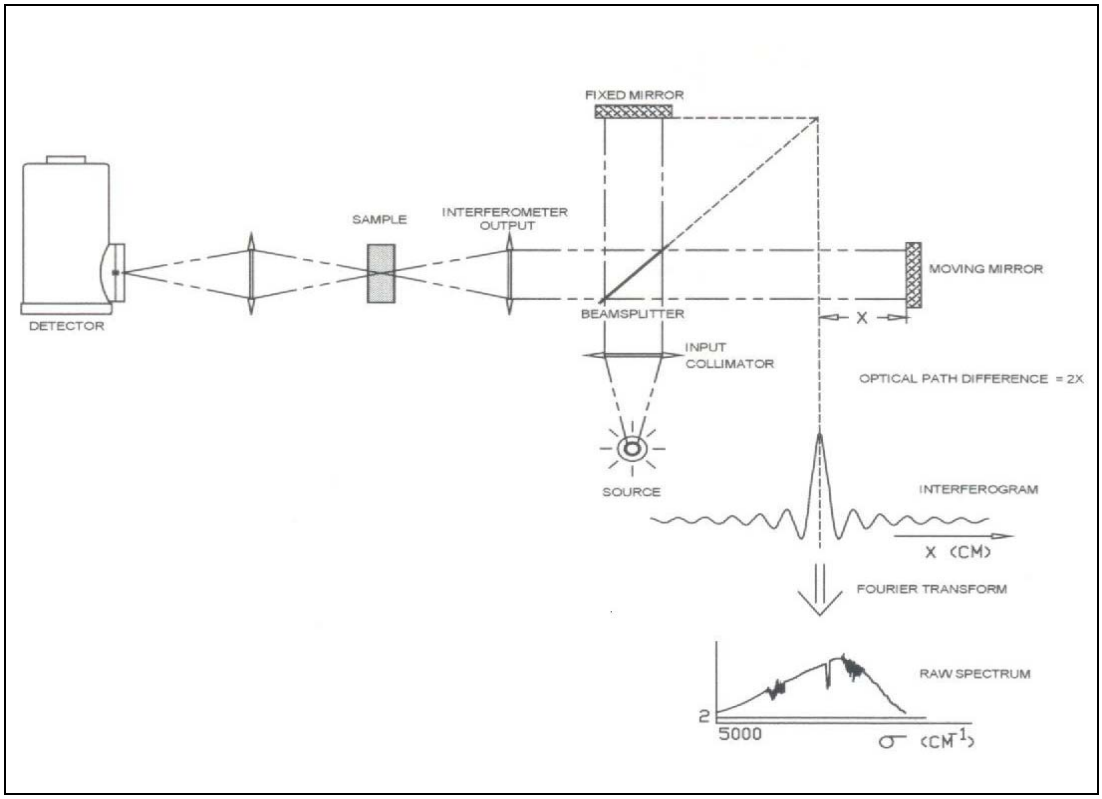

Figura 3.10 – Princípio de operação (Fonte: FTLA2000, 2002)

O movimento do espelho introduz continuamente um caminho ótico diferente entre os dois espelhos (ver o 2x na figura 3.10). Com o movimento do espelho, os dois raios refletidos interferem com caminhos diferentes, isto cria variações de intensidade.

Em certos pontos destes caminhos óticos, a interferência é construtiva para algumas freqüências e destrutiva para outras. Isto ocorre porque a diferença do caminho ótico varia constantemente, e várias freqüências presentes no raio incidente são moduladas em diferentes razões. A razão de modulação de cada freqüência é proporcional à freqüência ótica.

Depois de deixar o interferômetro, a luz modulada passa através da amostra. Se as amostras contiverem moléculas assimétricas estas serão absorvidas

pela radiação do infravermelho em freqüências específicas. A luz restante atinge o detector que a converte em um sinal elétrico (Manual FTLA 2000-2002).

3.1.7 Marcador

A Portaria ANP Nº 274, de 01/11/2001 *estabelece a obrigatoriedade de adição de marcador a solventes e a derivados de petróleo eventualmente indicados pela ANP bem como a proibição da presença de marcador na gasolina*.

Segundo a portaria, marcador é uma substância que permite, através dos métodos analíticos estabelecidos pela ANP, a identificação de sua presença na gasolina e que, ao ser adicionada aos PMC (Produtos de Marcação Compulsória), em concentração não superior a 1ppm não altere suas características físico-químicas, e não interfira no grau de segurança para manuseio e uso desses produtos.

Como se trata de um assunto sigiloso e de grande interesse comercial, a metodologia utilizada neste trabalho para identificação da presença de marcador na gasolina foi descrita de forma sucinta.

Na identificação do marcador foi utilizada a técnica de cromatografia em fase gasosa com um equipamento da marca Shimadzu, modelo GC-17A. Para a calibração do cromatógrafo foi utilizado um padrão do marcador, e foram construídas curvas de calibração em concentrações adequadas. O teor de marcador na gasolina é determinado a partir de padronização externa, analisando volumes idênticos de gasolina e dos padrões utilizados nas curvas de calibração.

As amostras a serem analisadas são colocadas no amostrador, modelo AOC-20i, marca Shimadzu, e injetadas automaticamente no cromatógrafo. A amostra passa através da coluna cromatográfica capilar (apropriada para este tipo de análise) e a presença do marcador é indicada pela constatação de um pico característico, no

mesmo tempo em que o marcador elui. A concentração em ppb é determinada pela área do pico obtida no cromatograma e calculada através da curva de calibração.

3.2 Métodos Estatísticos Multivariáveis

A análise estatística multivariável preocupa-se com dados que consistem de múltiplas medidas para um grande número de indivíduos, objetos ou amostras de dados. Os dados dos espectros NIR, usados neste trabalho, são dados multivariáveis em que as variáveis (comprimento de ondas) têm uma forte correlação com outros comprimentos de ondas. O tratamento estatístico usado para resolver problemas analíticos tem sido o maior fator no crescimento do interesse pela espectroscopia NIR (Siesler et al., 2002). Neste item serão apresentadas as técnicas para a classificação e para a compressão dos dados.

3.2.1 Classificação

A análise discriminante é a técnica estatística apropriada para classificar variáveis dependentes de categorias nominais e variáveis independentes métricas. Em muitos casos a variável dependente consiste de dois grupos ou classificações, por exemplo, masculino e feminino ou alto e baixo. Quando dois grupos de classificação estão envolvidos na análise, o problema de classificação é comumente chamado de análise discriminante de dois grupos (binária). Quando três ou mais grupos de classificação estão envolvidos a técnica é conhecida como análise discriminante múltipla (MDA – *multiple discriminant analysis*) (Hair et al. 1998).

A análise discriminante é uma técnica estatística cuja finalidade é testar a hipótese de que a média de um grupo de variáveis independentes é igual à de dois ou mais grupos. Basicamente ela consiste em multiplicar cada variável independente por pesos (*scores*) e em somá-las. A média do grupo é calculada pelos *scores* discriminantes individuais.

A significância estatística da função discriminante depende da distância entre as médias dos grupos discriminados. A comparação é realizada pela distribuição dos *scores* discriminantes nos grupos. Se a sobreposição da distribuição entre os grupos é pequena, a função discriminante separa bem os grupos (figura 3.11a). Se a sobreposição é grande na distribuição dos *scores* discriminantes, então a função discriminante não separa bem os grupos (figura 3.11b).

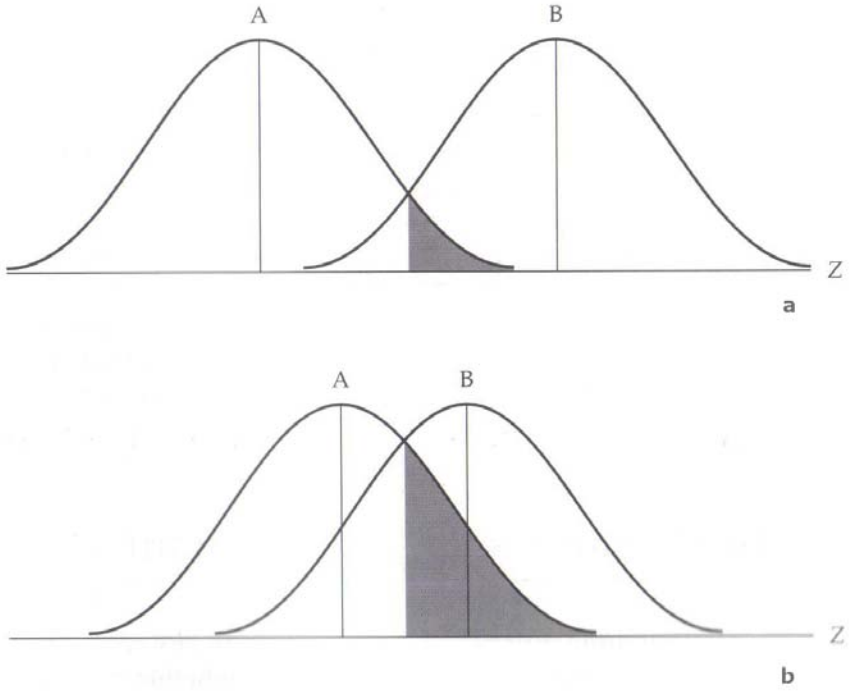

Figura 3.11 – Distribuição dos *scores* discriminantes dos grupos A e B.

(a) representa uma separação boa dos grupos. (b) representa uma separação ruim dos grupos

(Fonte: Hair, et al., 1998).

A figura 3.12 apresenta graficamente a análise discriminante entre dois grupos A e B, e duas medidas V_1 e V_2. As elipses ao redor dos pontos menores (grupo B) e pontos maiores (grupo A) representa a proporção de objetos (normalmente 95% de intervalo de confiança) dentro de cada grupo. As distribuições A' e B' no eixo Z representa a informação condensada sobre diferentes grupos dentro de um conjunto de pontos (Z *scores*).

Resumindo, em uma análise discriminante, uma combinação linear de variáveis independentes é obtida, resultando numa série de *scores* discriminantes para cada objeto em cada grupo. Estes *scores* (ou limites entre classes) são calculados de acordo com uma regra estatística de maximização de variância entre grupos e minimização de variância dentro deles (Hair, et al., 1998, Otto, 1999 e Sharma, 1996). Se a variância entre grupos é relativamente grande em comparação com a variância dentro dos grupos, então a função discriminante separa bem os grupos. A performance de um processo de discriminação ou análise discriminante para a identificação de características de uma mistura química utilizando espectros NIR depende da combinação de dois fatores: das características da extração de informação dos espectros (item 3.2.2) e do método de classificação.

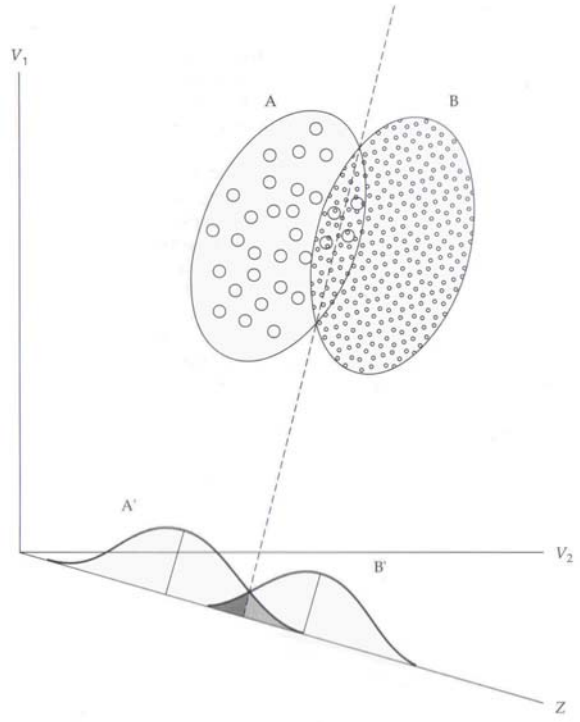

Figura 3.12 – Análise discriminante entre dois grupos A e B. A' e B' são as distribuições dos grupos, V1 e V2 são as medidas e Z é o eixo dos *scores* discriminantes (Fonte: Hair, et al., 1998).

Na classificação de objetos desconhecidos (ou uma amostra nova) suas características são inseridas no modelo de função discriminante e transformadas em coordenadas da mesma forma que foi feito para o conjunto de dados originais. Então

o objeto é classificado no grupo ao qual o centróide (centro de uma determinada classe) tem a menor distância (Otto, 1999).

Os algorítmos dos classificadores LDA e QDA foram implementados em Matlab e seguiram o procedimento descrito por Wu, et al., 1996. Daqui em diante convenciona-se que cada amostra é designada como x_i e é baseada em d medidas diferentes $x_i=(x_{i1},...,x_{id})^T$ para cada tipo de classe **K**.

3.2.1.1 QDA – Análise Discriminante Quadrática

O QDA recebe este nome porque baseado em uma regra de classificação quadrática, o limite de classes é uma forma quadrática. Os *scores* de classificação são calculados da seguinte forma:

$$cf_k(x_i) = (x_i - \mu_k)^T \sum_k^{-1} (x_i - \mu_k) + \ln\left|\sum_k\right| - 2\ln \pi_k \qquad (3.9)$$

onde:

Σ_k é a matriz de covariância de classe k,

μ_k é a média dos vetores de classe k, e

π_k é a probabilidade da classe k.

A covariância, a média do vetor e a probabilidade foram estimadas através das seguintes equações:

$$\hat{\sum}_k = 1/n_k \sum_{i=1}^{n_k} (x_i - \mu_k)(x_{i-}\mu_k)^T, \qquad (3.10)$$

$$\hat{\mu}_k = 1/n_k \sum_{i=1}^{n_k} x_i, \qquad (3.11)$$

$$\hat{\pi}_k = n_k / n, \qquad (3.12)$$

onde:

n_k é o número de objetos dentro da classe k, e

n é o total de objetos do conjunto de treinamento (calibração).

Na equação (3.9), existem três termos, e o primeiro termo corresponde à distância de Mahanalobis.

3.2.1.2 LDA – Análise Discriminante Linear

No caso do LDA, as matrizes de covariância de cada classe são consideradas iguais entre si. Então, a matriz de covariância de classe k na equação (3.9) é substituída pela matriz de covariância combinada (equação 3.13).

$$\sum_{\text{combinada}} = 1/n \sum_{k=1}^{k} n_k \sum_k \tag{3.13}$$

Omitindo os termos constantes, os *scores* de classificação para LDA são calculados da seguinte forma:

$$cf_k(x_i) = (x_i - \mu_k)^T \sum_{\text{combinada}}^{-1} (x_i - \mu_k) - 2\ln \pi_k \tag{3.14}$$

Quando a probabilidade π_k a priori é constante, a equação acima corresponde à distância de Mahanalobis. Segundo Johnson, et al. (1998) quando se comparam duas classes com $\Sigma_1=\Sigma_2$ e com $\pi_1 = \pi_2$, pode ser demonstrado que utilizar a equação (3.14) é equivalente a comparar:

$$\Phi(x_i) = (\mu_1 - \mu_2)^T \sum_{\text{combinada}}^{-1} (x_i) \tag{3.15}$$

com o número:

$$\frac{1}{2}(\mu_1 - \mu_2)^T \sum_{\text{combinada}}^{-1} (\mu_1 + \mu_2) \tag{3.16}$$

A equação (3.15) é linear em x_i, daí o nome de LDA.

Dentro de um conjunto de dados "fraco", no qual o número de variáveis é comparável ao número de parâmetros a ser estimado, a estimativa da matriz de covariância torna-se altamente incerta e os pequenos autovalores têm desvios baixos e os grandes autovalores estimados tem desvios elevados. Para conjuntos mal condicionados, onde os números de variáveis são elevados quando comparados ao número de objetos do grupo (n_k) e baixos quando comparados ao número total de objetos (n), a QDA não pode ser aplicada, porque a matriz de covariância de classe k ($\sum_k$) é singular. Se o número de variáveis é maior do que o número total de objetos (n), a QDA e a LDA não podem ser usadas, porque $\sum_k$ e $\sum_{combinada}$ seriam singulares.

3.2.1.3 Critério para classificação das classes

A classificação de cada amostra, tanto pela QDA quanto pela LDA, é definida comparando-se os *scores* de classificação, aquele que tem menor valor fornece a classe de classificação daquela amostra.

3.2.2 Compressão de Dados

3.2.2.1 PCA – Análise em Componentes Principais

Os dados espectroscópicos, após o pré-tratamento, são introduzidos no PCA sendo então que os *scores* dos componentes principais (PC) podem ser utilizados como variáveis de entrada para os métodos de classificação.

Usualmente, quando se trata de dados de dimensão elevada, a primeira etapa na análise dos dados é a redução da sua dimensionalidade. Há diferentes razões para isto, como a dificuldade na interpretação e a visualização direta de suas estruturas. As variáveis redundantes criam problemas computacionais. A ferramenta mais usual

para resolver estes problemas é a PCA. A idéia principal da PCA é a projeção dos dados de um espaço de alta dimensão em um espaço de dimensão pequena. Se a compressão dos dados for suficiente, o número elevado de variáveis é substituído por um número pequeno de fatores latentes os quais podem ser suficientes para explicar a estrutura dos dados. Os fatores latentes, também chamados de componentes principais (PCs) são obtidos pela maximização da variância dos dados projetados, em outras palavras, pelo cálculo dos autovalores da matriz de covariância das amostras (Stanimirova, 2004).

A PCA encontra combinações de variáveis ou fatores, que descrevem as maiores tendências dos dados. Matematicamente, a PCA depende da transformação dos dados originais para novos sistemas de coordenadas, ou seja, da decomposição da matriz original **X** a partir da matriz de covariância ou da matriz de correlação das variáveis de processo (Otto, 1999). A estrutura dos dados pode então ser visualizada diretamente em um gráfico que corresponde à projeção dos objetos no espaço definido pelos PCs selecionados. A classificação pode ser obtida também pela análise dos primeiros PCs (ex.: plotando PC1 x PC2) que explicam a maior parte da variância dos dados originais.

O software PLS_Toolbox do Matlab^{TM} calcula a matriz de covariância, os *scores* e *loadings* e seus resultados podem ser obtidos numericamente ou graficamente. Basicamente, o cálculo pelo software é realizado como descrito a seguir (equações 3.17 a 3.23).

Para uma dada matriz **X** com *n* linhas e *m* colunas, com cada variável iniciando uma coluna e cada amostra uma linha, a matriz de covariância **X** é definida como:

$$\mathrm{cov}(\mathrm{X}) = \frac{\mathrm{X}^{\mathrm{T}}\mathrm{X}}{\mathrm{n}-1} \qquad (3.17)$$

As colunas de **X** foram normalizadas, sendo centradas na média, cada coluna é ajustada para média zero pela subtração da média original de cada coluna. A PCA decompõe os dados da matriz **X** como a soma do produto externo de vetores $\mathbf{t_i}$ e $\mathbf{p_i}$ mais uma matriz residual ***E*** (PLS_Toolbox, 2002).

$$X = t_1 p_1^T + t_2 p_2^T + ... + t_k p_k^T + E \qquad (3.18)$$

Aqui **k** é menor ou igual à menor dimensão de **X**, por exemplo, $\mathbf{k} \leq \min\{n, m\}$.

Outra forma, análoga a esta, apresentada por Otto (1999) diz que a PCA aproxima a matriz **X** pelo produto de duas matrizes pequenas:

$${}_n X^m = {}_n T^d {}_d P'^m \qquad (3.19)$$

onde **X** é a matriz original com *n* linhas (objetos) e *m* colunas (características das variáveis); **T** é a matriz *scores* com *n* linhas e *d* colunas (números de componentes principais) e **P'** é a matriz *loading* (transposta) com *d* colunas e *m* linhas. Os componentes principais (*scores* **T**) podem ser considerados como projeções dos dados da matriz original, **X**. Para isto os *scores* são convertidos para o lado esquerdo da equação (3.19), conforme descrito a seguir:

$${}_n T^d = {}_n X^m {}_d P^m \qquad (3.20)$$

Na decomposição do PCA, os vetores $\mathbf{p_i}$ são autovetores da matriz de covariância, então para cada $\mathbf{p_i}$ temos:

$$\mathrm{cov}(X)\, p_i = \lambda_i p_i \qquad (3.21)$$

onde,

λ_i é o autovalor associado com o autovetor $\mathbf{p_i}$

O vetor $\mathbf{t}_i$ forma um conjunto ortogonal ($\mathbf{t}_i^T\mathbf{t}_j=0$ para $i \neq j$), enquanto que $\mathbf{p}_i$ é ortonormal ($\mathbf{p}_i^T\mathbf{p}_j=0$ para $i \neq j$, $\mathbf{p}_i^T\mathbf{p}_j=1$ para $i=j$). Para $\mathbf{X}$ e algum par de $\mathbf{t}_i$, $\mathbf{p}_i$ tem-se:

$$Xp_i = t_i \tag{3.22}$$

As novas coordenadas são combinações lineares das variáveis originais. Desta forma a determinação dos componentes principais é descrito a seguir:

$$\begin{aligned} t_{11} &= x_{11}p_{11} + x_{12}p_{21} + \ldots + x_{1m}p_{m1} \\ t_{21} &= x_{21}p_{11} + x_{22}p_{21} + \ldots + x_{2m}p_{m1} \\ &\vdots \\ t_{n1} &= x_{n1}p_{11} + x_{n2}p_{21} + \ldots + x_{nm}p_{m1} \end{aligned} \tag{3.23}$$

Os vetores $\mathbf{t}_i$ são conhecidos como *scores* e contem informações de como as *amostras* se relacionam com os componentes principais. Os vetores $\mathbf{p}_i$ são conhecidos como *loadings* (ou componentes principais) e contêm informações de como as *variáveis* se relacionam entre si (Sharma, 1996). Os vetores $\mathbf{t}_i$ (*scores*) são combinações lineares das variáveis da matriz original **X,** com coeficientes definidos por $\mathbf{p}_i$. Outra maneira de interpretar a PCA é imaginar que $\mathbf{t}_i$ é a projeção de $\mathbf{X}$ em $\mathbf{p}_i$. Os pares de $\mathbf{t}_i$, $\mathbf{p}_i$ são arranjados em ordem decrescente de acordo com o autovalor associado. O autovalor (λ_i) é uma medida da quantidade de variância descrita pelo par $\mathbf{t}_i$, $\mathbf{p}_i$. Neste contexto pode-se pensar na variância como informação. Por causa do par $\mathbf{t}_i$, $\mathbf{p}_i$ estarem em ordem decrescente de λ_i, o primeiro par captura a maior quantidade de informação de algum par da decomposição (PLS_Toolbox, 2002).

A interpretação dos resultados da análise dos componentes principais é usualmente realizada pela visualização dos *scores* e *loadings* (Otto, 1999, Khattree, et al., 2000, Johnson, et al., 1998, Sharma, 1996, Martens, et al., 1989). Um exemplo de interpretação geométrica, chamada de *biplot*, é ilustrada na figura 3.13. Quando apresentadas em três dimensões (três variáveis), é fácil de se notar que todas as amostras formam um plano e podem ser fechadas por uma elipse. É aparente também, que as amostras variem mais ao longo de um eixo da elipse do que de outro.

O primeiro componente principal (PC) descreve a direção da maior variação do conjunto de dados, o qual é o maior eixo da elipse. O segundo PC descreve a segunda maior variação (menor eixo da elipse). Neste caso, um modelo PCA (*scores*, vetores *loadings* e autovalores associados) com dois componentes principais descreve adequadamente todas as variações na medição.

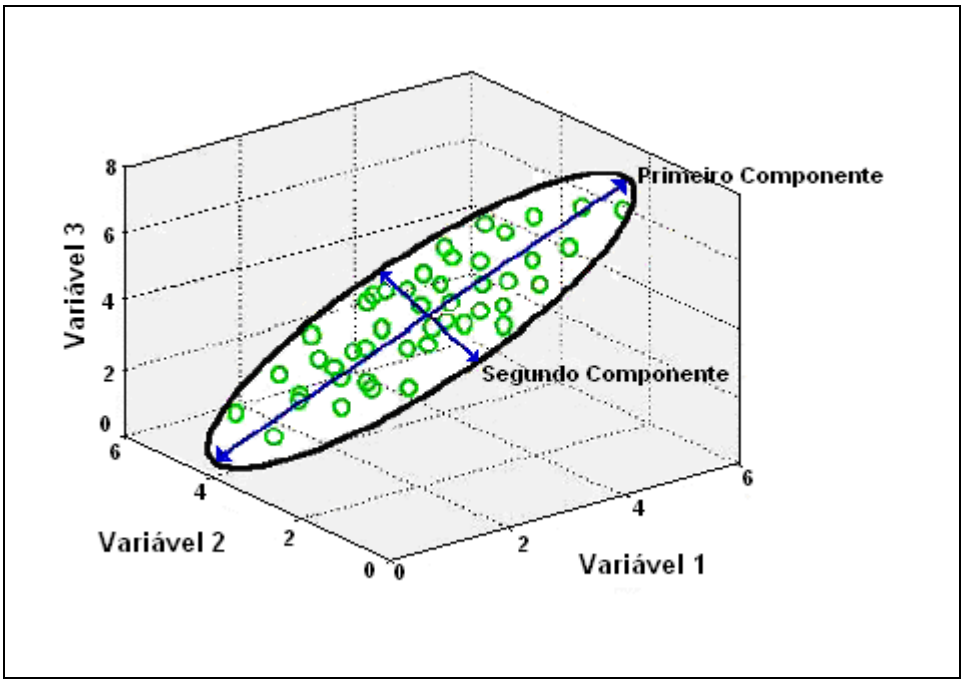

Figura 3.13 – Representação gráfica da PCA

3.2.2.2 PLS – Mínimos Quadrados Parciais

De maneira semelhante à PCA, o PLS também tem a função de redução de dimensionalidade de dados, porém o algoritmo PLS, além dos dados espectrais **X**, requer um segundo conjunto de dados de entrada: uma matriz de variáveis dependentes. Caso exista somente uma variável de interesse, o segundo conjunto é representado por um vetor **y** (n x 1); caso existam $s > 1$ variáveis dependentes, então uma matriz **Y** (n x s) é requerida. Normalmente **X** e **y** ou **Y** têm suas médias centradas como primeira etapa (Kemsley, 1996).

Portanto, o PLS é um método de regressão baseado em um modelo bilinear em relação aos objetos e variáveis das matrizes **X** e **Y**. A idéia do PLS é encontrar fatores que consigam capturar a variância dos dados originais e também obter correlação com a variável dependente. Isto pode ser conseguido maximizando a covariância, ou seja, as variáveis latentes da direção da matriz **X** são modificadas para que a covariância entre ela e a matriz **Y** seja maximizada (PLS_Toolbox, 2002, Otto, 1999, Martens, et al., 1989).

Os dados espectroscópicos, após o pré-tratamento, são introduzidos no PLS sendo então que os *scores* das variáveis latentes (LV) podem ser utilizados como variáveis de entrada para os métodos de classificação. A classificação pode ser obtida também pela análise dos primeiros LVs (ex.: plotando LV1 x LV2) que explicam a maior parte da variância dos dados originais. Assim, as variáveis latentes têm uma interpretação análoga aos componentes principais da PCA.

Apesar do PLS ser muito utilizado como método de regressão multivariável, os seus *scores* ou variáveis latentes podem ser usadas para estudar a classificação de amostras de gasolinas. Nesta metodologia é aplicada a PCA às amostras (resultados NIR) para a redução das variáveis e cálculo dos componentes principais. Os componentes principais são utilizados para se fazer a análise de classificação (LDA e QDA).

Segundo Otto (1999) ambas matrizes **X** e **Y** são decompostas em pequenas matrizes conforme as equações (3.24) e (3.25):

$$ {}_{n}X^{m} = {}_{n}T^{d}\,{}_{d}P'^{m} + {}_{n}E^{m} \qquad (3.24) $$

$$ {}_{n}Y^{p} = {}_{n}U^{d}\,{}_{d}Q'^{p} + {}_{n}F^{p} \qquad (3.25) $$

onde **X** é a matriz das variáveis independentes com n linhas (objetos) e m colunas (características das variáveis espectrais); **Y** é a matriz das variáveis dependentes com n linhas (objetos) e m colunas (propriedades); **T** e **U** são as matrizes *scores* com n

linhas e d colunas ortogonais; e **P** é a matriz d x m *loading* (transposta) da matriz **X**; **E** é a matriz *n* x *m* dos erros residuais da matriz **X**; **Q** é a matriz *d* x *p* loading (transposta) da matriz **Y**; **F** é a matriz *n* x *p* dos erros residuais da matriz **Y.**

O significado e estimativa da matriz peso **W** podem ser entendidos a partir do algoritmo que é apresentado a seguir (PLS_Toolbox, 2002, Otto, 1999, Martens, et al., 1989): Primeiramente, centralizam-se as variáveis:

$l = 0$

$$X = X_{original} - \bar{x} \tag{3.26}$$

$$Y = Y_{original} - \bar{y} \tag{3.27}$$

onde: $\bar{x}$ e $\bar{y}$ são os vetores colunas médios das respectivas matrizes.

As próximas dimensões l = 1 a l = d são calculadas com base em um conveniente critério de parada, usualmente o erro padrão de predição é utilizado. O *loop* para o número de dimensões é $l = l$+1.

Os *scores*, *loadings* e pesos são estimados iterativamente, como mostrado nas etapas a seguir com o algoritmo NIPALS. A iteração é interrompida se a precisão requerida é atingida.

(a) A primeira coluna da matriz **Y** atual é usada como um vetor início para o *y-score* vetor *u*:

$u = y_1$

(b) Os pesos de **X** são estimados:

$$\mathrm{w}^{\mathrm{T}} = \frac{\mathrm{u}^{\mathrm{T}}\mathrm{X}}{\mathrm{u}^{\mathrm{T}}\mathrm{u}} \tag{3.28}$$

(c) Os pesos são escalonados para um vetor de comprimento um:

$$w^T = \frac{w^T}{\left(w^T w\right)^{1/2}} \tag{3.29}$$

(d) Os *scores* da matriz **X** são estimados:

$$t = Xw^T \tag{3.30}$$

(e) Os *loadings* da matriz **Y** são estimados:

$$q^T = \frac{t^T Y}{t^T t} \tag{3.31}$$

(f) O *y-score* vetor *u* é gerado:

$$u = \frac{Yq}{q^T q} \tag{3.32}$$

Nesta etapa o vetor *u* (da iteração passada) é comparado com o *u* (novo).
Se $//\, u(antigo) - u(novo) \,// < // \, u(novo) \,// \times tolerância$, a convergência é obtida, caso contrário a iteração é re-iniciada na etapa (a). Uma vez convergido, utiliza-se o valor de *u* obtido na etapa (f), para calcular:

(g) O coeficiente de regressão *b* para a correlação é calculado:

$$b = \frac{u^T t}{t^T t} \tag{3.33}$$

(h) Os *loadings* da matriz **X** são estimados:

$$p^T = \frac{t^T X}{t^T t} \tag{3.34}$$

(i) Finalmente, os novos resíduos das matrizes **X** e **Y** podem ser calculados:

$$E = X - btp \tag{3.35}$$

$$F = Y - btq \tag{3.36}$$

Estes valores são substituídos a **X** e **Y**, fazendo-se $l = l+1$ retoma-se a partir de (a), e calcula-se a variáveis latentes seguintes.

3.2.4 Redução de Dimensionalidade e Discriminação

Nesta metodologia são aplicadas a PCA ou a PLS às amostras (resultados do NIR) para a redução das variáveis é feito o cálculo dos *scores* (componentes principais ou variáveis latentes) e os *loadings* correspondentes a cada amostra. Os *scores* correspondentes a cada amostra são utilizados para se fazer a análise de classificação (LDA e QDA), podendo se reter um número reduzido de variáveis latentes.

3.2.5 Quadro Resumo

Alguns dos métodos estatísticos multivariáveis utilizados neste trabalho foram implementados em Matlab™. O PLS_Toolbox (versão 3.0, 2002, *eigenvectors Inc.*) para uso com o Matlab™ foi utilizado para a Análise dos Componentes Principais (PCA) e dos Mínimos Quadrados Parciais (PLS).

Matrizes de dados foram geradas com os resultados obtidos nas análises físico-químicas, marcador e espectros NIR. Utilizando as especificações da Portaria

da ANP foi possível classificar as amostras ensaiadas em duas classes distintas: gasolina conforme e gasolina não conforme (ver anexo I).

A figura 3.14 apresenta um esquema da estrutura utilizada neste trabalho para a classificação das gasolinas. Resumidamente, temos duas etapas distintas: (1) a calibração, onde uma quantidade razoável de amostras é utilizada para a concepção dos modelos, e (2) a validação externa, onde 50 novas amostras são utilizadas para verificar a performance destes modelos. Os quadros (cinzas) em destaque representam os principais métodos estatísticos multivariáveis utilizados no trabalho.

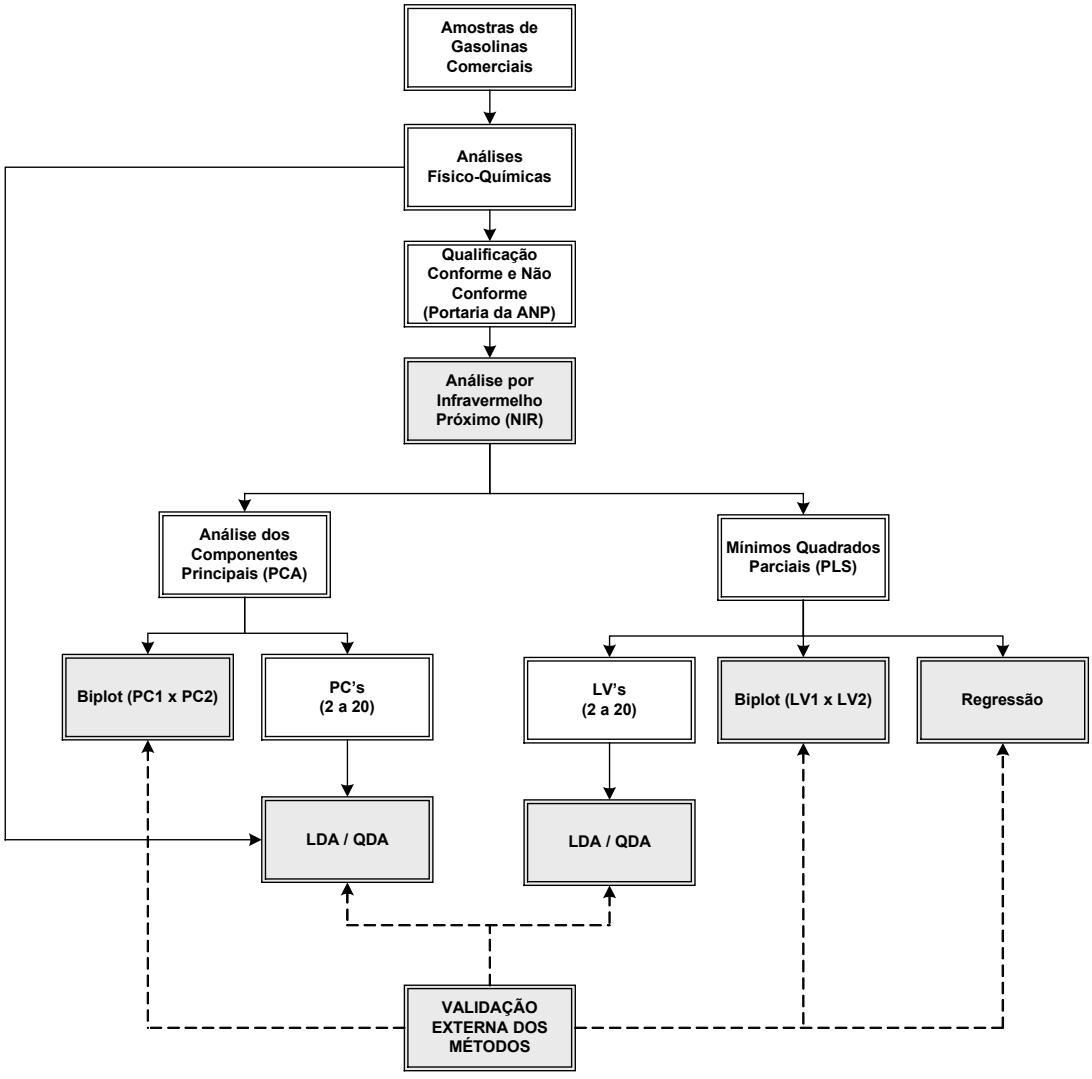

Figura 3.14 – Estrutura para classificação de gasolinas

Antes da aplicação dos classificadores QDA e LDA a matriz de dados dos resultados físico-químicos, os componentes principais e as variáveis latentes foram normalizados para o intervalo de –1 a 1, utilizando os valores máximos e mínimos de cada variável.

CAPÍTULO 4

4 RESULTADOS E DISCUSSÃO

Um dos objetivos deste trabalho é minimizar o número de análises físico-químicas a serem realizadas, substituindo estas análises por análises baseadas em NIR. Para tal, é interessante implementar uma metodologia baseada nesta análise que permita classificar as amostras em conformes e não conformes.

Durante a classificação das amostras podem ocorrer dois tipos de erro:

- erro do tipo 1: amostras conformes classificadas como não conformes;
- erro do tipo 2: amostras não conformes classificadas como conformes.

Nesta metodologia seria desejável que se evite ao máximo erros do tipo 2, ou seja, erros em que se considerem conformes amostras não conformes. Isto porque, a ocorrência de erros do tipo 1 levaria apenas à realização de análises físico-químicas inutilmente, enquanto que a ocorrência de erros do tipo 2 levaria à aceitação de amostras não conformes que não seriam submetidas a análises físico-químicas.

A classificação proposta pela ANP é baseada em limites máximos e mínimos para uma série de diferentes propriedades físico-químicas. Inicialmente, apresentamos a aplicação das técnicas de LDA e QDA a estas mesmas propriedades, com o intuito de ilustrar a aplicação destas técnicas.

A seguir são apresentadas três diferentes metodologias baseadas nas análises espectrais NIR das amostras:

- aplicação de técnicas de compressão de dados (PCA e PLS) aos espectros e aplicação de QDA e LDA aos dados comprimidos;
- aplicação de técnicas de regressão (PLS) para a predição da conformidade das amostras;

- utilização de ferramentas gráficas (*biplot*) para a classificação das amostras.

Finalmente, para verificar a validade destas técnicas, elas são aplicadas a um conjunto de dados de validação.

4.1 Métodos analíticos

Foram realizados ensaios analíticos em 284 amostras de gasolina comum coletadas aleatoriamente no período de fevereiro a maio de 2004, nos postos de vendas de combustíveis da grande São Paulo e região. Deste total, 95 amostras são provenientes da fiscalização realizada pela Agência Nacional do Petróleo e as 183 amostras restantes são oriundas do programa de monitoramento da qualidade realizado pelo IPT.

Após a execução dos ensaios analíticos e comparação dos resultados com as especificações da portaria da ANP, 192 (68%) amostras foram consideradas conformes, ou seja, atendiam as especificações estabelecidas pela portaria da ANP. As outras 92 (32%) amostras foram consideradas não conformes. Estes valores não servem como parâmetro para avaliação de adulterações dos postos de combustíveis da grande São Paulo, pois na composição das amostras está incluída uma grande quantidade de amostras de fiscalização, que normalmente é realizada quando se tem certeza de que há adulterações, ou por denúncias ou pela indicação dos ensaios do monitoramento da qualidade.

Pela análise dos resultados pode-se observar que as adulterações de combustíveis ocorrem principalmente pela adição de solventes e álcool, portanto os ensaios que melhor indicaram adulterações nestas amostras coletadas em ordem decrescente de importância são: marcador de solventes, octanagem, teor de álcool e destilação. A tabela 4.1 mostra numericamente esta análise.

Pode-se notar pela tabela que das 92 amostras não conformes, a grande maioria foi apontada pelo ensaio de marcador (90%), porém outros ensaios complementam esta avaliação. Das 27 amostras não conformes indicadas pelo ensaio de teor de álcool (NBR 13992), 8 (9%) delas não continham presença de marcador e o 1% restante foi indicado pelo ensaio de destilação (T10%).

É importante observar que apesar de não estar especificado na portaria da ANP o ensaio de RON (octanagem) indicou uma grande quantidade de amostras adulteradas (55%). Muitas vezes o índice antidetonante (IAD) está dentro da especificação, pois a combinação (média aritmética) entre um valor alto de MON e um valor baixo de RON dá esta condição.

Tabela 4.1 – Quantidade de amostras não conformes detectadas nos ensaios

Ensaio		Número de amostras consideradas não conformes
Marcador		78
Octanagem	MON	6
	RON	51
	IAD	23
Teor de álcool – analisador portátil		28
Teor de álcool – método proveta		27
Destilação	T10%	17
	T50%	1
	T90%	1
	Ponto final	5
Benzeno – analisador portátil (I.V.)		13

O ensaio de massa específica não foi relacionado na tabela porque não existe uma especificação adotada pela portaria, o resultado serve apenas como referência e deve ser anotado. Fazendo uma analogia entre valores de massa específica de

gasolinas adulteradas com gasolinas boas não foi possível fazer nenhuma classificação baseada apenas nesta propriedade.

Diante disto, os dados foram organizados em quatro matrizes. Cada matriz contem as análises de 284 amostras. A última coluna corresponde à classificação da amostra quanto a sua conformidade obtida pelos métodos analíticos, comparando sempre com os limites estabelecidos pelas portarias da ANP.

As outras colunas das matrizes são formadas com os resultados obtidos na destilação (T10%, T50%, T90% e PF), octanagem (MON, RON e IAD), teor de álcool (proveta e analisador portátil) e benzeno (10 resultados). A densidade é utilizada em dois casos.

As quatro matrizes se distinguem pela variedade de ensaios levados em conta para cada amostra de gasolina, incluindo-se, ou não, a densidade, e na maneira como a conformidade da amostra é definida a partir destes ensaios. As matrizes são definidas como segue:

- *Dadosfq* – A matriz contém os dados brutos da análise de cada amostra, definidos com base nos resultados físico-químicos. Na classificação da conformidade da amostra não é considerada a incerteza de medição dos dados físico-químicos, apenas o valor nominal da análise;
- *Dadosfqsd* – A matriz contém os dados brutos da análise de cada amostra, definidos com base nos resultados físico-químicos, sem a densidade. Não é considerada a incerteza de medição na classificação da conformidade da amostra;
- *Dadosfqmarc* – A matriz contém os dados brutos da análise de cada amostra, definidos com base nos resultados físico-químicos, incluindo a densidade. Foram considerados a incerteza de medição e o limite de detecção dos equipamentos, principalmente nos ensaios de marcador e teor de álcool, na classificação da conformidade da amostra, e não apenas o seu valor nominal;
- *Dadosfqmarcsd* – Esta matriz foi construída de maneira similar à anterior, no entanto a informação de densidade foi excluída dela.

4.1.1 Análises por NIR

Das 284 amostras de gasolina comum analisadas pelos ensaios físico-químicos, 266 delas foram analisadas por espectroscopia NIR, sendo que, 216 delas (anexo I) foram utilizadas para a construção do modelo e 50 amostras (anexo II) para a validação dos modelos. Das 216 amostras utilizadas 145 (67%) são conformes e 71 (33%) são não conformes. Como comentado no item 4.1 estes valores não servem como parâmetro para avaliação dos postos de combustíveis da grande São Paulo.

Foram realizadas duas leituras de espectro NIR para cada amostra ensaiada, sendo as suas médias calculadas e utilizadas no trabalho. Inicialmente o espectro foi obtido em toda a região do infravermelho próximo (4000 cm^{-1} a 14000 cm^{-1}), em seguida foi construído um gráfico das absorbâncias em função do número de onda (cm^{-1}), onde foi possível delimitar a região espectral que possivelmente representaria estas amostras de gasolina (figura 4.1). Eliminando as áreas correspondentes a picos de saturação, a região espectral ficou definida nos intervalos de 4471 cm^{-1} a 5740 cm^{-1} e 5921 cm^{-1} a 9000 cm^{-1}.

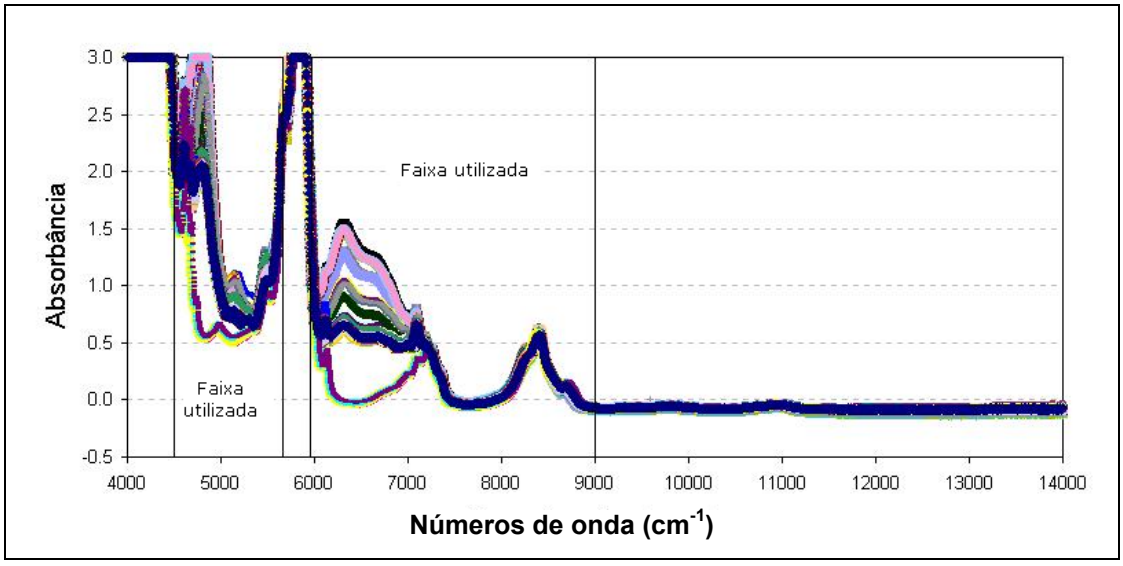

Figura 4.1 – Espectro total da análise NIR

A figura 4.2 apresenta os espectros das amostras de gasolinas após a definição das regiões acima. Inicialmente a quantidade de variáveis (comprimentos de ondas) contida no espectro era de aproximadamente 8000 e após a seleção da região este valor diminuiu para 2256.

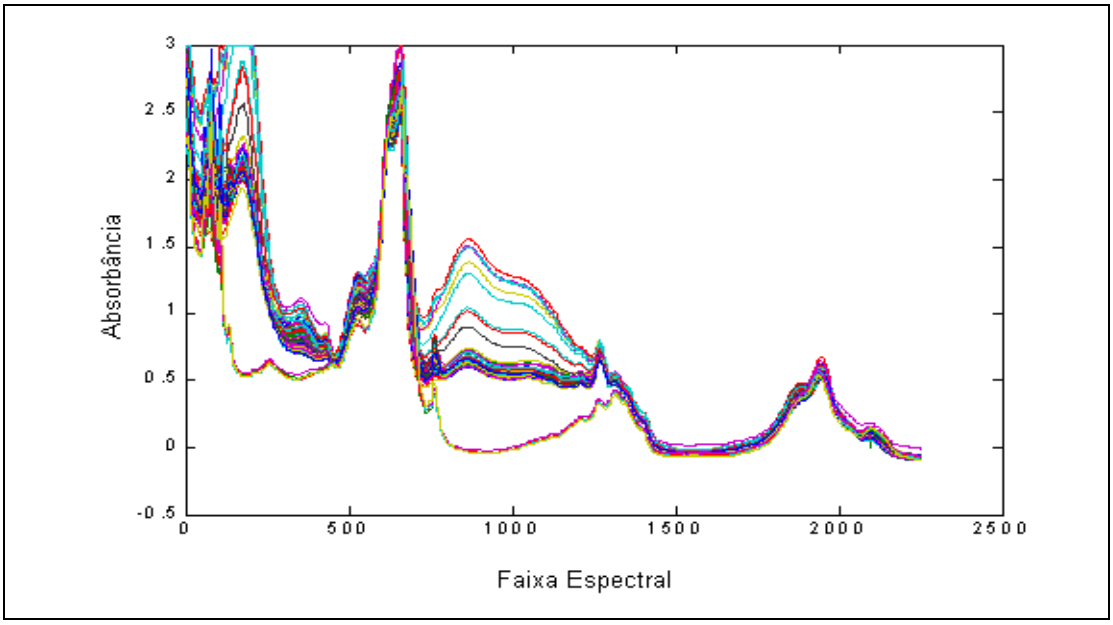

Figura 4.2 – Espectros NIR utilizados na PCA e PLS

4.2 Classificadores

Os algoritmos para os classificadores LDA e QDA foram implementados no programa Matlab (versão 6.5 – release 13) de acordo com a metodologia apresentada por Wu et al. (1996) descrita no capítulo 3. Inicialmente é feita a normalização dos dados físico-químicos utilizando os valores mínimo e máximo, tanto para o QDA quanto para o LDA. No anexo III são apresentados os arquivos fonte referente a estas rotinas.

4.2.1 LDA

Foi implementado o algoritmo para o cálculo dos *scores* de classificação da LDA. A primeira etapa foi classificar as gasolinas em duas classes: conforme e não

conforme e em seguida calcular a média de cada classe.

A segunda etapa consistiu em calcular as variâncias de cada classe, e depois combiná-las de forma a obter a variância combinada de acordo com equação (3.13) do capítulo 3.

Na etapa final foi obtido o *score* de classificação (equação 3.14), e finalmente foi investigado quais amostras não tiveram classificação correta.

4.2.1.1 LDA (Dados físico-químicos)

Aplicando a análise discriminante linear às quatro matrizes de dados físico-químicos foram obtidos os resultados apresentados na tabela 4.2:

O melhor resultado foi obtido utilizando os dados para os quais foram observados os limites de detecção para os equipamentos e verificada a incerteza de medição para se estabelecer a conformidade. Nestes dados estava incluída também a densidade entre os dados físico-químicos utilizados.

Todas as 9 amostras analisadas erroneamente a partir da matriz *Dadosfqmarc* apresentaram erro do tipo 2, que corresponde a 3,2% do total de amostras analisadas. As amostras classificadas erroneamente neste caso podem ser interpretadas da seguinte maneira:

- uma das amostras contem 28% ± 1% de álcool em sua composição, valor este que está muito próximo do limite estabelecido para o teor de álcool (25% ± 1%);
- quatro amostras ficaram no limite de classificação do ensaio de marcador;
- as demais amostras classificadas erroneamente também eram amostras não conformes devido ao ensaio de marcador (três) e teor de álcool (uma).

Fazendo o teste sem os resultados da massa específica, observa-se que o número de amostras erradas aumenta, tanto para os dados em que se considera a

incerteza das medidas na definição da conformidade quanto para os que não se considera. Isto indica que a densidade é um dado importante para os métodos de classificação puramente estatísticos baseados em LDA.

Tabela 4.2 – Resultados obtidos com a LDA nos dados físico-químicos

Matriz	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
Dadosfq	20	N.O.[a]	20 (7,0%)[b]
Dadosfqsd	22	N.O.	22 (7,8%)
Dadosfqmarc	09	N.O.	09 (3,2%)
Dadosfqmarcsd	15	N.O.	15 (5,3%)

[a] N.O. - Não Observado
[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

4.2.2 QDA

O algoritmo para a QDA também está apresentado no anexo III. O que diferencia os métodos QDA do LDA é a matriz de variância que é diferente para o grupo conforme e não conforme. No entanto como a coluna correspondente aos resultados do marcador é praticamente zero para os dados considerados conforme, a matriz de variância torna-se singular, o que torna o cálculo do classificador impossível. Para evitar este problema, a matriz de dados utilizada na análise QDA não leva em conta as informações referentes aos resultados do marcador.

4.2.2.1 QDA (Dados físico-químicos)

Os resultados da análise discriminante quadrática nas quatro matrizes de dados físico-químicos são apresentados na tabela 4.3. Diferentemente do obtido através da LDA, os resultados obtidos com a QDA, com e sem a coluna da massa

específica, foram praticamente os mesmos no primeiro caso e no segundo caso. Porém, no contexto geral a quantidade de amostras classificadas erroneamente foi maior no caso da QDA. Assim como na LDA a melhor matriz testada foi a *Dadosfqmarc*.

Do total de 284 amostras testadas, 7% (20 amostras) delas foram classificadas erroneamente. Sendo que, 1% teve erro do tipo 1 e 6% erro do tipo2.

Fazendo uma análise somente dos erros, das 20 amostras classificadas erroneamente 85% tiveram erro do tipo 2 e 15% tiveram erro do tipo 1. A principal razão para este pior desempenho foi o fato de não ter sido utilizada a informação referente às análises de marcador.

Tabela 4.3 – Resultados obtidos com a QDA nos dados físico-químicos

Matriz	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
Dadosfq	28 (9,9%)	04 (1,4%) [b]	24 (8,4%)
Dadosfqsd	29 (10,2%)	05 (1,8%)	24 (8,4%)
Dadosfqmarc	20 (7,0%)	03 (1,0%)	17 (6,0%)
Dadosfqmarcsd	21 (7,4%)	02 (0,7%)	19 (6,7%)

[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

4.2.3 Resumo LDA e QDA (Dados físico-químicos)

Na tabela 4.4 é apresentado um resumo dos resultados obtidos aplicando os classificadores LDA e QDA aos dados das análises físico-químicas. O melhor resultado de cada método é apresentado. Nota-se que a porcentagem de acertos é insatisfatória.

Tabela 4.4 – Porcentagem de acertos para os classificadores LDA e QDA

Método	Total de amostras	N° de acertos	Classificação % acertos
LDA	284	275	96,8
QDA		264	93,0

4.3 PCA-LDA e PCA-QDA

Como o objetivo é minimizar o número de análises físico-químicas a serem realizadas, substituindo estas análises por análises baseadas em NIR, é interessante que se evite ao máximo erros do tipo 2, ou seja em que se considerem conformes amostras não conformes. Nesta etapa, foram utilizadas 216 amostras de gasolina comum na análise por componentes principais. A PCA foi aplicada utilizando-se o Toolbox_PLS (*Eigenvector*) para Matlab. Foi construída uma matriz (dimensão de 216 x 2256) com todos os dados dos espectros da região selecionada anteriormente.

Os dados foram pré-tratados utilizando os recursos do Toolbox_PLS. Utilizou-se a centralização pela média para normalização dos dados (*Preprocess – mean*). Em seguida foram obtidos os componentes principais, cujos resultados estão apresentados na figura 4.3. Os dois primeiros componentes já explicam praticamente 95% da variância dos dados originais.

No procedimento, ao invés dos resultados das análises físico-químicas, são utilizados os *scores* (projeções) correspondentes a "n" primeiros componentes principais.

A figura 4.4 apresenta os resultados obtidos utilizando os componentes principais na análise discriminante linear (LDA). A melhor classificação utilizando a PCA-LDA foi obtida utilizando 19 e 20 componentes principais. Para estes PC's o número de erros foi de 11. Sendo que somente a amostra número 1 apresentou erro

do tipo 1, onde a amostra era conforme e foi classificada como não conforme. As demais amostras (tabela 4.5) apresentaram erro tipo 2, as amostras eram não conformes e foram classificadas como conformes. A porcentagem total de erros (tipo 1 + tipo 2) foi de 5,1%.

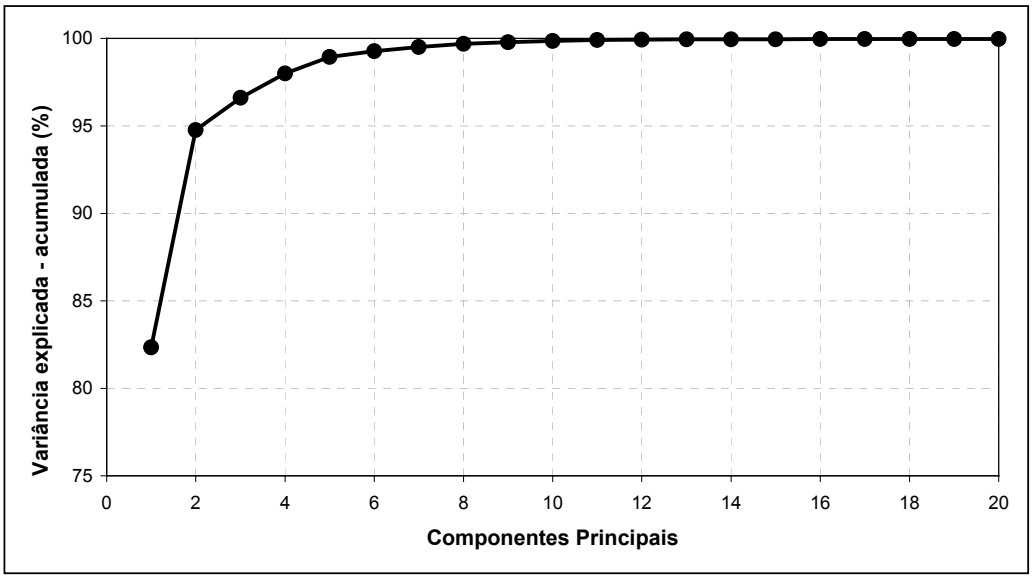

Figura 4.3 – Resultado da PCA (variância explicada pelo modelo PCA)

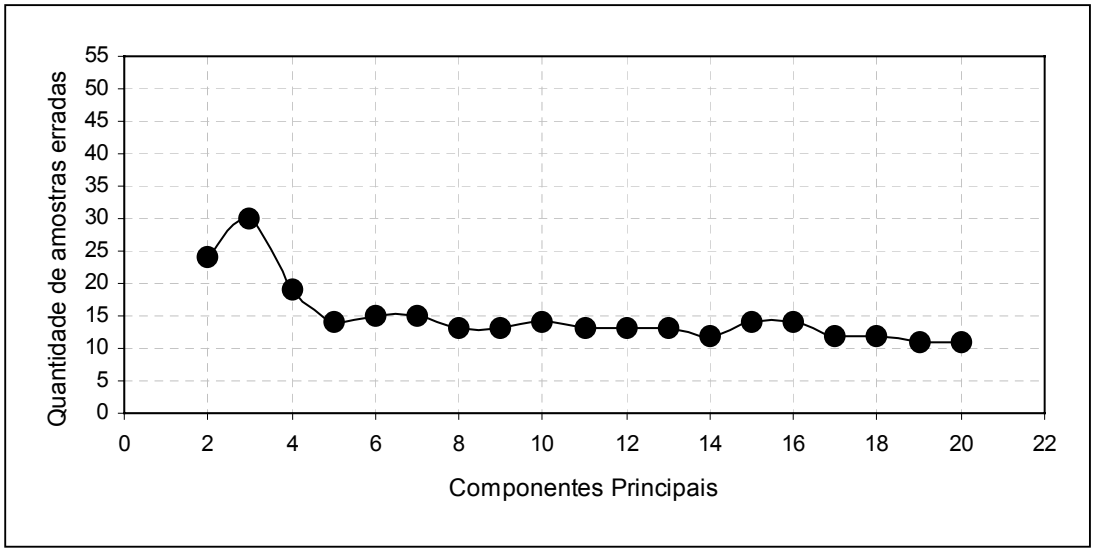

Figura 4.4 – Resultado utilizando PCA-LDA

Tabela 4.5 – Resultados da PCA-LDA

N° de componetes principais (PC) utilizados	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
19	1, 17, 48, 116, 121, 152, 170, 188, 194, 195, 209	01(0,5%)[b]	10 (4,6%)
20	1, 17, 48, 116, 121, 152, 170, 188, 194, 195, 209	01(0,5%)	10 (4,6%)

[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

Na figura 4.5 são apresentados os resultados obtidos utilizando os componentes principais na análise discriminante quadrática (QDA). O melhor resultado obtido foi utilizando 10 componentes principais, onde o número de amostras classificadas erroneamente foi de 7.

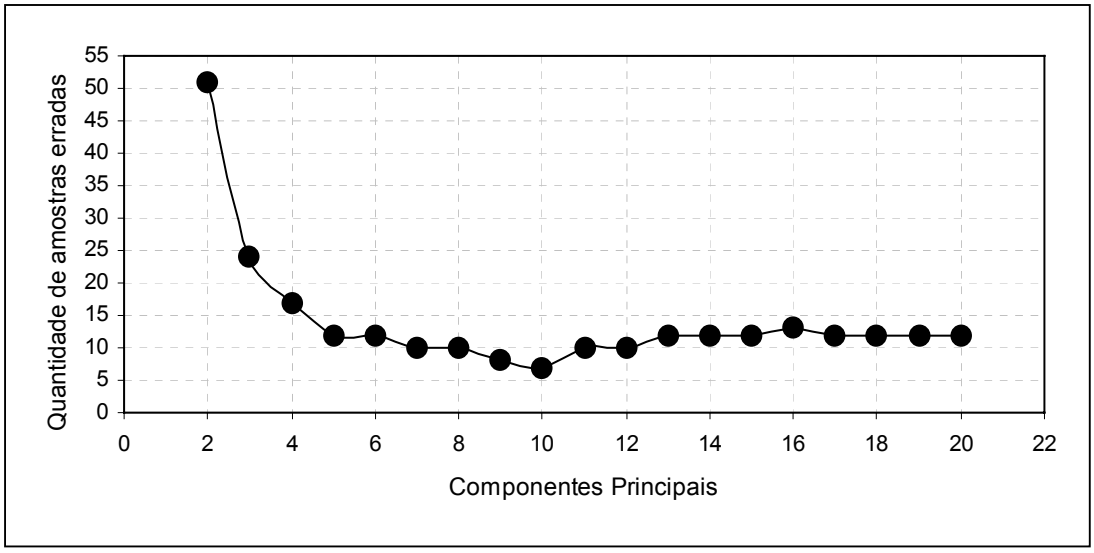

Figura 4.5 – Resultado utilizando PCA-QDA

As amostras classificadas erroneamente na análise PCA-QDA são apresentadas na tabela 4.6. Como aconteceu na análise LDA a amostra 1 teve erro do

tipo 1 e as demais tiveram erro do tipo 2. A porcentagem total de erros (tipo 1 + tipo 2) foi de 3,3%.

Tabela 4.6 – Resultados da PCA-QDA

N° de componetes principais (PC) utilizados	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
10	1, 17, 48, 116, 170, 195, 209	01(0,5%)[b]	06 (2,8%)

[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

A tabela 4.7 mostra a porcentagem de acertos dos classificadores LDA e QDA com melhor desempenho utilizando os componentes principais. A QDA teve um melhor desempenho que a LDA, utilizando somente 10 componentes principais (que explicam 99,85% da variância das variáveis originais) contra 19 componentes utilizados pela LDA. Porém o tipo de erro foi similar nos dois casos, onde na maioria dos casos as amostras eram não conformes e foram classificadas como conformes.

Ao se analisar os resultados físico-químicos das 6 amostras classificadas erroneamente (erro tipo 2) pela QDA (2,8% do total de amostras analisadas) percebe-se que as amostras 17, 116, 195 e 209 são não conformes devido à presença de marcador. Este erro pode ter ocorrido pela distribuição dos resultados do marcador, como foi visto no começo deste capítulo. O ensaio de marcador classificou 78 amostras como não conformes (com resultados variando de 20 ppb a 270 ppb) e 138 amostras conformes com valores, na maioria dos casos, com 0 ppb. Talvez, esta grande variação faça com que amostras com valores próximos ao limite de detecção do equipamento apresentem este erro na classificação.

A amostra 48 pode ser considerada suspeita, pois tem 27% de álcool e se considerarmos 1% de incerteza no resultado ela estaria dentro do limite da portaria da ANP que é de 25% ± 1%. Com relação às amostras 1 e 170, a primeira está dentro

das especificações da portaria em todos os ensaios e a segunda está fora pois foram detectados 147 ppb de marcador. Uma explicação para a classificação errônea da amostra 1, é que ela possui uma temperatura baixa do ponto T10% da destilação. Apesar de estar dentro das especificações da portaria da ANP, ela tem um valor menor do que 1 desvio padrão em relação à média do conjunto analítico.

Tabela 4.7 – Porcentagem de acertos para os classificadores LDA e QDA na PCA

Método	**Total de amostras**	**N° de acertos**	**Classificação % acertos**
LDA	216	205	94,9
QDA		209	96,8

4.4 PLS-LDA e PLS-QDA

Aqui também foram utilizadas as mesmas 216 amostras de gasolina comum na análise através dos mínimos quadrados parciais (PLS). O PLS foi aplicado utilizando-se o Toolbox_PLS do Matlab. A mesma matriz de dados espectrais (dimensão de 216 x 2256) da PCA foi utilizada para o PLS. A utilização do PLS justifica-se pois este método propõe uma decomposição da matriz de dados espectrais em que se leva em conta a informação sobre variáveis dependentes (Y).

Três blocos diferentes (valores de Y) foram utilizados no PLS: (1) Utilizou-se o vetor coluna das amostras conformes e não conformes (conformidade), (2) utilizou-se dos resultados do marcador e (3) utilizou-se da combinação dos dois. Estes blocos foram escolhidos pois representam a informação que se quer discriminar.

Os dados foram pré-tratados utilizando os recursos do Toolbox_PLS. Após testes, utilizou-se a centralização pela média para normalização dos dados (*Preprocess – mean*). No procedimento, ao invés dos resultados das análises físico-químicas, são utilizados os *scores* (projeções) correspondentes a “n” primeiras variáveis latentes.

A figura 4.6 apresenta a porcentagem de variância explicada pelas variáveis latentes (LV) obtidos utilizando como bloco de variáveis dependentes (Y) os dados da conformidade. Analogamente à PCA todas as variáveis latentes de 1 a 20 foram testadas tanto para a LDA, como para a QDA (figuras 4.7 e 4.8 respectivamente).

Os melhores resultados obtidos na LDA-conformidade foram obtidos utilizando 18, 19 e 20 variáveis latentes. O método classificou erroneamente uma amostra em cada conjunto de variáveis (figura 4.7), porém as amostras classificadas erroneamente foram diferentes. Todas tiveram erro do tipo 2, ou seja, as amostras eram não conformes e foram classificadas como conformes. Utilizando as primeiras 18 LV's, a amostra n° 48 teve sua classificação errada, e utilizando 19 e 20 LV's a amostra errada foi a n° 195 em ambos os casos.

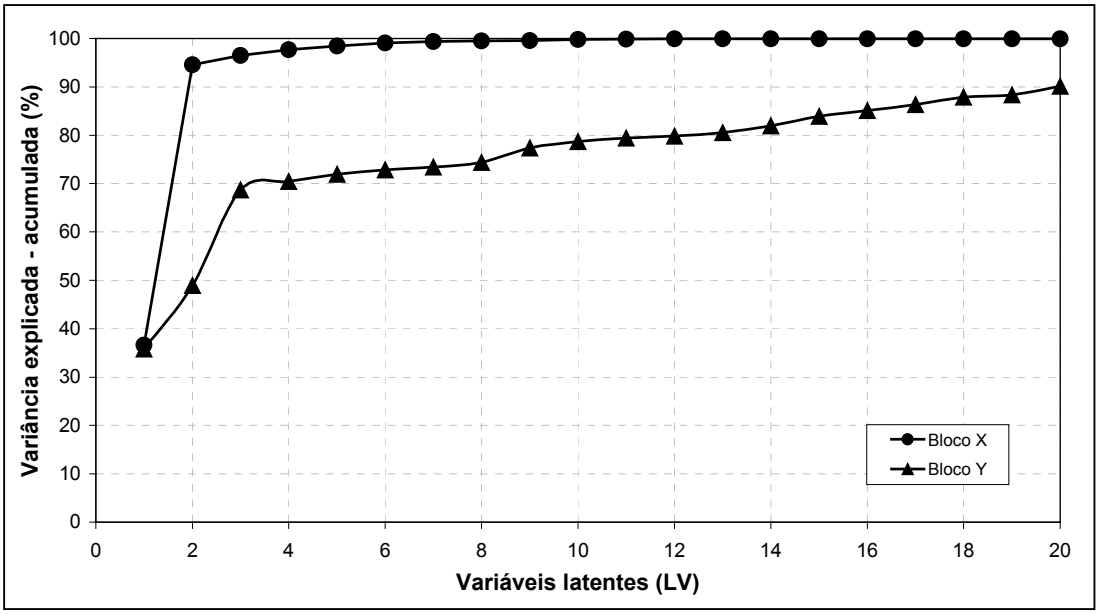

Figura 4.6 – Resultado da PLS (Bloco Y - conformidade)

Já a QDA teve uma melhor performance utilizando as 20 LV's. A quantidade de erros foi de 4, e as amostras erradas foram a 1, 17, 48 e a 195. A amostra 1 apresentou erro do tipo 1 e as demais tiveram erro do tipo 2.

Como as amostras classificadas erroneamente são as mesmas que em relação ao PCA, o mesmo comentário sobre a possível falha na classificação destas amostras pode ser aplicado.

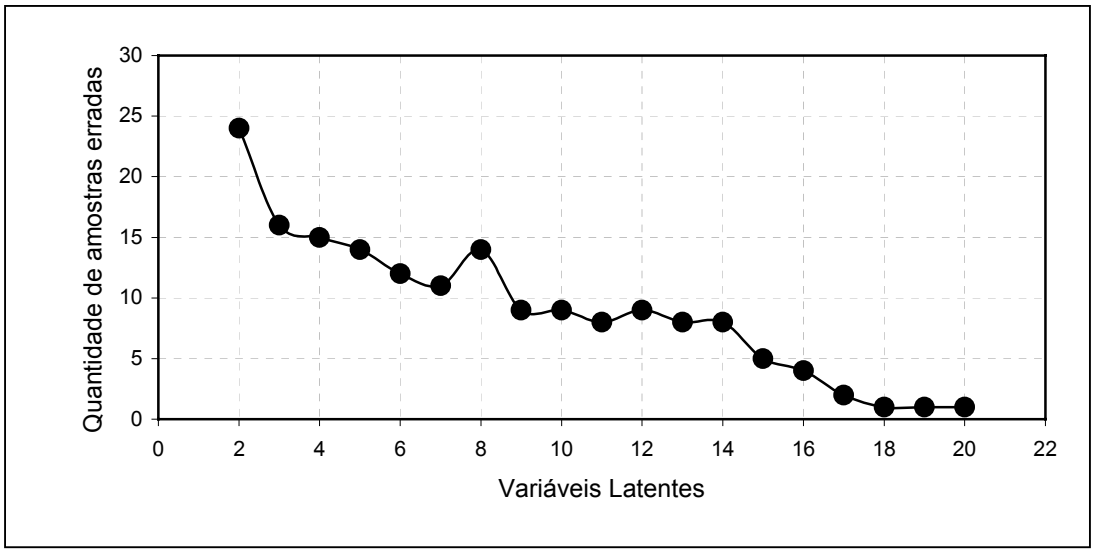

Figura 4.7 – Resultado utilizando LDA-conformidade

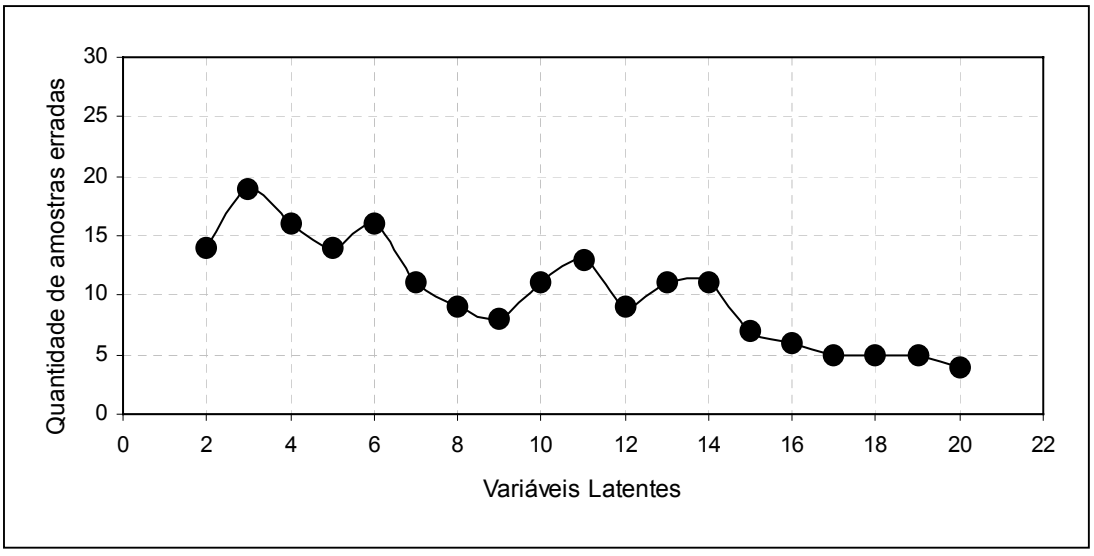

Figura 4.8 – Resultado utilizando QDA-conformidade

A porcentagem de variância explicada pelas variáveis latentes, obtidas utilizando o bloco de variáveis dependentes (Y) como sendo os dados do marcador

são apresentadas na figura 4.9. Aqui de 1 a 20 variáveis latentes foram testadas tanto para a LDA como para a QDA (figuras 4.10 e 4.11 respectivamente).

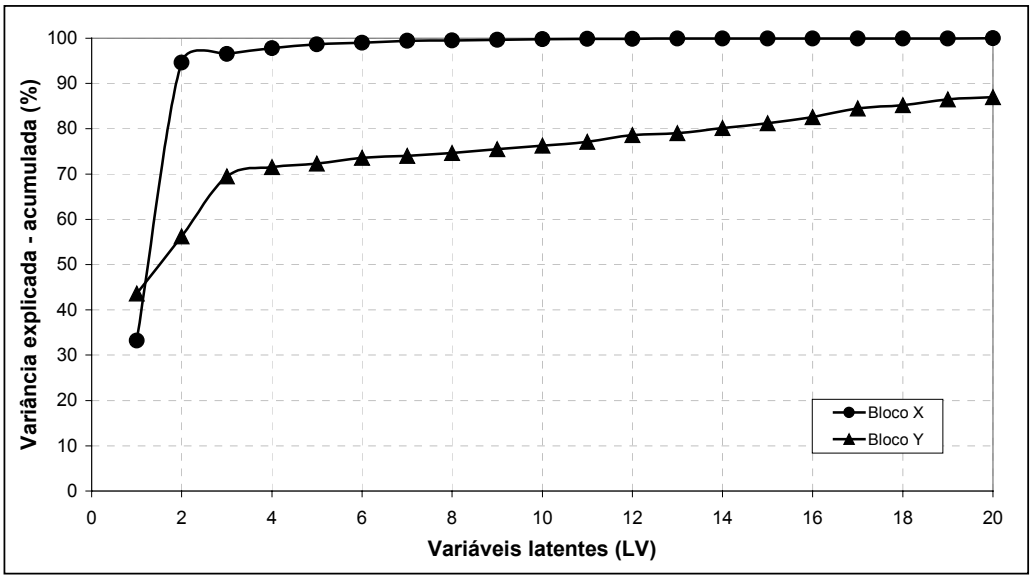

Figura 4.9 – Resultado da PLS (Bloco Y - marcador)

Tanto para a LDA (32) como para a QDA (35), a quantidade de amostras classificadas erroneamente foi muito grande.

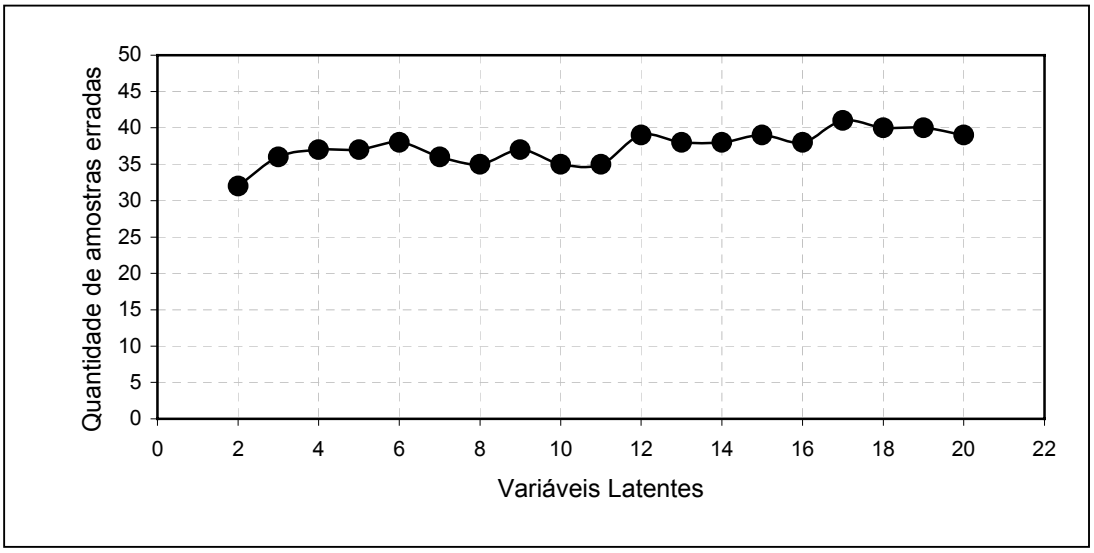

Figura 4.10 – Resultado utilizando LDA-marcador

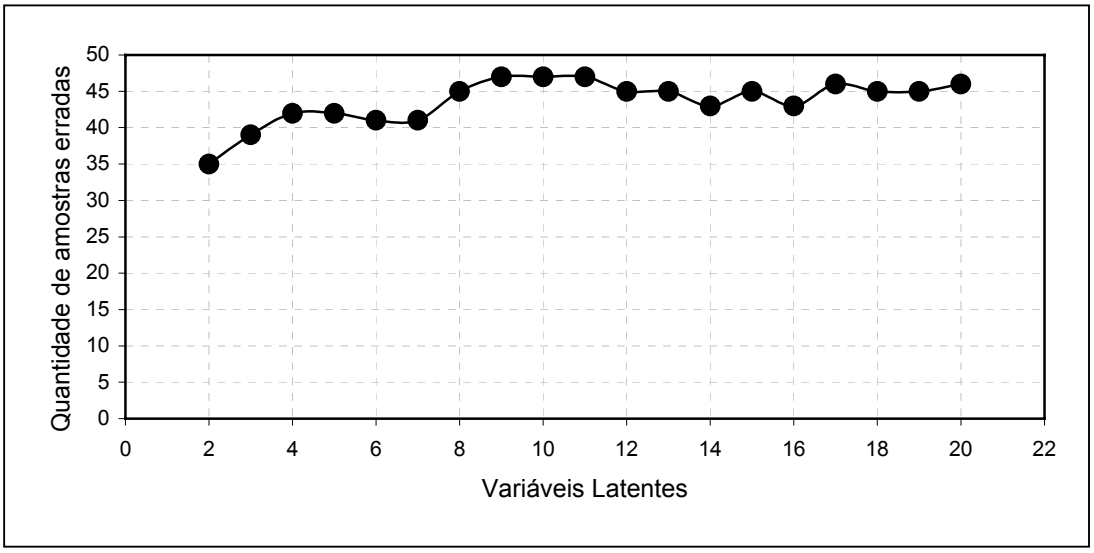

Figura 4.11 – Resultado utilizando QDA-marcador

A porcentagem de variância explicada pelas variáveis latentes, obtidas utilizando o bloco de dados de variáveis dependentes (Y) como sendo a combinação (conformidade + marcador) são apresentadas na figura 4.12. Aqui foram testadas de 1 a 20 variáveis latentes tanto para a LDA, como para a QDA (figuras 4.13 e 4.14 respectivamente).

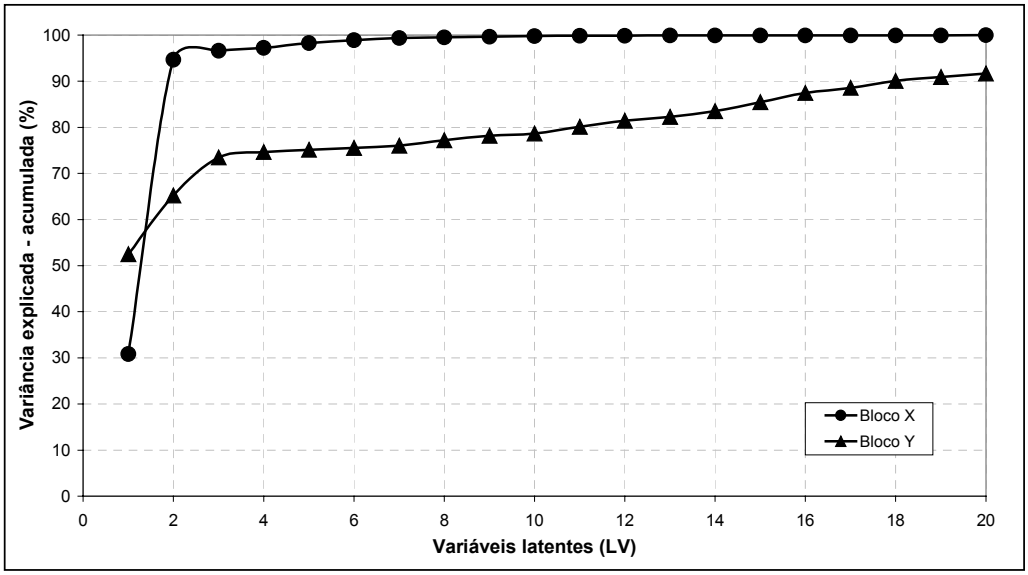

Figura 4.12 – Resultado da PLS (Bloco Y - combinação)

O resultado da utilização da combinação enquanto variáveis dependentes teve uma performance similar ao da utilização da conformidade unicamente. Aparentemente, a utilização do marcador teve uma influência negativa nos resultados, pois o número de amostras erradas na LDA-combinação foi de 3, utilizando 20 LV's (todas com erro tipo 2) e 5 amostras na QDA-combinação utilizando 17 e 20 LV's (uma com erro tipo 1 e 4 com erro do tipo 2).

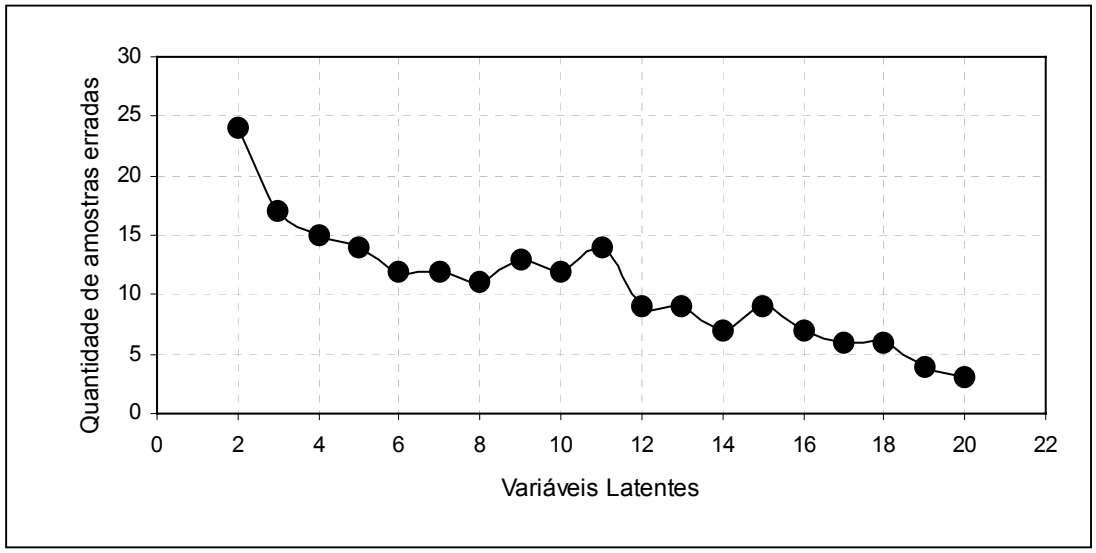

Figura 4.13 – Resultado utilizando LDA-combinação

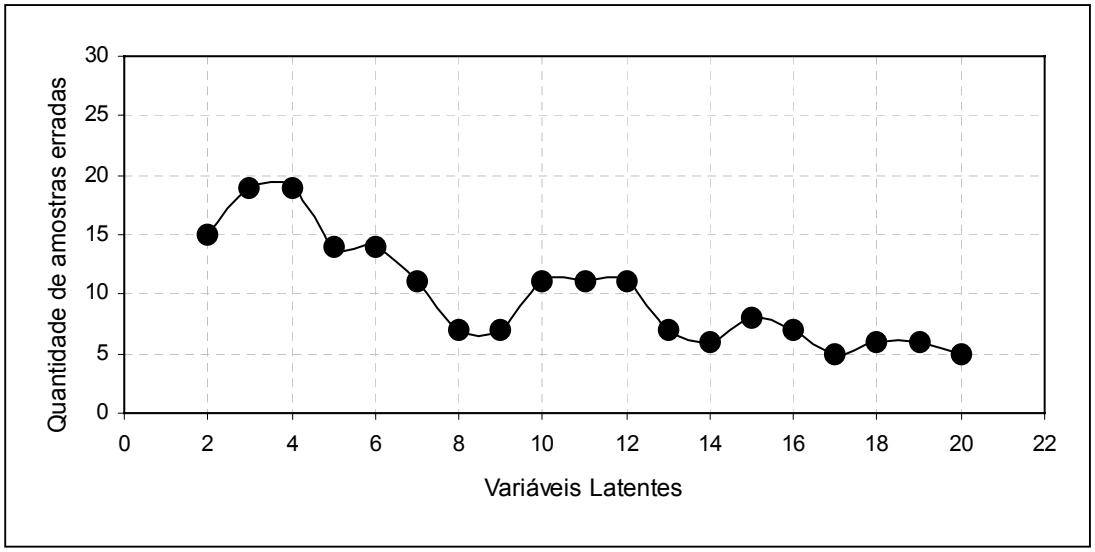

Figura 4.14 – Resultado utilizando QDA-combinação

Na tabela 4.8 é apresentado um resumo dos resultados obtidos com os classificadores LDA e QDA no PLS, utilizando os três blocos (valores de Y) definidos anteriormente. O melhor resultado de cada método é apresentado. Nota-se que a porcentagem de acertos para a LDA-conformidade é satisfatória. Em síntese, PLS-LDA dá os melhores resultados.

Tabela 4.8 – Porcentagem de acertos para os classificadores LDA e QDA no PLS

Método	**Total de amostras**	**N° de acertos**	**Classificação % acertos**
LDA-conformidade	216	215	99,5
LDA-marcador		184	85,2
LDA-combinação		213	98,6
QDA-conformidade		212	98,1
QDA-marcador		181	83,8
QDA-combinação		211	97,7

4.5 Regressão utilizando PLS

Uma outra técnica proposta para inferir a conformidade das amostras é através de um modelo para a predição de propriedades. Baseado nos resultados anteriores foi utilizado um modelo para a predição da conformidade. O modelo foi construído utilizando-se os mesmos 216 espectros de amostras de gasolinas (Bloco X) e os respectivos resultados de conformidade obtidos a partir das análises físico-químicas e de marcador (Bloco Y). O gráfico de paridade dos valores medidos em função dos valores preditos está apresentado na figura 4.15. Para as gasolinas não conformes (0) os valores mínimo e máximo foram: -0,2422 e 0,5173. Para as gasolinas conformes (1) os valores mínimo e máximo foram: 0,6432 e 1,2491. Pôde-se notar que as duas regiões ficaram bem definidas, e nesta fase (calibração) houve apenas duas amostras classificadas erroneamente, e ambas com erro do tipo1. A

amostra 1 (0,6432) já comentada nos modelos anteriores e a amostra 138 (0,6543) que pela análise físico-química está dentro das especificações e não tem nenhum desvio em relação ao conjunto.

Para dividir os grupos foi utilizado o método do valor de corte (*cutoff-value*; Sharma, 1996). Este método consiste em dividir o espaço discriminante em duas regiões e foi calculado como segue:

$$Valor\ de\ corte = \frac{n_1\overline{X}_1 + n_0\overline{X}_0}{n_1 + n_0} = \frac{145 \times 0{,}9675 + 71 \times 0{,}0664}{145 + 71} = 0{,}67$$

onde: $n_0 = 71$ (n° de amostras não conformes) e $n_1 = 145$ (n° de amostras conformes)

$\overline{X}_1 = 0{,}9675$ (média das amostras conformes)

$\overline{X}_0 = 0{,}0664$ (média das amostras não conformes)

Portanto a faixa para a classificação de amostras de gasolinas ficou definida assim: amostras não conformes < 0,67 < amostras conformes.

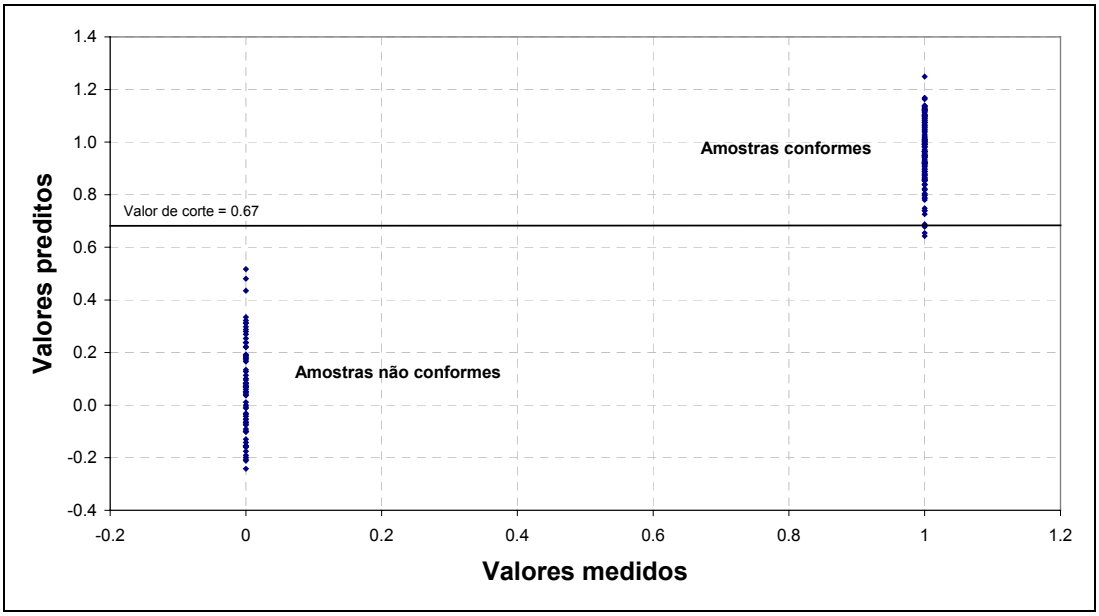

Figura 4.15 – Gráfico de paridade (PLS – conformidade)

• amostras testadas (abaixo do valor de corte são não conformes e acima são conformes.)

4.6 Análise dos gráficos (*biplot*) dos componentes principais e variáveis latentes

Outra maneira muito utilizada por pesquisadores para a classificação das amostras e identificação de *outliers*, é plotar o gráfico dos primeiros componentes principais (no caso da PCA) ou das primeiras variáveis latentes (no caso do PLS). Como estes *scores*, na maioria dos casos, explicam aproximadamente 90% da variância total das variáveis originais, é possível uma boa visualização dos pontos pertencentes a diferentes regiões.

Aqui também foram utilizadas as mesmas 216 amostras de gasolina comum utilizadas anteriormente. A mesma matriz de dados espectrais (dimensão de 216 x 2256) foi utilizada para o PCA e para o PLS. O PLS foi aplicado utilizando-se o marcador, a conformidade e a combinação de ambas enquanto variáveis dependentes (Y).

Em todos os casos estudados a seguir não foi feito um estudo a priori para eliminar possíveis *outliers*, pois além do conjunto de amostras ter um tamanho considerável, poucas amostras tiveram valores discrepantes das demais. Nos casos em que houve um distanciamento maior em relação a uma concentração maior de amostras, é porque existem grandes diferenças de composição da gasolina. Por exemplo, teor alto de álcool e adição de solventes adulterantes.

4.6.1 Componentes principais

A partir do gráfico do PC1 (componente principal 1) em função do PC2 é possível classificar o espaço em 7 regiões (figura 4.16), como descrito a seguir:

- região 1: contem amostras de gasolina do tipo A, ou seja, amostras sem adição de álcool etílico anidro. Nota-se que a região ficou bem definida e separada dos demais grupos;
- região 2: também ficou bem definida, as amostras pertencentes a este grupo contêm altos teores de álcool, acima dos estabelecidos pela portaria da ANP;

- região 3: contem amostras com dois tipos de adulterações, adição de álcool e solvente (marcador);
- regiões 4 e 5: formada por amostras conformes, porém a região 4 possui uma população maior do que a 5. Elas diferem entre si provavelmente pela diferença de massa específica, na região 5 prevalecem amostras com massa específica maior;
- região 6: composta por amostras não conformes, devido principalmente à presença de marcador e índice antidetonante alto;
- região 7: composta de amostras conformes e não conformes, ela foi considerada como região suspeita, ou seja, por segurança as amostras deste conjunto são não conformes e deverão passar pelos ensaios físico-químicos para avaliação final.

Mesmo com esta divisão podemos verificar que houve algumas amostras mal classificadas. Dentro da região 4 aparecem sete pontos vermelhos que são amostras não conformes (17, 154, 156, 170, 181, 183 e 209) todas pela presença de marcador. Excluindo a amostra 170 (147 ppb) as demais continham marcador na faixa de 20 ppb a 80 ppb. Algumas destas amostras já foram comentadas nos resultados da PCA-LDA e PCA-QDA. A amostra 48 considerada suspeita (item 4.3) ficou exatamente entre os grupos 4 e 5.

Outra classificação errada ocorreu na região 6, onde a amostra 1 (conforme) aparece isolada entre as amostras não conformes. Detalhes sobre esta amostra foram discutidos anteriormente no item 4.3.

Na maioria dos métodos foi possível observar uma falha de classificação quando as amostras continham apenas adulteração por adição de marcador, ou seja, presença de marcador na gasolina na faixa entre 20 e 80 ppb.

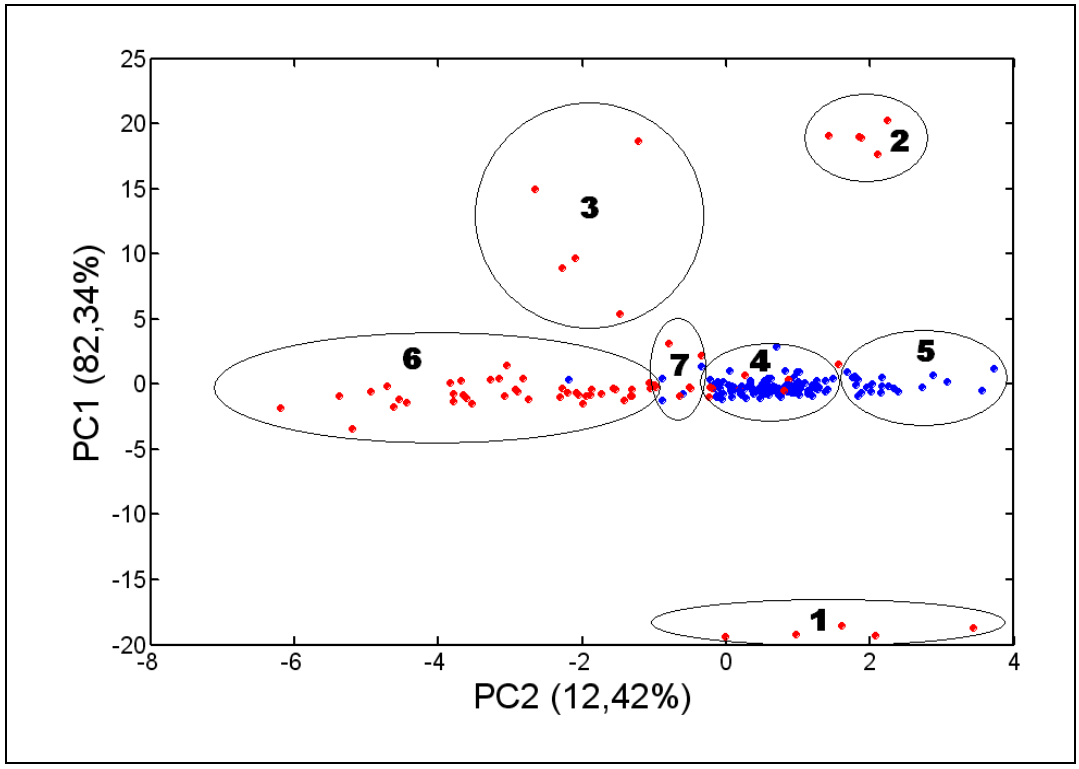

Figura 4.16 – Gráfico dos primeiros dois componentes

● Amostras conformes, ● amostras não conformes

4.6.2 Variáveis Latentes

Neste caso foram testadas as combinações de *biplot* entre as variáveis latentes para os modelos da conformidade, marcador e combinação, conforme citado anteriormente. A seguir são apresentados os melhores resultados de cada modelo.

A figura 4.17 apresenta o gráfico das duas primeiras variáveis latentes (LV1xLV2), onde a LV1 explica 36,6% da variância dos dados originais e a LV2 explica 57,96% para o modelo da conformidade. As duas juntas explicam 94,56% da variância dos dados originais. Observando esta figura pode-se notar que foi possível definir praticamente as mesmas regiões que com o PCA, com uma única alteração, as regiões de amostras conformes anteriormente separadas (4 e 5) agora estão fundidas na região 4. A região das amostras não conformes é a região 5, e a região 6 é

composta por amostras suspeitas. As regiões 1, 2 e 3 seguem a mesma definição explicada quando da análise do PCA.

Dentro da região 4 (amostras conformes) aparecem cinco pontos vermelhos que são amostras não conformes (17, 48, 156, 181, 183), ocorrendo aqui erro do tipo 2. Todas estas amostras já foram comentadas anteriormente na PCA, assim como a amostra número 1 que teve erro do tipo 1 e que aparece na região 5.

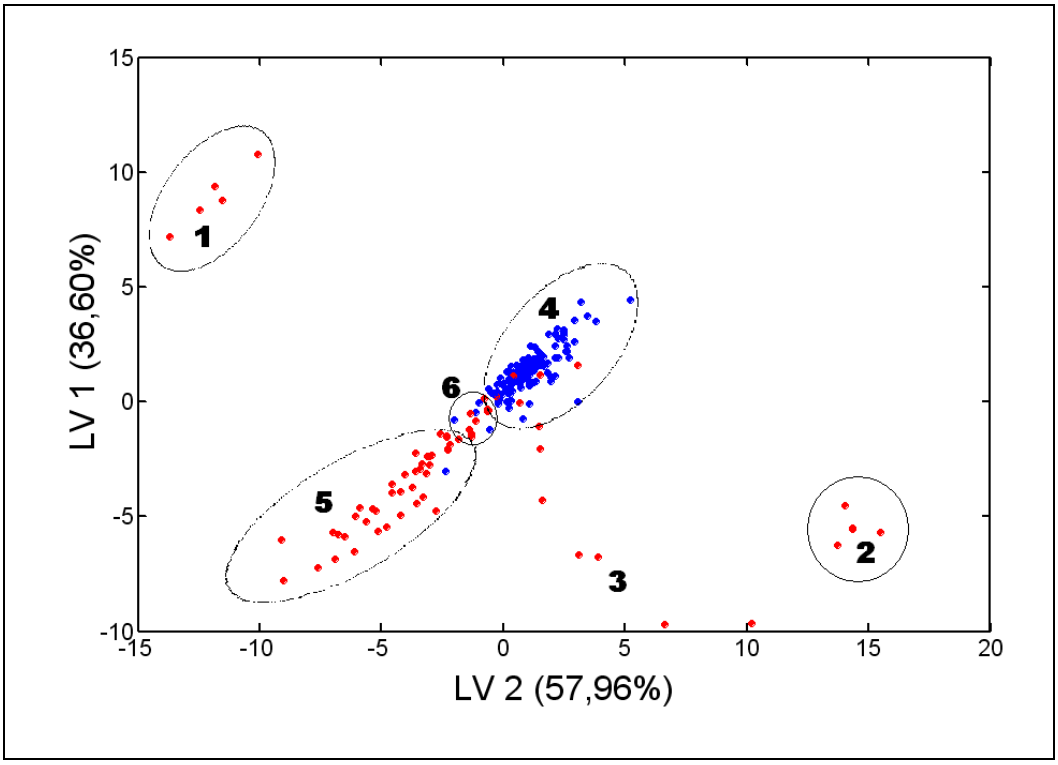

Figura 4.17 – Gráfico das duas primeiras variáveis da Conformidade

● Amostras conformes, ● amostras não conformes

Uma outra boa combinação utilizando o modelo da conformidade foi a combinação das variáveis latentes LV1 (36,60%) em função da LV3 (1,89%). Ambas explicando 38,49% da variância dos dados originais. A figura 4.18 apresenta esta combinação e as regiões encontradas.

Apesar de não estarem tão claras como anteriormente, foi possível definir 4 regiões distintas. As regiões 1 e 2 foram definidas como anteriormente na PCA, a

primeira com amostras de gasolina comum do tipo A e a segunda com amostras adulteradas com teor de álcool elevado. A região 3 (retângulo cinza) é composta por amostras conformes e a região 4 pelas demais amostras não conformes. A região 3 tem uma característica interessante, sua definição é bem clara as amostras conformes estão dentro de um retângulo com o limite superior de valor 5 e limite inferior de valor 0.

Assim como na análise anterior (LV1xLV2), aparecem cinco amostras não conformes (17, 48, 181, 183 e 209) na região das conformes (erro do tipo 2), porém a amostra que aparecia anteriormente (156) agora está fora da região das conformes, mas próxima do limite inferior (linha do zero). A amostra 209 foi classificada como não conforme, pois foi detectada presença de marcador de solventes. Com a definição do retângulo para as amostras conformes, ocorreram mais erros do tipo 1 na região 4 em comparação com o modelo anterior (ver pontos azuis na região 4).

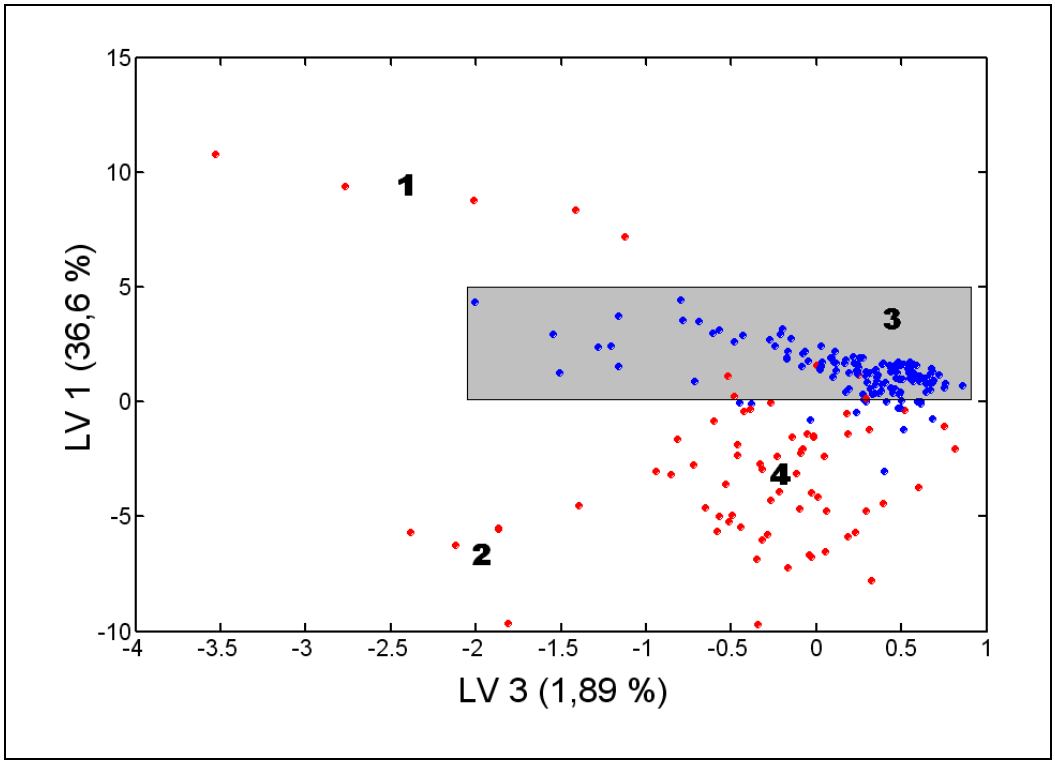

Figura 4.18 – Gráfico das variáveis latentes LV1xLV3 da Conformidade

● Amostras conformes, ● amostras não conformes

O resultado do modelo marcador é apresentado pela figura 4.19, pode-se observar que ele se aproxima do modelo conformidade, porém existe uma inversão do conjunto de amostras conformes com não conformes. No caso do modelo marcador as duas primeiras variáveis latentes (LV1 - 30,83% e LV2 - 63,86%) explicam juntas 94,69% da variância das variáveis originais. A linha próxima do zero pode ser utilizada simplesmente para dividir o espaço em duas regiões, uma conforme e outra não conforme. A região acima da linha formada pelas regiões 2, 4 e 5 são de amostras não conformes e a região abaixo da linha formada pelas regiões 1 e 3 são de amostras conformes. No contexto do estudo a região 1, formada por amostras de gasolina comum do tipo A, é considerada não conforme, pois ela é muito diferente de uma amostra de gasolina comum tipo C de boa qualidade. Como a gasolina A é de boa procedência, se adicionarmos 25% de álcool etílico anidro à sua composição, ela provavelmente iria para a região 3.

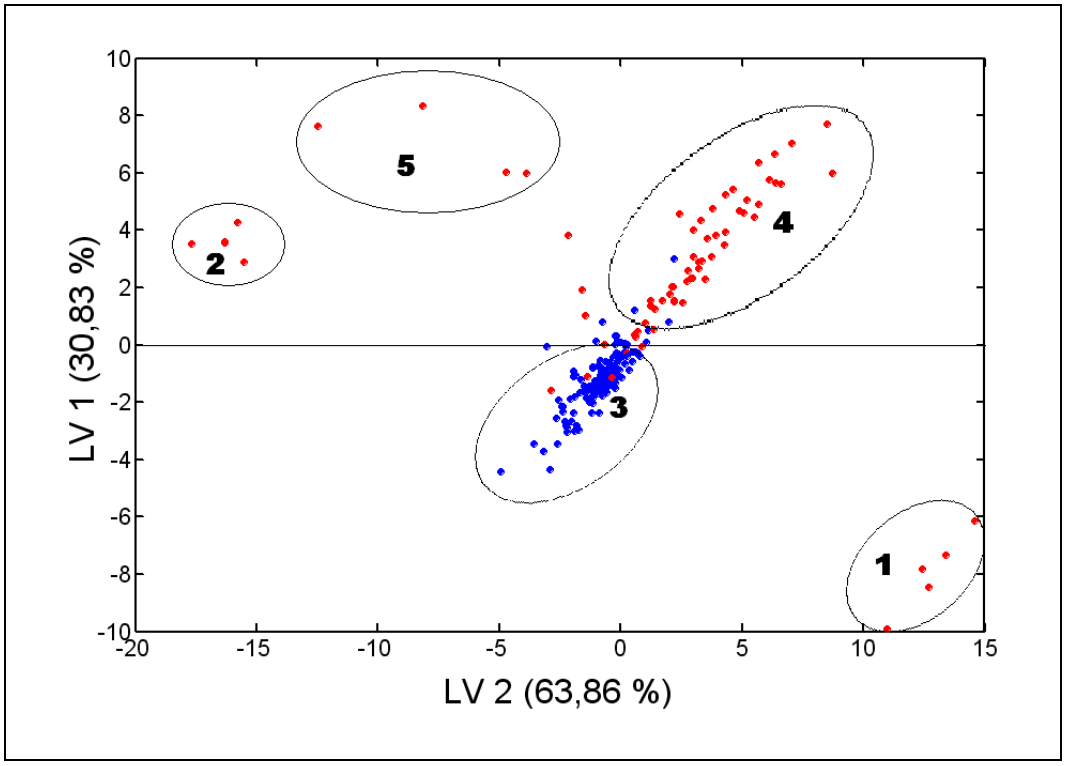

Figura 4.19 – Gráfico das variáveis latentes LV1xLV3 do Marcador

● Amostras conformes, ● amostras não conformes

No modelo marcador o erro do tipo 2 ocorreu em 6 casos (17, 48, 156, 181, 183 e 209) já comentados em outros modelos. Quanto ao erro do tipo 1, ele ocorreu em grande quantidade se for assumido a linha do zero como valor de corte, por outro lado se a região 4 for utilizada somente 3 amostras estariam erradas. De qualquer forma existe uma região (suspeita) entre as regiões 3 e 4 que em um procedimento de decisão deveria ser assumida como de amostras não conformes e confirmadas posteriormente através dos ensaios físico-químicos.

O último modelo estudado foi a combinação dos dois primeiros. Como o bloco Y possuía tanto dados da conformidade e do marcador, com escalas diferentes, foi aplicado o pré-tratamento de escalonamento nestes dados. Já, aos dados espectrais foi aplicada a centralização pela média.

O resultado obtido é apresentado na figura 4.20, onde as duas primeiras variáveis latentes (LV1 - 33,23% e LV2 - 61,42%) explicam 94,65% da variância dos dados originais, praticamente o mesmo resultado obtido no modelo conformidade (LV1xLV2) e marcador. A figura também se assemelha muito às anteriores, porém há uma inversão vertical quando comparada à figura 4.17 (conformidade) e inversão horizontal quando comparada à figura 4.19 (marcador). Os mesmos 6 casos ocorridos no modelo marcador, erros e comentários podem ser aplicados ao modelo de combinação.

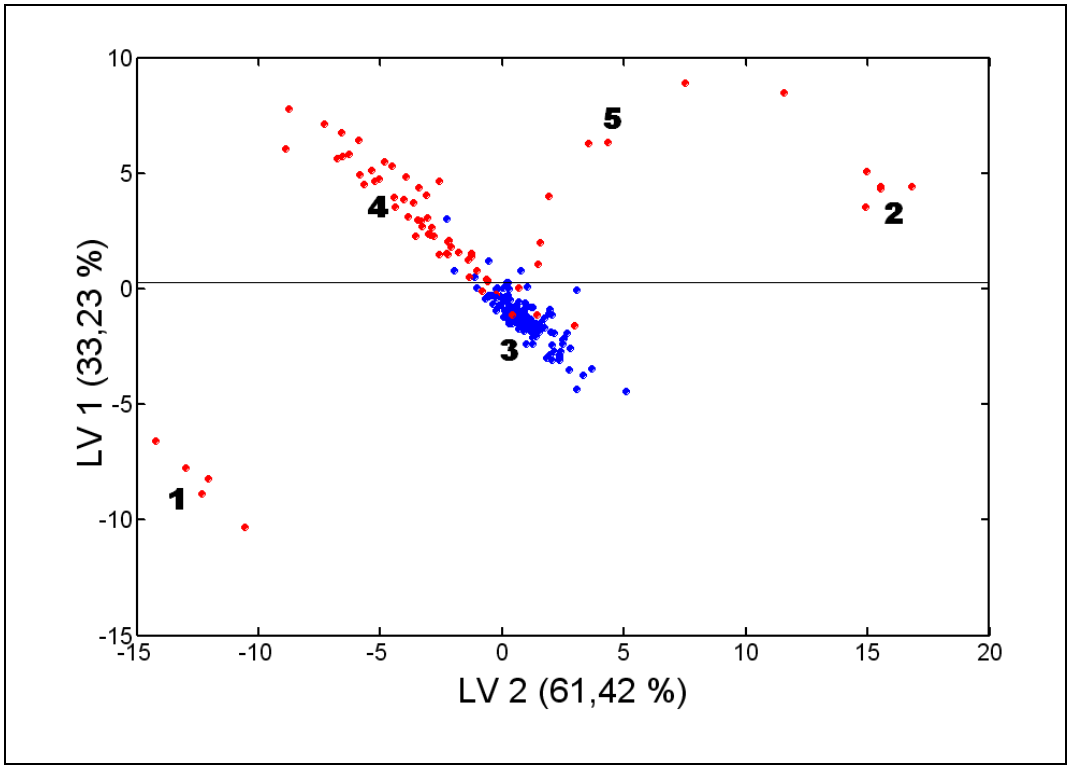

Figura 4.20 – Gráfico das variáveis latentes LV1xLV3 da Combinação

● Amostras conformes, ● amostras não conformes

4.7 Validação dos métodos

Para validação dos métodos foi utilizada a validação externa, composta por um conjunto suplementar de 50 amostras de gasolinas comerciais. Como no primeiro conjunto de calibração, estas amostras foram coletadas aleatoriamente nos postos revendedores de combustíveis da grande São Paulo. Os espectros destas amostras foram coletados de acordo com os itens 3.16 e 4.1.1.

Os resultados e os tipos de não conformidade apontadas nestas amostras são apresentados no anexo II. Das 50 amostras coletadas 32 (64%) eram conformes e 18 (36%) não conformes, como já comentado anteriormente estes valores não servem como parâmetro para avaliação dos combustíveis em São Paulo.

Nesta etapa foram testados os métodos com melhor desempenho, ou seja, o PCA-QDA, o PLS-LDA, o PLS-QDA, os primeiros componentes principais (PC1xPC2), o modelo de conformidade (LV1xLV2 e LV1xLV3), o modelo marcador (LV1xLV2), o modelo combinação (LV1xLV2) e o PLS regressão (conformidade).

Para cada método citado acima, foi introduzido o novo conjunto de dados NIR (50 amostras) e os novos valores obtidos foram comparados com os obtidos durante a calibração dos respectivos modelos.

4.7.1 PCA-QDA

Na tabela 4.9 são apresentadas as 7 amostras classificadas erroneamente utilizando 10 componentes principais. Na amostra 36 ocorreu erro do tipo 1 e nas demais (6 amostras) ocorreu erro do tipo 2. As amostras 16, 19 e 35 eram não conformes pela presença de marcador, as amostras 11 e 48 também eram não conformes pela presença de marcador, porém próximo ao limite de detecção do equipamento. A amostra 5 estava fora dos limites estabelecidos pela Portaria devido à octanagem (MON e RON) e destilação (T10%). Durante a etapa de calibração o método apresentou apenas 3,3 %.

Tabela 4.9 – Resultados da validação PCA-QDA

N° de componentes principais (PC) utilizados	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
10	5, 11, 16, 19, 35, 36, 48	01(2,0%)[b]	06 (12,0%)

[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

4.7.2 PLS-LDA

Confirmando o bom desempenho da calibração utilizando os dados de conformidade no algoritmo PLS-LDA, a validação teve uma melhor performance comparado aos demais métodos. Das 50 amostras utilizadas 5 amostras foram classificadas erroneamente, conforme tabela 4.10. Todas elas apresentaram erro do tipo 2, perfazendo um total de 10% de erros de predição em relação ao conjunto de validação. No entanto, na etapa de calibração o método apresentou apenas 0,5% de erro de classificação.

Tabela 4.10 – Resultados da validação PLS-LDA

N° de variáveis latentes (LV) utilizadas	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
19	11, 16, 19, 35, 48	N.O.[a]	05 (10%)[b]

[a] Não Observado
[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

4.7.3 PLS-QDA

Praticamente os mesmos erros ocorreram na validação da PLS-QDA (tabela 4.11). O total foi de 8 amostras classificadas erroneamente, e somente a amostra 8 que não pertencia ao grupo anterior (PCA-QDA). A amostra 8 também apresentou erro do tipo 2. A porcentagem total de amostras classificadas erroneamente (tipo 1 + tipo 2) foi de 16,0%, sendo que na etapa de calibração foi de 1,9 %.

Tabela 4.11 – Resultados da validação PLS-QDA

N° de variáveis latentes (LV) utilizadas	Amostras classificadas erroneamente	Erros do Tipo 1	Erros do Tipo 2
20	5, 8, 11, 16, 19, 35, 36, 48	01 (2,0%)[b]	07 (14,0%)

[b] Os valores entre parênteses correspondem a porcentagem de erros em relação ao total de amostras analisadas

4.7.4 Componentes Principais (PC1xPC2)

A figura 4.21 apresenta o resultado obtido na validação da análise dos dois primeiros componentes principais. Das 50 amostras testadas apenas 4 (8,0%) amostras não tiveram sua classificação correta. As amostras 11 e 48 (não conformes) foram projetadas na região 4 das amostras conformes, ocorrendo erro do tipo 2. Analisando os resultados físico-químicos destas duas amostras foi verificado que ambas foram consideradas não conformes devido a presença de marcador, porém a quantidade detectada estava próxima do limite de detecção do equipamento. Esse tipo de erro poderia ocorrer, uma vez que o modelo construído também apresentou este desvio.

As amostras 02 e 36 não se enquadraram em nenhuma região pré-estabelecida e foram consideradas como não conformes. Já a amostra 36 apesar de ser conforme, ela é diferente das outras 49 amostras do grupo validação, pois possui uma octanagem (MON, RON e IAD) mais elevada que as demais, e comparando com grupo das amostras de calibração está fora do desvio padrão das três propriedades de octanagem. Portanto, pode-se até dizer que no total foram 3 (6,0%) amostras classificadas erroneamente, duas com erro do tipo 2 e uma com erro do tipo 1. Já as amostras 04, 06 e 07 estão alocadas corretamente em suas regiões.

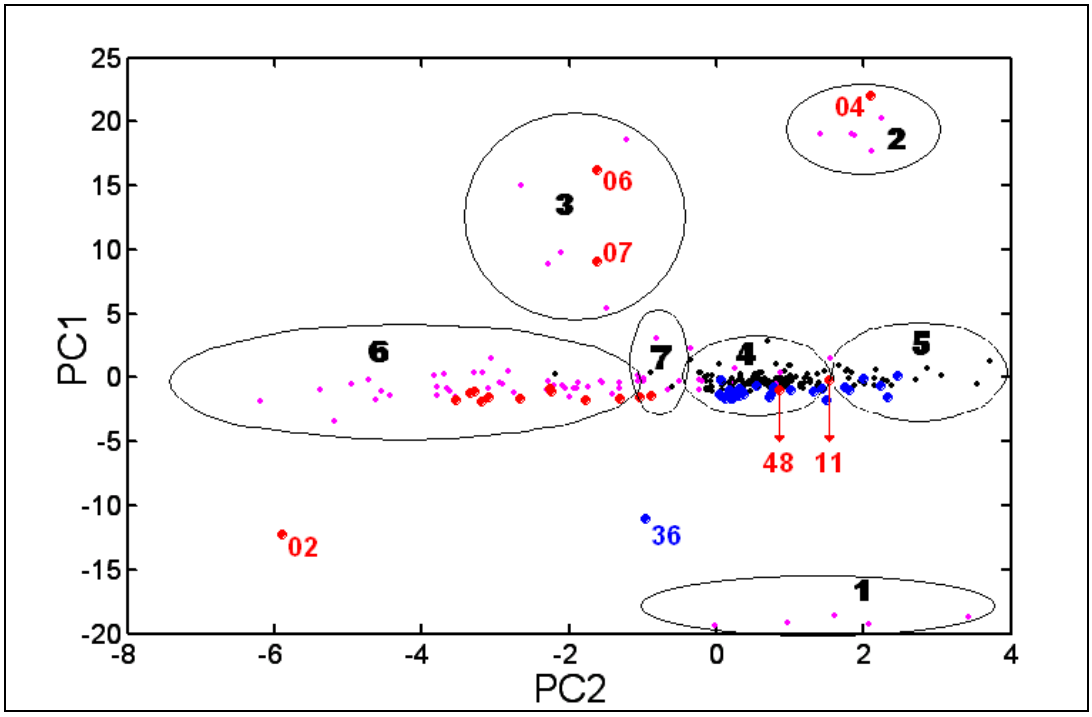

Figura 4.21 – Resultado da validação do modelo dos componentes principais

● Amostras conformes (validação), ● amostras não conformes (validação)
● Amostras conformes (calibração), ● Amostras não conformes (calibração)

4.7.5 Variáveis Latentes

Pode-se observar pela figura 4.22 que a validação do modelo conformidade (LV1xLV2) se comportou exatamente como o modelo PCA. As amostras 04, 06 e 07 se encontram dentro de suas regiões pré-estabelecidas na calibração. As amostras 02 e 36 não estão plotadas em nenhuma região definida, e as amostras 11 e 48 (não conformes) estão dentro da região 4 das conformes. Os mesmos comentários efetuados anteriormente se aplicam aqui, com exceção da amostra 36 que se encontra acima da linha retilínea do zero, esta linha indica que se uma amostra cair acima de zero é conforme e abaixo é não conforme. Portanto, a porcentagem de erro quando da utilização deste modelo foi de 4,0%.

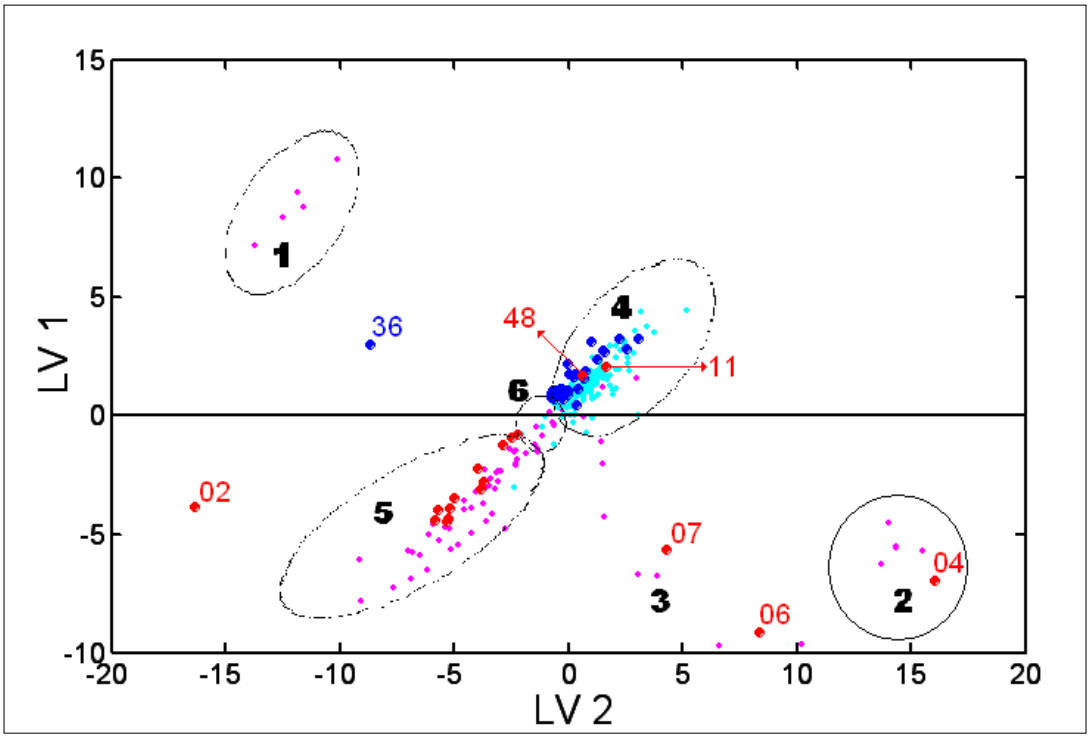

Figura 4.22 – Resultado da Validação do modelo PLS-Conformidade

● Amostras conformes (validação), ● amostras não conformes (validação)
● Amostras conformes (calibração), ● Amostras não conformes (calibração)

A validação do modelo conformidade (LV1xLV3) é apresentada na figura – 4.23. Pode-se observar que somente as amostras 11 e 48 foram classificadas erradamente. Como dito anteriormente, estas são amostras não conformes e estão dentro da região 3 (retângulo cinza) das amostras conformes. Considerando toda a região fora do retângulo como não conforme, as demais amostras tiveram uma boa classificação. Comparando com os outros modelos não é possível definir regiões por tipo de adulteração, por exemplo, a amostra 04 pertence à região 2 (teor de álcool elevado), mas está afastada. Já a amostra 06 pertence à região das amostras adulteradas com teor de álcool elevado e marcador de solventes. Portanto, a porcentagem de erro quando na utilização deste modelo foi de 4,0%.

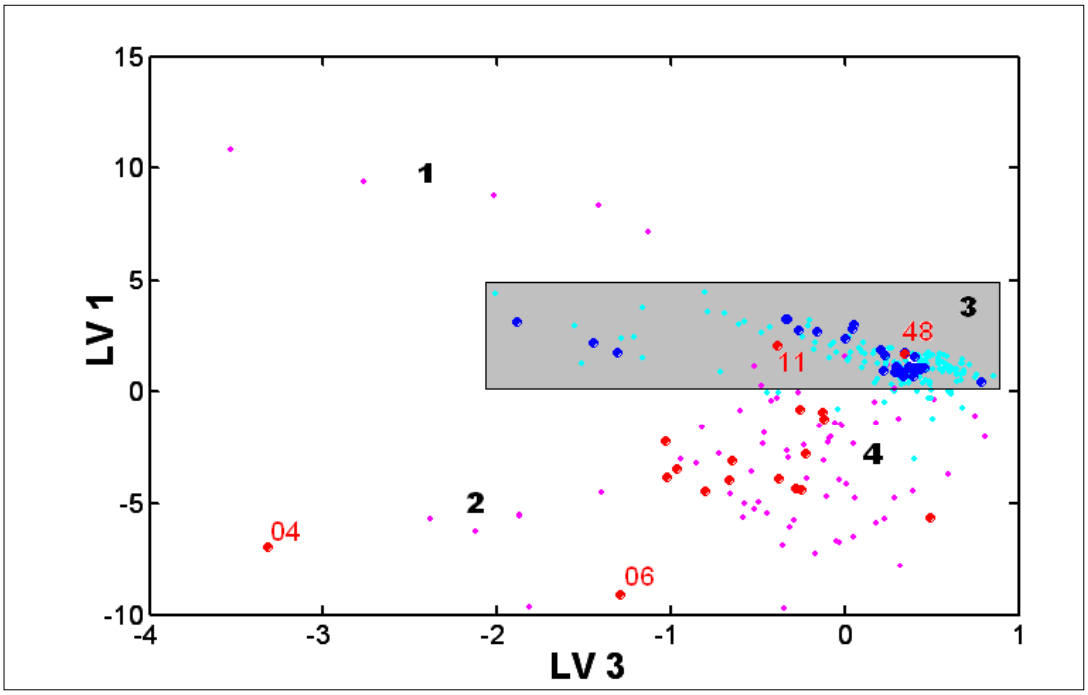

Figura 4.23 – Resultados da validação do modelo PLS-Conformidade

● Amostras conformes (validação), ● amostras não conformes (validação)
● Amostras conformes (calibração), ● Amostras não conformes (calibração)

A figura 4.24 apresenta o gráfico (LV1xLV2) do modelo marcador. Nota-se que o resultado é idêntico ao mostrado pelo gráfico do modelo conformidade (LV1xLV2). Portanto, a porcentagem de erro quando da utilização deste modelo foi de 6,0%. Os mesmos comentários efetuados anteriormente se aplicam aqui, a amostra que cair acima de zero é conforme e abaixo é não conforme.

Como o modelo conformidade e marcador deram o mesmo resultado, era de se esperar que a combinação repetisse este fato. Isto realmente aconteceu, pode-se observar pela figura 4.25 que as amostras em destaques estão alocadas nas mesmas regiões dos modelos anteriores. Por isso os mesmos comentários anteriores também podem ser aplicados aqui.

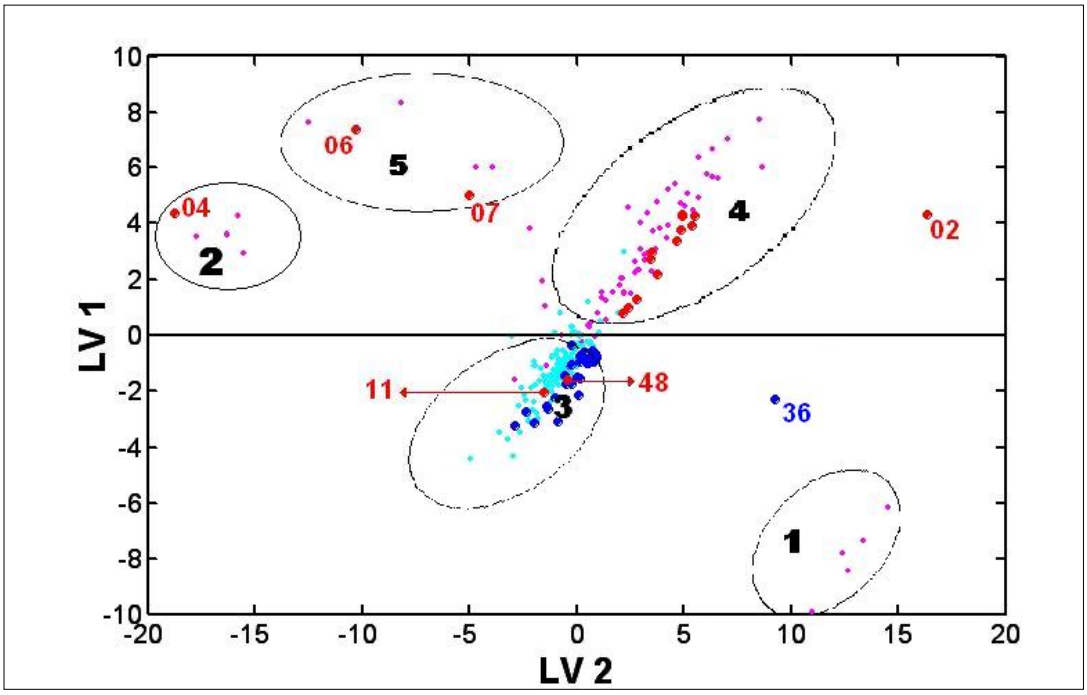

Figura 4.24 – Resultados da validação do modelo PLS-Marcador

● Amostras conformes (validação), ● amostras não conformes (validação)
● Amostras conformes (calibração), ● Amostras não conformes (calibração)

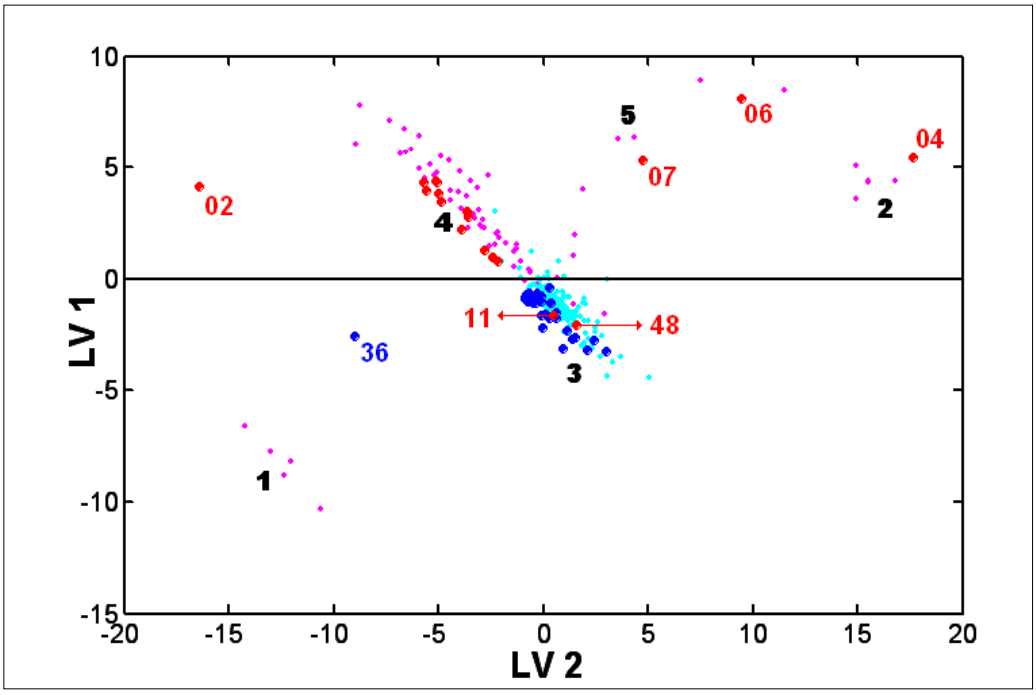

Figura 4.25 – Resultados da validação do modelo PLS-Combinação

● Amostras conformes (validação), ● amostras não conformes (validação)
● Amostras conformes (calibração), ● Amostras não conformes (calibração)

4.7.6 Validação PLS (Regressão)

Das 50 amostras utilizadas para validação do modelo da regressão PLS, 4 delas não tiveram boa predição e ficaram fora da linha de corte, 2 apresentando erro do tipo 1 e 2 erro do tipo 2, conforme mostra a figura 4.26. Portanto, a porcentagem de erro quando da utilização deste modelo foi de 8,0%.

As amostras com erro do tipo 1 foram as de número 30 (0,55) e 38 (0,54). Avaliando os resultados de suas propriedades físico-químicas constata-se que ambas estão dentro das especificações da ANP, porém a amostra 30 possui uma pequena quantidade de marcador e quatro ensaios que estão dentro dos limites da portaria mas fora do desvio padrão da população. Já as amostras 11 (0,74) e 48 (1,14) foram consideradas conformes provavelmente porque possuem marcador na faixa próxima ao limite de detecção do equipamento.

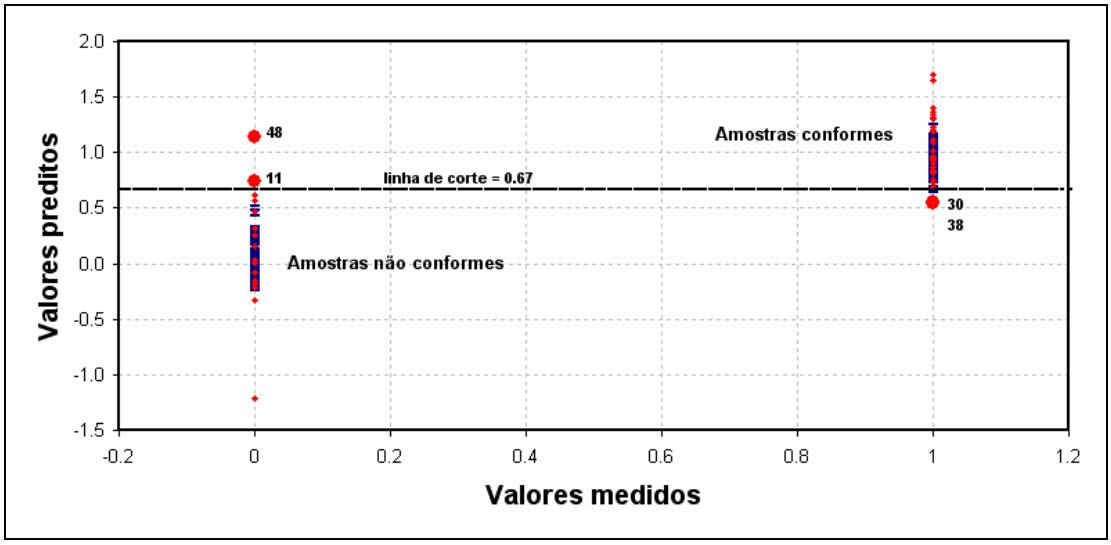

Figura 4.26 – Resultados da validação da regressão (PLS)

• Amostras da calibração, • amostras da validação

4.8 Conclusões

Neste capítulo foram implementadas algumas metodologias estatísticas multivariáveis para minimizar o número de análises físico-químicas, substituindo por análises baseadas em espectroscopia NIR.

Inicialmente, foi testada a aplicação das técnicas LDA e QDA nos resultados físico-químicos. O melhor resultado obtido foi utilizando a matriz com os dados brutos da análise de cada amostra, incluindo a massa específica, a incerteza de medição e o limite de detecção dos equipamentos. A LDA apresentou 3,2% de erro (todos do tipo 2), tendo portanto um melhor desempenho que a QDA que errou 7,0%, sendo 1,0% de erro tipo 1 e 6,0% de erro tipo 2. A ANP baseia a classificação sobre estes mesmos dados, utilizando faixas para definir os limites de conformidade. Isto ilustra que existem características muito particulares nos métodos LDA e QDA que fazem com que os mesmos não possam reproduzir exatamente a maneira como as especificações da ANP são expressas.

Em seguida, foi realizada a calibração para concepção dos modelos de classificação e os melhores resultados são apresentados à seguir:

(1) A PCA foi aplicada aos dados NIR para extração de informação e redução dos dados. Os componentes principais obtidos foram testados na LDA e QDA. A PCA combinada com a QDA apresentou o melhor resultado utilizando 10 componentes principais, que explicam 99,85% das variâncias dos dados originais. A QDA apresentou 3,3% de erro, sendo 0,5% de erro tipo 1 e 6,0% de erro tipo 2. Já a LDA apresentou 5,1% de erro.

(2) O PLS foi aplicado aos dados NIR (X) e às variáveis dependentes (Y) como sendo conformidade, marcador e combinação. As variáveis latentes obtidas foram testadas na LDA e QDA. Nos três casos acima a LDA teve uma melhor performance que a QDA. No melhor caso, enquanto a LDA-conformidade

apresentou apenas 0,5% de erro (ocorreu somente erro tipo 2) a QDA-conformidade apresentou 1,9% de erro.

(3) O resultado da regressão apresentou 0,9% de erro de predição, porém erros do tipo 1.

(4) Na análise dos gráficos (*Biplot*) foi possível além de classificar as amostras em conforme e não conforme, dividir o espaço amostral em regiões que permitem qualificar as amostras em outras classes. Esta idéia será melhor discutida no capítulo 5. Em relação ao desempenho dos modelos utilizados para classificar as amostras com exceção do gráfico dos PCs que teve um índice maior de erros do tipo 2 (3,7%), os gráficos das LVs obtiveram resultados muito próximos entre si, em torno de 2,3% de erro tipo 2. Só foi analisado aqui o erro do tipo 2, pois dependendo da forma de análise este número pode variar. Por exemplo, no gráfico dos PCs se passarmos uma linha vertical no valor zero de PC2, teremos definido duas regiões (lado esquerdo não conforme e lado direito conforme) que diminuirá a quantidade de erro tipo 2, porém aumentará drasticamente a quantidade de erro tipo 1.

Na etapa de calibração o melhor resultado foi apresentado pelo modelo PLS-LDA-conformidade com 0,5% de erro, seguido pelo modelo de regressão PLS-conformidade com 0,9% de erro e pelo modelo PLS-QDA conformidade com 1,9% de erro.

Na etapa de validação dos modelos com melhor desempenho através da validação externa, todos os modelos testados tiveram desempenho inferior à obtida durante a etapa de calibração.

Na validação a análise dos gráficos (*biplot*), seguida da regressão apresentaram melhor performance do que os modelos combinados de compressão de dados e análise discriminante.

Os resultados apresentados pela análise dos gráficos dos *scores* ficaram muito próximos. Todos os gráficos das variáveis latentes apresentaram 4% de erro, e o gráfico dos componentes principais apresentou 6% de erro.

A regressão PLS-conformidade apresentou 8% de erro, seguida pelas combinações PLS-QDA (10%), PCA-QDA (14%) e PLS-QDA (16,0%).

CAPÍTULO 5

5 CONCLUSÕES E PERSPECTIVAS

O objetivo deste trabalho foi aplicar ferramentas estatísticas multivariáveis à classificação de gasolinas comerciais com o intuito de gerar uma metodologia para fazer uma triagem das amostras coletadas durante o Programa de Monitoramento da Qualidade dos Combustíveis (PMQC). Esta triagem evitaria a necessidade de ensaiar amostras, que já se sabe que são de boa qualidade.

Os estudos objetivaram avaliar a possibilidade de fazer a classificação baseada em análises espectroscópicas em NIR. Foram avaliadas as técnicas de PCA-LDA, PCA-QDA, PLS-LDA, PLS-QDA, Regressão (PLS) para a predição da conformidade e de gráficos de *scores* (Biplot) utilizando PCA e PLS.

Fazendo o teste sem os resultados da massa específica, observa-se que o número de amostras erradas aumenta, tanto para os dados físico-químicos como para os dados em que se considera a incerteza da medida. Isto indica que a densidade é um dado importante para os métodos de classificação puramente estatísticos baseados em LDA.

Na etapa de calibração o melhor resultado foi apresentado pelo modelo PLS-LDA-conformidade com 0,5% de erro, seguido pelo modelo de regressão PLS-conformidade com 0,9% de erro e pelo modelo PLS-QDA conformidade com 1,9% de erro. Na etapa de validação dos modelos com melhor desempenho através da validação externa, todos os modelos testados tiveram desempenho inferior à obtida durante a etapa de calibração.

Globalmente, na validação a porcentagem de classificações corretas foi de 84 % a 96 %. Em muitas publicações este desempenho é considerado adequado, no entanto como neste caso trata-se de monitoramento para suporte à fiscalização, o desempenho deverá

ser melhorado para a sua implementação prática. Um aspecto importante é que em todos os modelos de classificação prevaleceram os erros do tipo 2 tanto na etapa de calibração quanto de validação.

Na maioria dos métodos foi possível observar uma falha de classificação quando as amostras contêm uma adulteração revelada pela presença de marcador na gasolina na faixa entre 20 e 80 ppb. Seria interessante fazer um estudo para verificar a sensibilidade dos espectros NIR em relação à presença de marcador na gasolina, podendo-se para tal preparar amostras padronizadas com solvente marcado, nas mais variadas concentrações.

A classificação através dos gráficos de *scores* sugeriu a presença de classes bastante diferenciadas mesmo entre combustíveis de uma mesma conformidade. Isto sugere a possibilidade de utilização de uma metodologia multi-classes, que certamente deve melhorar o desempenho da classificação , conforme sugerido por Tominaga (1999).

Outra questão que deve ser melhor estudada é com relação à quantidade e qualidade das amostras. Amostras com grandes desvios em relação ao conjunto utilizado durante a calibração podem influenciar de forma negativa o modelo de classificação, gerando um modelo deficiente. O número de dados utilizado também deve ser estudado, já que não é difícil aumentar o número de amostras analisadas, pois este cresce constantemente.

Devem ser realizados estudos de sensibilidade visando otimizar alguns parâmetros dos classificadores (em particular π_k) de forma a se atingir o objetivo de minimizar os erros tipo 2, e visando chegar a uma porcentagem de acertos previamente estabelecida.

Como visto anteriormente, existem ainda algumas etapas a serem alcançadas para tornar prática a utilização da classificação de gasolinas comerciais através de NIR, no entanto, a contribuição deste estudo é importante dentro deste objetivo.

REFERÊNCIAS BIBLIOGRÁFICAS

ASTM D2699 – Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel, 1999.

ASTM D2700 – Standard Test Method for Motor Octane Number of Spark-Ignition Engine Fuel, 1999.

ASTM D4052 - Standard Test Method for Density and Relative Density of Liquids by Digital Density Meter, 2002.

ASTM D6277 Standard Test Method for Determination of Benzene in Spark-Ignition Engine Fuels Using Mid Infrared Spectroscopy, 1999.

ASTM D86 Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure, 2001$^{\epsilon 1}$.

Campos, A. C., Leontsinis, E., *Petróleo e Derivados*, JR Editora Técnica Ltda., 1990.
Carrilo Le Roux, G.A.; Sotelo, F.F., Caracterização do petróleo bruto através da espectroscopia NIR e da quimiometria, *XV Congresso Brasileiro de Engenharia Química (COBEQ)*, Curitiba - PR, 2004.

Chung, H.; Ku, M-S.; Lee, J-S., Comparation of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene, *Vib. Spectrosc.*,vol. 20, p. 155-163, 1999.

DMA 4500, Density; Specific Gravity; Concentration Meter, Instruction Manual. Anton Paar GMBH, Áustria, 2003.

FTLA2000, ABB BOMEM Inc., Series Laboratory FT-IR Spectrometers – revision 1-2, november-2002.

Hair, J.F.; et al., Multivariate data analysis with readings. 5th edition, Upper Saddle River, N.J., *Prentice Hall*, 1998.

Honigs, D.E.; Hirschfeld, T.; Hieftje, G.M., Near-Infrared Determination of Several Physical Properties of Hydrocarbons, *Anal. Chem.*, vol.57, p. 443-445, 1985.

Johnson, R. A.; Wichern, D. W. Applied multivariate statistical analysis. Fourth edition New Jersey, *Prentice Hall Inc.*, Upper Saddle River, 1998.

Kelly, J.J.; Callis, J.B. Nondestructive analytical procedure for simultaneous estimation of major classes of hydrocarbon constituents of finished gasolines. *Analytical Chemistry*, vol. 62, nº 14, p. 1444-1451,1990.

Kemsley, E.K., Discriminant analysis of high-dimensional data: a comparation of principal components analysis and partial least squares data reduction methods, *Chemometrics Intelligent Laboratory Systems*, vol.33, p. 47-61, 1996.

Khattree, R.; Naik, D.N., Multivariate data reduction and discrimination with SAS® software, *Wiley Inter-Science*, 2000.

Kim, M.; Lee, Y.-H.; Han, C., Real-time classification of petroleum products using near-infrared spectra, *Computers and Chemical Engineering*, vol. 24, p. 513-517, 2000.

Litani-Barzilai, I.; Sela, I.; Bulatov, V.; Zilberman, I.; Schechter, I. On-line remote prediction of gasoline properties by combined optical methods. *Analytica Chimica Acta*, vol. 339, p. 193-199, 1997.

Lu, J.; Plataniotis, K.N.; Venetsanopoulos, A.N., Regularized discriminat analysis for the small sample size problem in face recognition, *Elsevier Science*, p. 1-14, 2003.

Martens, H.; Naes, T., Multivariate Calibration, *John Wiley & Sons*, 1989.

MB 457 Combustível – Determinação das características antidetonantes – Índice de octano – Método motor.

McClure, W. F., 204 years of near infrared technology: 1800-2003, *J. Near Infrared Spectrosc.*, vol. 11, p. 487-518, 2003.

NBR 13992 - Determinação do teor de álcool etílico anidro combustível (AEAC) existente em gasolina automotiva, *ABNT*, 1997.

NBR 14065 – Destilados de Petróleo e Óleos Viscosos – Determinação da Massa Específica e Densidade Relativa pelo Densímetro Digital, *ABNT*, 1998.

NBR 7148 – Petróleo e produtos de petróleo – Determinação da massa específica, densidade relativa e ºAPI – Método do densímetro, *ABNT*, 2000.

NBR 9619 Produtos de petróleo – Determinação das propriedades de destilação, *ABNT*, 1998.

Otto, M., Chemometrics: Statistics and Computer Application in Analytical Chemistry, *Wiley-VCH*, 1999.

Pavia, D.L.; Lampman, G.M.; Junior, G.S.K. Introduction to Spectroscopy: A guide for students of organic chemitry, *Saunders Golden Sunburst Series*, 1979.

Petrospec, *Manual de operação do GS1000*, 1999.

PLS_Toolbox version 3, *Eigenvector Research*, Inc., 2002.

Sharma, S., Applied multivariate techniques, *Jonh Wiley & Sons*, INC., 1996.

Siesler, H.W.; Ozaki, Y.; Kawata, S.; Heise, H.M., Near-Infrared Spectroscopy Principles, Instruments, Applications, *Wiley-VCH*, 2002.

Stanimirova, I.; Walcsak, B.; Massart,D.L.; Simeonov, V., A comparation between two robust PCA algorithms, *Chemometrics Intelligent Laboratory Systems*, vol. 71, p. 83-95, 2004.

Tominaga, Y. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and *k*-NN, *Chemometrics and Intelligent Laboratory System* vol. 49, p. 105-115, 1999.

Wu, W.; Massart, D.L.; Jong, S., Kernel-PCA algorithms for wide data. Part I: fast cross-validation and application in classification of NIR data, *J. Chemometrics and Intelligent Laboratory System,* vol. 37, p. 271-280, 1997.

Wu, W.; Walczak, B.; Penninckx, W.; Massart, D.L., Comparation of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data, *Analytica Chimica Acta*, vol. 329, p. 257-265, 1996.

ANEXO I

Tabela das amostras utilizadas na etapa de calibração para a concepção dos modelos, conforme descrito no capítulo 4 (Item 4.1.1 e Item 4.3 a 4.6).

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
1	Conforme	---
2	Não Conforme	Marcador
3	Não Conforme	Marcador, RON e IAD
4	Não Conforme	MON, RON, IAD e Destilação (T10%, T50% e T90%)
5	Não Conforme	Marcador
6	Conforme	---
7	Conforme	---
8	Não Conforme	Marcador, MON, RON e IAD
9	Não Conforme	Marcador
10	Conforme	---
11	Conforme	---
12	Não Conforme	Marcador, RON e IAD
13	Conforme	---
14	Conforme	---
15	Conforme	---
16	Conforme	---
17	Não Conforme	Marcador (limite de detecção do equipamento)
18	Conforme	---
19	Conforme	---
20	Conforme	---
21	Conforme	---
22	Conforme	---
23	Conforme	---
24	Conforme	---
25	Conforme	---
26	Não Conforme	Marcador
27	Conforme	---
28	Conforme	---
29	Conforme	---
30	Conforme	---
31	Não Conforme	Marcador e Destilação (T10%)
32	Conforme	---
33	Conforme	---
34	Não Conforme	Marcador (limite de detecção do equipamento) e Teor de Álcool
35	Conforme	---
36	Não Conforme	Marcador, RON e Destilação (T10% e T90%)
37	Conforme	---
38	Não Conforme	Marcador, MON, RON e IAD
39	Conforme	---
40	Conforme	---
41	Conforme	---
42	Conforme	---

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
43	Conforme	---
44	Conforme	---
45	Conforme	---
46	Conforme	---
47	Conforme	---
48	Não Conforme	Teor de Álcool - 27% (suspeita)
49	Conforme	---
50	Conforme	---
51	Conforme	---
52	Conforme	---
53	Conforme	---
54	Conforme	---
55	Conforme	---
56	Conforme	---
57	Conforme	---
58	Conforme	---
59	Conforme	---
60	Conforme	---
61	Conforme	---
62	Conforme	---
63	Conforme	---
64	Conforme	---
65	Conforme	---
66	Conforme	---
67	Conforme	---
68	Conforme	---
69	Conforme	---
70	Conforme	---
71	Conforme	---
72	Conforme	---
73	Conforme	---
74	Conforme	---
75	Conforme	---
76	Não Conforme	Marcador
77	Conforme	---
78	Conforme	---
79	Conforme	---
80	Conforme	---
81	Conforme	---
82	Conforme	---
83	Conforme	---
84	Não Conforme	Marcador, RON, IAD e Destilação (T10% e PF)
85	Conforme	---
86	Conforme	---
87	Não Conforme	Marcador e RON
88	Conforme	---
89	Conforme	---
90	Conforme	---
91	Não Conforme	Marcador, RON, IAD e Destilação (T10%, T90% e PF)
92	Não Conforme	Marcador e IAD
93	Conforme	---

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
94	Conforme	---
95	Conforme	---
96	Não Conforme	Marcador, Teor de Álcool, RON, IAD e destilação (PF)
97	Conforme	---
98	Conforme	---
99	Conforme	---
100	Conforme	---
101	Não Conforme	Teor de Álcool, MON, RON, IAD e Destilação (T90%)
102	Conforme	---
103	Conforme	---
104	Conforme	---
105	Conforme	---
106	Não Conforme	Marcador
107	Conforme	---
108	Conforme	---
109	Não Conforme	Marcador, Teor de Álcool, RON e destilação (T90%)
110	Conforme	---
111	Não Conforme	Teor de Álcool
112	Conforme	---
113	Conforme	---
114	Não Conforme	Marcador
115	Conforme	---
116	Não Conforme	Marcador
117	Conforme	---
118	Conforme	---
119	Não Conforme	Marcador e RON
120	Conforme	---
121	Não Conforme	Marcador
122	Não Conforme	Marcador e RON
123	Conforme	---
124	Conforme	---
125	Conforme	---
126	Não Conforme	Teor de Álcool
127	Conforme	---
128	Não Conforme	Marcador
129	Conforme	---
130	Conforme	---
131	Conforme	---
132	Não Conforme	Marcador, RON, IAD e Destilação (T10%)
133	Conforme	---
134	Conforme	---
135	Conforme	---
136	Conforme	---
137	Conforme	---
138	Conforme	---
139	Conforme	---
140	Não Conforme	Marcador
141	Não Conforme	Marcador, RON e Destilação (T10%)
142	Conforme	---
143	Conforme	---

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
144	Não Conforme	Marcador, Teor de Álcool, MON, RON, IAD e Destilação (T10% e T90%)
145	Não Conforme	Marcador
146	Conforme	---
147	Conforme	---
148	Não Conforme	Marcador, Teor de Álcool (28%) e RON
149	Conforme	---
150	Conforme	---
151	Conforme	---
152	Não Conforme	Teor de Álcool (28%)
153	Conforme	---
154	Não Conforme	Marcador
155	Conforme	---
156	Não Conforme	Marcador
157	Conforme	---
158	Conforme	---
159	Conforme	---
160	Conforme	---
161	Conforme	---
162	Não Conforme	Marcador, RON, IAD e Destilação (T10%)
163	Não Conforme	Marcador e RON
164	Não Conforme	Marcador e RON
165	Não Conforme	Marcador e RON
166	Não Conforme	Marcador e RON
167	Conforme	---
168	Não Conforme	Marcador, RON e Destilação (T10%)
169	Conforme	---
170	Não Conforme	Marcador e RON
171	Não Conforme	Marcador, RON e Destilação (T10%)
172	Não Conforme	Marcador, RON e Destilação (T10%)
173	Não Conforme	Teor de Álcool, RON e IAD
174	Não Conforme	Marcador e RON
175	Conforme	---
176	Conforme	---
177	Não Conforme	Marcador
178	Não Conforme	Marcador, RON e Destilação (T10%)
179	Não Conforme	Teor de Álcool, RON e IAD
180	Não Conforme	Marcador e RON
181	Não Conforme	Marcador
182	Conforme	---
183	Não Conforme	Marcador
184	Conforme	---
185	Não Conforme	Marcador, Teor de Álcool e Destilação (T10%)
186	Não Conforme	Marcador, RON e Destilação (T10% e PF)
187	Conforme	---
188	Não Conforme	Marcador
189	Não Conforme	Marcador
190	Não Conforme	Marcador e Destilação (PF)
191	Conforme	---
192	Conforme	---
193	Conforme	---

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
194	Não Conforme	Marcador
195	Não Conforme	Marcador (limite de detecção do equipamento)
196	Conforme	---
197	Conforme	---
198	Conforme	---
199	Conforme	---
200	Conforme	---
201	Conforme	---
202	Conforme	---
203	Conforme	---
204	Conforme	---
205	Conforme	---
206	Conforme	---
207	Conforme	---
208	Conforme	---
209	Não Conforme	Marcador (limite de detecção do equipamento)
210	Não Conforme	Marcador
211	Conforme	---
212	Não Conforme	Teor de Álcool (0%)
213	Não Conforme	Teor de Álcool (0%)
214	Não Conforme	Teor de Álcool (0%)
215	Não Conforme	Teor de Álcool (0%)
216	Não Conforme	Teor de Álcool (0%)

ANEXO II

Tabela das amostras utilizadas na validação dos modelos, conforme descrito no capítulo 4 (Item 4.7).

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
1	Não conforme	Marcador e RON
2	Não conforme	Marcador, Teor de Álcool e Destilação (T10%)
3	Conforme	---
4	Não conforme	Teor de Álcool, MON, RON, IAD e Destilação (T10%)
5	Não conforme	Marcador, RON e Destilação (T10%)
6	Não conforme	Marcador, Teor de Álcool e RON
7	Não conforme	Marcador e Teor de Álcool
8	Não conforme	Marcador
9	Conforme	---
10	Conforme	---
11	Não conforme	Marcador (limite de detecção do equipamento)
12	Conforme	---
13	Conforme	---
14	Conforme	---
15	Conforme	---
16	Não conforme	Marcador
17	Conforme	---
18	Conforme	---
19	Não conforme	Marcador
20	Conforme	---
21	Não conforme	Marcador e RON
22	Não conforme	Marcador e RON
23	Conforme	---
24	Conforme	---
25	Conforme	---
26	Conforme	---
27	Conforme	---
28	Conforme	---
29	Conforme	---
30	Conforme	---
31	Conforme	---
32	Conforme	---
33	Conforme	---
34	Conforme	---
35	Não conforme	Marcador
36	Conforme	---
37	Conforme	---
38	Conforme	---
39	Conforme	---
40	Conforme	---
41	Não conforme	Marcador, RON e Destilação (T10%)
42	Conforme	---
43	Conforme	---

Número da amostra	Classificação	Ensaios físico-químicos que detectaram a não conformidade
44	Conforme	---
45	Conforme	---
46	Não conforme	Marcador, RON e IAD
47	Conforme	---
48	Não conforme	Marcador (limite de detecção do equipamento)
49	Não conforme	Marcador, RON e IAD
50	Não conforme	Marcador, RON e IAD

ANEXO III

Os algoritmos para os classificadores LDA e QDA apresentados a seguir, foram implementados no programa Matlab e foram utilizados para a classificação das amostras de gasolinas comerciais tanto pelos resultados físico-químicos quanto pelos dados NIR (PCA e PLS).

Algoritmo para Análise Discriminante Linear (LDA)

Normalização dos dados

```
mínimo=min(A);
máximo=max(A);
[n m]=size(A);

for i=1:m
norma(:,i)=2*(A(:,i)-(mínimo(i)+máximo(i))/2*ones(n,1))/(-mínimo(i)+máximo(i));
end
```

Primeira etapa:Cálculo da média de cada classe

```
ib=0; (gasolina boa)
ir=0; (gasolina ruim)
for i=1:n
   if norma(i,m) == -1
ir=ir+1;
      x0(ir,:)=norma(i,1:m-1);
  else
      ib=ib+1;
      x1(ib,:)=norma(i,1:m-1);
  end
end

mediac0=mean(x0); (gasolina ruim)
mediac1=mean(x1); (gasolina boa)
```

Segunda etapa: Cálculo das variâncias de cada classe

Para variável 0 (não conforme)

```
var0=zeros(m-1,m-1);
for i=1:ir
   var0=var0+((x0(i,1:m-1)-mediac0(1,1:m-1))'*(x0(i,1:m-1)-mediac0(1,1:m-1)));

end
var0=var0/n;
pi0=ir/n;
```

Para variável 1 (conforme)

```
var1=zeros(m-1,m-1);
for i=1:ib

var1=var1+((x1(i,1:m-1)-mediac1(1,1:m-1))'*(x1(i,1:m-1)-mediac1(1,1:m-1)));
end
var1=var1/n;
pi1=ib/n;
```

Etapa final: Cálculo do score de classificação

varpool=(var0+var1)/n (variância combinada)

```
for i=1:n
cf0(i)=(norma(i,1:m-1)-mediac0(1,1:m-1))*inv(varpool)*(norma(i,1:m-1)-mediac0(1,1:m-1))'-
2*log(pi0);
cf1(i)=(norma(i,1:m-1)-mediac1(1,1:m-1))*inv(varpool)*(norma(i,1:m-1)-mediac1(1,1:m-1))'-
2*log(pi1);
```

Classificação das amostras (quais não tiveram classificação correta e quantidade - k)

```
if cf0(i) < cf1(i)
      classe(i)=-1;
   else
      classe(i)=1;
   end
end
numero=(1:n)';
```

Continuação da etapa final

```
difer=(cf0-cf1)/1e3;
[numero norma(:,m) classe' cf0'/1e3 cf1'/1e3 difer']
k=0
for i=1:n
   deu=norma(i,m)*classe(i);

if deu < 0
      k=k+1;
      i
      pause
end
end
k
```

Algoritmo para Análise Discriminante Quadrática (QDA)

O que diferencia o método QDA do LDA é a matriz de variância que é diferente para o grupo conforme e não conforme. No entanto como a coluna correspondente aos resultados do marcador é praticamente zero para os dados considerados conforme, a matriz de variância torna-se singular e torna o cálculo do *score* impossível. Para evitar este problema, a matriz de dados utilizada na análise QDA não leva em conta as informações referentes aos resultados do marcador. A primeira e a segunda etapa segue os mesmos passos da LDA.

Etapa final: Cálculo do score de classificação

```
logdet0=log(det(var0));
logdet1=log(det(var1));
for i=1:n
cf0(i)=(norma(i,1:m-2)-mediac0(1,1:m-2))*inv(var0)*(norma(i,1:m-2)-mediac0(1,1:m-2))'+logdet0-
2*log(pi0);
 cf1(i)=(norma(i,1:m-2)-mediac1(1,1:m-2))*inv(var1)*(norma(i,1:m-2)-mediac1(1,1:m-2))'+logdet1-
2*log(pi1);
```

Continuação da etapa final

Classificação das amostras (quais não tiveram classificação correta e quantidade - k)

```
 if cf0(i) < cf1(i)
       classe(i)=-1;
   else
       classe(i)=1;
   end
end
numero=(1:n)';

difer=(cf0-cf1)/1e3;
[numero norma(:,m) classe' cf0'/1e3 cf1'/1e3 difer']
k=0
for i=1:n
   deu=norma(i,m)*classe(i);
   if deu < 0
      k=k+1;
i
      pause
   end
end
k
```