UNIVERSIDADE DE SANTA CRUZ DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E PROCESSOS INDUSTRIAIS – MESTRADO
ÁREA DE CONCENTRAÇÃO EM
CONTROLE E OTIMIZAÇÃO DE PROCESSOS INDUSTRIAIS

Gustavo Post Sabin

OTIMIZAÇÃO DE MODELOS DE REGRESSÃO MULTIVARIADOS EMPREGANDO MÉTODOS DE SELEÇÃO DE VARIÁVEIS

Santa Cruz do Sul, março de 2007.

Gustavo Post Sabin

OTIMIZAÇÃO DE MODELOS DE REGRESSÃO MULTIVARIADOS EMPREGANDO MÉTODOS DE SELEÇÃO DE VARIÁVEIS

Dissertação apresentada ao Curso de Pós-Graduação em Sistemas e Processos Industriais – Mestrado – da Universidade de Santa Cruz do Sul, para a obtenção do título de Mestre em Sistemas e Processos Industriais.

Orientador: Prof. Dr. Marco Flôres Ferrão

Co-orientador: Prof. Dr. João Carlos Furtado

Santa Cruz do Sul, março de 2007.

Gustavo Post Sabin

OTIMIZAÇÃO DE MODELOS DE REGRESSÃO MULTIVARIADOS EMPREGANDO MÉTODOS DE SELEÇÃO DE VARIÁVEIS

Esta dissertação foi submetida ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado, Área de Concentração em Controle e Otimização de Processos Industriais, Universidade de Santa Cruz do Sul – UNISC, como requisito para a obtenção do título de Mestre em Sistemas e Processos Industriais.

Dr. Marco Flôres Ferrão
Professor Orientador

Dr. João Carlos Furtado
Professor Co-orientador

Dr. Jarbas José Rodrigues Rohwedder

Dr. Rolf Fredi Molz

AGRADECIMENTOS

Inicialmente agradeço ao meu orientador, Prof. Dr. **Marco Flôres Ferrão**, por ter sugerido este projeto, auxiliado na escolha de uma bibliografia adequada, pela orientação e paciência na correção das minhas falhas.

Agradeço também ao meu co-orientador, **João Carlos Furtado**, por ter aceitado despender do seu tempo para auxiliar na elaboração do projeto em questão.

Gostaria de agradecer também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – **CAPES** pela bolsa concedida, viabilizando minha participação neste Programa de Pós-Graduação.

Agradeço à Universidade de Santa Cruz do Sul – **UNISC** e a todos os docentes e colaboradores do **Programa de Pós-Graduação em Sistemas e Processos Industriais** que de alguma forma contribuíram para esta pesquisa. Faz-se necessário agradecer à secretaria deste programa de pós-graduação e, especialmente, às secretarias Janaina Iochims Ramires e Claudia de Souza Lopes que, por incontáveis vezes, foram prestativas e eficientes, ultrapassando o limite de suas obrigações para que os mestrandos não se desviassem de suas pesquisas mais do que o necessário, para realizarem atividades burocráticas.

Os demais mestrandos deste programa de pós-graduação com quem tive contato também merecem os meus mais sinceros agradecimentos, pois foram muito mais que colegas e, ao contrário do esperado, formaram um conciso grupo de amigos. Valeu pessoal!

A minha família (ao contrário do que dizem, considero meu sogro, minha sogra e meu cunhado como parte dela) pelo suporte psicológico, financeiro e afetivo. Sem vocês eu nem se quer teria concluído a graduação.

Apesar de sua contribuição não ser direta, é pertinente que eu agradeça meu irmão, **Guilherme Post Sabin**, por ter me motivado a ingressar neste mestrado e por auxiliar no entendimento de alguns conceitos referentes à quimiometria, além de servir como referência de profissionalismo e capacidade intelectual. Não teria concluído essa pesquisa sem o teu apoio “meu irmão mais buchudo”!

Agradeço ao meu primo-*irmão*, **Leônidas Post Ferreira**, que apesar de não ter contribuído com a pesquisa, foi uma companhia agradável e necessária nos momentos mais difíceis prestando o famoso “apoio moral” e ajudando a desopilar os meus pensamentos, geralmente com muita cerveja.

Agradeço à minha mãe, avô e avó, **Juçara Post Sabin**, **Frederico Post** e **Sueli Lia Post** respectivamente, por me ensinarem a ser um homem estudioso e honesto. Esta tarefa também foi dividida com o meu pai, **Ernani Maydana Sabin**, que se estivesse entre nós, tenho certeza que ficaria orgulhoso dos seus filhos como sempre o fez durante a vida. Saudades...

Não posso deixar de mencionar, com imensa satisfação, os ex-colegas de graduação e eternos amigos **Fábio** “o insano” **Pasini**, **André** “o grande” **de Almeida Barros**, **Guilherme** “o alemão” **Frederico Rohde** e **Cássio Soares Carvalho** pela amizade e companheirismo sem limites, além de terem me ensinado muito de computação durante a época da universidade. A parceria é infinita mesmo! **Luciano Vaghetti de Oliveira**, **Jean Paulo Sandri Orengo** e **Marcelo Santos Linder**, apesar do distanciamento, também merecem reconhecimento pela influência positiva que exerceram na minha vida.

Por último e mais importante, ao meu grande e único amor, **Eveline do Amor Divino**, companheira de todas as horas. Obrigado pela compreensão, afeto, companheirismo e auxílio nas correções de português. Te amo!

RESUMO

Há um grande aumento na utilização de técnicas de espectroscopia no infravermelho para análises químicas na indústria, devido à rapidez, baixo custo e mantém a integridade das amostras neste tipo de análise. Com isso, é desejável um estudo de técnicas de obtenção de espectros no infravermelho, de regressão multivariadas e de métodos de seleção de variáveis. Esta dissertação tem como objetivo o estudo e implementação de um algoritmo genético, aliado a técnica de regressão multivariada de mínimos quadrados parciais por intervalo (iPLS), capazes de selecionar as variáveis mais pertinentes a propriedade que se deseja medir e assim criar modelos de regressão multivariados mais robustos. Nesta pesquisa efetuou-se a determinação de hidroxilas de polióis de óleo de soja, onde os resultados obtidos foram 14,97% menores em relação ao erro de predição e 15,63% menores em relação ao erro médio percentual dos valores calculados para as amostras de predição em comparação com os resultados encontrados através do método iPLS. Também se fez a determinação de cloridrato de propranolol em comprimidos, onde os resultados obtidos foram 76,1% menores em relação ao erro de predição e 73,99% menores em relação ao erro médio percentual dos valores calculados para as amostras de predição em comparação com os resultados encontrados através do método iPLS. Observando tais valores, pode-se concluir que a utilização de algoritmos genéticos conjuntamente com o método iPLS foi capaz de otimizar as soluções, selecionando de forma eficiente as variáveis espectrais envolvidas, encontrando modelos mais preditivos e robustos.

Palavras-chave: algoritmos genéticos, mínimos quadrados parciais por intervalo, otimização combinatorial, métodos heurísticos, espectroscopia no infravermelho

ABSTRACT

It has a great increase in the use of infrared spectroscopy techniques for chemical analyses in the industry, due to rapidity, low cost and preservation of the samples in this kind of analysis. Thus it is desirable a study of infrared spectra acquisition techniques, multivariate regression and variable selection methods. This dissertation has as objective the study and implementation of a genetic algorithm, jointly with interval partial least-squares multivariate regression technique (iPLS), capable to select the variables most pertinent the property that if it desires to measure and to create more robust multivariate regression models. In this research was made the determination of hydroxyl value of hydroxylated soybean oils, where the gotten results had been 14.97% betters in relation to the prediction error and 15.63% betters in relation to the average percentile error of the values calculated for the prediction samples in comparison with the results found through the iPLS method. Also it was made the propranolol hydrochloride determination in tablets, where the gotten results had been 76.1% betters in relation to the prediction error and 73.99% betters in relation to the average percentile error of the values calculated for the prediction samples in comparison with the results found through the iPLS method. Observing such values, it can be concluded that the use of genetic algorithms jointly with the iPLS method was capable to optimize the solutions, selecting of efficient way the involved spectral variables, finding more predictive and robust models.

Keywords: genetic algorithm, interval partial least-squares, combinatorial optimization, heuristic methods, infrared spectroscopy

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AOCS – American Oil Chemists' Society

ATR – Attenuated Total Reflectance

DRIFTS – Diffuse Reflectance Infra-red Fourier Transform

FT-IR – Fourier Transform – Infra-red

GA – Genetic Algorithm

GA-iPLS – Interval Partial Least-Squares Regression with Genetic Algorithm

HATR – Horizontal Attenuated Total Reflectance

iPLS – Interval Partial Least-Squares

MSC – Multiplicative Scatter Correction

PLS – Partial Least-Squares

RMSE – Root Mean Square Error

RMSEC – Root Mean Square Error of Calibration

RMSECV – Root Mean Square Error of Cross Validation

RMSEP – Root Mean Square Error of Prediction

RMSEV – Root Mean Square Error of Validation

SUMÁRIO

1 INTRODUÇÃO

Para que uma indústria obtenha destaque em seu ramo de atividades, são necessários um baixo custo de produção e um controle no processo que garanta a qualidade do produto desenvolvido. A utilização de ferramentas que possibilitem um controle rigoroso sobre a qualidade do produto e tenham um baixo custo operacional pode ser a diferença entre uma indústria líder no seu setor e uma mera concorrente sem muita expressão no mercado.

Neste sentido, muitas indústrias, como as químicas, as de alimentos, as farmacêuticas, etc, vem cada vez mais utilizando a espectroscopia no infravermelho como alternativa para realizar as análises inerentes aos seus produtos.

Este tipo de espectroscopia é capaz de obter informações da amostra através do uso da radiação infravermelha. A obtenção de dados sobre a estrutura das moléculas presentes em um determinado composto é feita a partir da análise de algumas bandas dessa radiação detectadas pelo equipamento, que são características de certos grupos de átomos.

Existem vários tipos de espectroscopia, este trabalho tem por foco a espectroscopia no infravermelho médio, por ser bastante difundida e apresentar um grande número de sinais de vários grupos funcionais presentes nos compostos ou misturas que constituem os insumos ou produtos da indústria.

1.1 Justificativa

Muitas indústrias que necessitam de análises químicas e físico-químicas podem encontrar na espectroscopia no infravermelho uma solução de baixo custo, rápida, com uma boa precisão e que não gera resíduos em suas análises, contribuindo assim com uma questão que pesa cada vez mais nas decisões das indústrias, a preservação ambiental.

Outra vantagem deste tipo de análise é a possibilidade de manter a integridade das amostras, já que outros métodos bastante utilizados, como a cromatografia e a titulação, provocam a destruição das amostras analisadas.

A utilização de espectroscopia combinada com o tratamento dos dados por métodos quimiométricos é bastante utilizada em laboratórios, possuindo uma bibliografia abundante de suas aplicações para a identificação e quantificação dos componentes de uma amostra, principalmente quando a espectroscopia no infravermelho é aliada a métodos como o de Reflexão Total Atenuada (ATR), como referenciado por Costa Filho & Poppi (2002), Borin & Poppi (2004) e Christy & Egeberg (2006) ou Reflexão Difusa no Infravermelho com Transformada de Fourier (DRIFTS), conforme Konzen *et al* (2003).

A decisão de utilizar algoritmo genético como forma de auxiliar na escolha das variáveis mais significativas para o modelo analisado deve-se a sua capacidade de guiar o processo de busca por melhores soluções, sem necessidade de avaliar todas as soluções possíveis. Este método, auxiliando a Regressão por Mínimos Quadrados Parciais (PLS), tem sido aplicado com sucesso em muitos casos, como citado por Ferrão *et al* (2004).

1.2 Objetivos

1.2.1 Objetivo geral

Tendo em vista a necessidade das indústrias em agilizar as suas análises, esta pesquisa tem como objetivo o estudo de técnicas de otimização empregadas na quimiometria através de levantamento bibliográfico e a sua implementação.

1.2.2 Objetivos específicos

A construção do algoritmo genético aplicado em conjunto com o iPLS (GA-iPLS) para determinar as regiões espectrais mais representativas à análise, também é objetivo almejado por esta pesquisa.

Também se deseja avaliar o comportamento de tais ferramentas em comparação com a metodologia oficial, que depende do tipo de propriedade ou substância que se deseja analisar, assim verificando a pertinência das soluções alcançadas. As metodologias utilizadas neste trabalho serão descritas quando os problemas estudados forem abordados.

Por fim, através da execução deste projeto, visa-se realizar um estudo comparativo entre os métodos propostos, verificando qual deles oferece mais vantagens em termos de convergência para uma solução otimizada.

1.3 Organização do texto

Esta dissertação está dividida em cinco capítulos: o primeiro capítulo introduz a análise espectroscópica, explanando sobre as aplicações deste tipo de análise, ainda justificando e mostrando o foco desta dissertação.

No Capítulo Dois apresenta-se a fundamentação teórica deste trabalho: métodos de regressão multivariada (PLS e iPLS) e o paradigma da computação evolutiva através dos Algoritmos Genéticos.

O Capítulo Três aborda a metodologia que foi utilizada nesta pesquisa e descreve a implementação detalhada da ferramenta desenvolvida (GA-iPLS) e das diferentes formas de avaliação das soluções.

Já no Capítulo Quatro são apresentados os problemas abordados nesta dissertação, explicando a importância desse estudo, informando detalhes sobre a aquisição dos espectros e revelando e analisando os resultados obtidos com o algoritmo aqui desenvolvido.

No Quinto e último capítulo são apresentadas algumas conclusões deste trabalho, expondo algumas dificuldades encontradas no decorrer desta pesquisa e as perspectivas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Como mencionado anteriormente, a espectroscopia por reflexão no infravermelho médio com transformada de Fourier ou a espectroscopia no infravermelho com reflexão total atenuada, aliada a técnicas de quimiometria, vem sendo cada vez mais utilizada (FERRÃO, 2004).

Para o desenvolvimento da presente pesquisa, torna-se necessária a explanação de alguns assuntos que dizem respeito à quimiometria, computação evolutiva e aos dados obtidos via instrumentação analítica.

2.1 Espectroscopia no infravermelho com transformada de Fourier (FT-IR)

A região espectral do infravermelho compreende radiação de números de onda no intervalo de aproximadamente 12800 a 10 cm^{-1}. Para fins de instrumentação, o espectro infravermelho é dividido em radiação do infravermelho próximo (12800 a 4000 cm^{-1}), médio (4000 a 200 cm^{-1}) e distante (200 a 10 cm^{-1}) (SKOOG *et al*, 2005).

Segundo Skoog *et al* (2005), a espectroscopia do tipo FT-IR é bastante utilizada atualmente porque possui poucos elementos óticos e nenhuma fenda para atenuar a radiação, tendo assim uma maior potência desta radiação incidindo no detector, tornando a relação sinal-ruído muito melhor. Esta vantagem é conhecida como eficiência de transporte ou vantagem de Jaquinot.

Outra vantagem deste tipo de instrumento de espectroscopia, se comparado com outros, é o seu alto poder de resolução e reprodutibilidade do comprimento de onda, o que torna possível a análise de espectros complexos.

Além destas duas vantagens da espectroscopia do tipo FT-IR, pode-se citar a sua rapidez, pois todos os elementos da fonte atingem o detector simultaneamente, possibilitando a obtenção de dados para um espectro inteiro em menos de um segundo (SKOOG *et al*, 2005).

Por outro lado, quando a amostra a ser analisada contém material biológico a espectroscopia no infravermelho médio pode ser problemática, pois esse tipo de material é opaco e contém uma quantidade elevada de água, podendo apresentar grande espalhamento de luz. Para resolver este problema, a utilização de reflectância difusa tem facilitado este tipo de análise.

2.2 Espectroscopia por reflexão difusa no infravermelho médio com transformada de Fourier (DRIFTS)

A reflexão difusa ocorre em superfícies não totalmente planas (como por exemplo, na forma de pó). Nestes casos, o feixe incidente penetra na superfície da amostra interagindo com a matriz. Depois de uma absorção parcial, ocorrem espalhamentos deste feixe retornando a superfície da amostra. Este efeito é ilustrado na Figura 1.

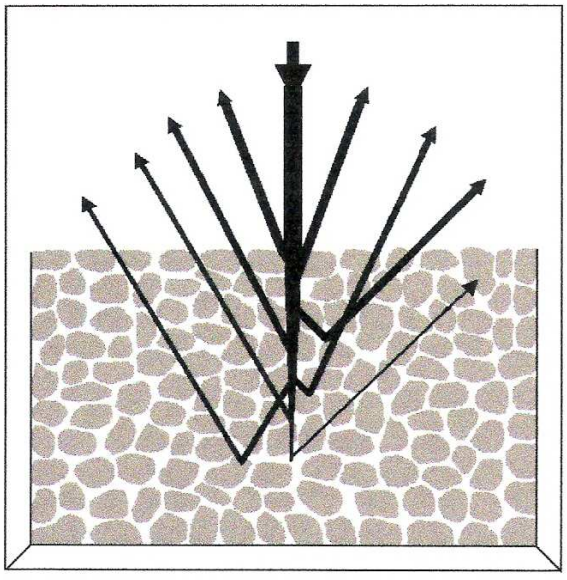


Figura 1 - Reflexão especular e difusa de uma onda eletromagnética em uma amostra.

Fonte: WETZEL, 1983

A luz refletida pela difusão no meio é composta por reflexão difusa e especular, mas para fins quimiométricos, apenas a reflexão difusa fornece informações relevantes, enquanto a reflexão especular pode causar algumas anomalias no espectro obtido, como o deslocamento de bandas.

Os diferentes tamanhos de partículas também podem afetar os resultados da análise, mas para reduzir estes efeitos indesejáveis existem algumas técnicas, tais como a transformação do espectro com a primeira e a segunda derivadas ou Correção do Espalhamento da Luz (MSC). Neste caso, somente o MSC será abordado por já estar disponível no software utilizado nesta pesquisa.

O MSC é o processo matemático que visa corrigir o espalhamento da luz presente nos espectros obtidos por técnicas de reflexão, pois este fenômeno altera a relação entre a intensidade da medida de reflexão e a concentração das espécies absorventes de uma matriz. Está correção é feita com base no espalhamento médio de todos os espectros formadores da matriz de dados **X**, retendo os resíduos e as informações químicas (FURTADO, 2002) (FERRÃO, 2001) (ZENI, 2005).

2.3 Reflexão total atenuada

Outro problema pode ocorrer quando o material (objeto da análise) é espesso e fortemente absorvente. Neste caso pode-se adotar o método de Reflexão Total Atenuada (ATR), elaborada por Fahrenfort em 1961, que é bastante rápido e não requer muita preparação da amostra. Este método tem a vantagem de obter espectros de materiais líquidos, sólidos e viscosos para muitos tipos diferentes de amostras (FERRÃO, 2001).

O ATR caracteriza-se por múltiplas reflexões da radiação infravermelha que ocorrem dentro de cristais com alto índice de refração (cristal ATR), interagindo somente com o material que estiver em contato com este cristal. Quando a radiação infravermelha passa através deste cristal e atinge a amostra, de densidade menor que a do cristal, ocorre uma reflexão de parte da radiação criando uma onda evanescente. Desta forma a amostra pode absorver a radiação incidente atenuando a sua intensidade, dando origem assim ao espectro

infravermelho. Exemplos de aplicações e mais detalhes sobre o ATR são descritos por Ferrão no artigo Técnicas de Reflexão no Infravermelho Aplicadas na Análise de Alimentos (FERRÃO, 2001).

Segundo Skoog (2002), através dessa técnica torna-se possível a análise quantitativa de amostras como pós, sólidos pouco solúveis e pastas, que são difíceis de serem analisadas através da espectroscopia por transmissão.

O cristal ATR pode ser de diferentes materiais e, segundo Ferrão (2001), a escolha deste cristal pode resultar em distorções da banda do espectro. Entre os materiais mais utilizados está o ZnSe, com baixo índice de refração e faixa de utilização de 20.000 até 650 cm^{-1}, e o Si, com alto índice de refração e faixa de utilização de 9.000 até 400 cm^{-1}.

Utilizando-se essa técnica, podem ocorrer alterações na intensidade das bandas devido à variação no contato da amostra com o cristal. Através deste efeito podemos obter informações sobre propriedades ou condições da superfície da amostra analisada (FERRÃO, 2001).

2.4 Regressão por mínimos quadrados parciais (PLS)

Para análises instrumentais, a calibração ou regressão é uma ferramenta poderosa e métodos de regressão do espectro inteiro, como o PLS, tem uma documentação abundante que comprova a sua eficiência em análises espectrais. Este procedimento tem o objetivo de descrever as relações quantitativas existentes entre as variáveis.

A calibração consiste em duas etapas. A primeira é a obtenção de padrões, através de medidas realizadas em uma série de amostras de concentrações conhecidas (etapa descritiva). A segunda utiliza o modelo obtido na etapa descritiva para prever as concentrações de novas amostras (etapa preditiva) (KONZEN *et al*, 2002).

O PLS foi desenvolvido por Herman Wold *apud* Konzen *et al* (2002) e trabalha simultaneamente com as informações espectrais e as concentrações no processo de calibração.

Este método de regressão está baseado na decomposição da matriz de dados X em várias matrizes M e uma matriz de resíduos E que corresponde ao erro, como demonstrado na Equação 1.

$$\mathbf{X} = \mathbf{M_1} + \mathbf{M_2} + ... + \mathbf{M_n} + \mathbf{E} \quad (1)$$

Na equação 1, **n** corresponde ao número de componentes principais ou variáveis latentes selecionadas para truncar a igualdade.

O PLS relaciona a matriz espectral dos padrões (**X**) com a matriz dos dados das concentrações (**Y**) como visto nas Equações 2 e 3.

$$\mathbf{X} = \mathbf{T}\,\mathbf{P^t} + \mathbf{E} \quad (2)$$

$$\mathbf{Y} = \mathbf{U}\,\mathbf{Q^t} + \mathbf{F} \quad (3)$$

Nas equações 2 e 3, **T** e **U** são, respectivamente, os *scores* de **X** e **Y, P** é o *loading* de **X** e **Q** é o *loading* de **Y**. As matrizes **E** e **F** representam os erros de modelagem de **X** e **Y**, respectivamente.

O método de regressão PLS tem como resultado uma equação linear que descreve a curva de calibração. A partir desta equação da calibração é feita uma correlação com o método de referência com base no coeficiente de correlação ($\mathbf{R^2}$) e dos erros de calibração (RMSEC) e de validação (RMSEV). A Equação 4 é responsável pelo cálculo destes erros, onde $\mathbf{y_i}$ e $\mathbf{\hat{y}_i}$ são, respectivamente, os valores de referência e estimado para a i-ésima amostra e **n** o número total de amostras.

$$RMSEC,\ RMSEV\ ou\ RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \quad (4)$$

Analogamente, quando é utilizado na validação o processo de validação cruzada (*crossvalidation*), obtém-se o RMSECV com base na Equação 4, onde **i** representa cada amostra retirada do conjunto de calibração ao longo da validação.

2.5 Regressão de mínimos quadrados parciais por intervalos (iPLS)

Este método constitui-se numa forma simples e determinística de realizar seleção de variáveis em modelos de regressão multivariada que empregam, por exemplo, dados espectrais.

O método iPLS consiste na aplicação do PLS em janelas eqüidistantes do espectro total, com o objetivo de fazer com que o processo avalie regiões importantes do espectro removendo as interferências de outras, aumentando assim a sensibilidade do método a ruídos presentes no espectro. Esta capacidade faz com que o iPLS seja adequado para reconhecer os locais do espectro onde estão as informações relevantes para a construção do modelo de calibração (LEARDI *et al*, 2004).

Ainda citando Leardi *et al* (2004), o iPLS divide o espectro em *n* intervalos, criando *n+1* modelos, um para cada intervalo, respectivamente, e um global, com todo o espectro. O desempenho de todos estes modelos é comparado, sendo esta comparação geralmente baseada no parâmetro de validação da média das raízes quadradas dos erros da validação cruzada (RMSECV).

2.6 Utilização de algoritmo genético para escolha de intervalos para aplicação do iPLS (GA-iPLS)

Como nem sempre as informações do espectro estão contidas em um único intervalo, pode ser necessário combinar diferentes intervalos para obter um modelo de calibração mais robusto.

Para auxiliar na resolução do problema da escolha dos intervalos mais relevantes do espectro que serão retidos, os dados espectrais podem ser submetidos ao processo de otimização através de Algoritmos Genéticos com o objetivo de refinar o modelo. A teoria necessária para um melhor entendimento dos Algoritmos Genéticos será descrita com maiores detalhes no item 2.8.

2.7 Teoria da evolução

Durante o século XIX Charles Darwin (1809-1882) lançou sua teoria sobre a evolução da vida, afirmando que o homem, assim como os demais animais, também seria resultado de uma evolução. Segundo Darwin citado por (YEPES, 2004), os seres vivos sofrem modificações ao longo de anos que os tornam cada vez mais adaptados ao ambiente em que estão inseridos. A determinação das características dos seres vivos que estarão presentes ou não em próximas gerações, seria dada por uma seleção natural. Essa seleção tem como princípio: Os seres vivos melhor adaptados a um determinado ambiente tendem a se reproduzir mais do que os restantes.

Um exemplo ilustrado por Darwin é o do aumento do número de girafas com pescoço longo. Inicialmente existia uma população de girafas com tamanhos variados de pescoços. Com o passar do tempo e com uma escassez de alimentos no solo, as girafas com pescoços mais compridos apresentaram uma maior facilidade para alimentar-se, passando a se reproduzirem com mais freqüência que as de pescoços curtos. Após um tempo, a existência de girafas, em sua grande maioria com pescoços longos, foi conseqüência de uma adaptação sofrida por esses animais de acordo com as necessidades impostas pelo ambiente.

Darwin, na época, não conseguiu explicar claramente a forma de transmissão dessas características entre seres vivos. Isso seria mais bem entendido posteriormente, com as descobertas das Leis de Mendel e Mutações.

2.8 Algoritmos genéticos (GA)

Durante os anos 60 e 70 John Holland citado por Yepes (2004)[1], com a colaboração de alguns colegas, desenvolveu a teoria denominada Algoritmos Genéticos. A idéia era trazer para os sistemas de computação os mecanismos de adaptação natural, utilizando os princípios da natureza como meio para solução de problemas. Os Algoritmos Genéticos são, na verdade, um dos paradigmas de computação evolutiva. Podendo ser citado ainda programação genética, programação evolutiva, estratégias evolutivas e sistemas classificadores.

O uso de Algoritmos Genéticos para o tratamento de otimizações vem apresentando resultados pertinentes para aplicações em que métodos tradicionais de programação não são capazes de solucionar em um tempo computacionalmente viável.

Um Algoritmo Genético funciona, então, da seguinte forma: Tem-se um problema cuja solução possui um padrão conhecido, ou seja, se conhece o formato da solução para este problema.

Assim, qualquer solução que respeite essas características conhecidas será considerada uma solução válida para o problema em questão. Essa solução pode ser avaliada quanto ao seu erro ou quanto ao seu grau de acerto para o caso em estudo, e de acordo com esse resultado, pode ser considerada uma solução bem adaptada ou pouco adaptada ao problema.

O princípio dos Algoritmos Genéticos é criar um conjunto inicial com várias possíveis soluções para um problema. Esse conjunto é chamado de população inicial e cada solução pertencente a esta população é chamada de indivíduo. Esses indivíduos devem ser codificados de forma a constituir seu material genético. Neste caso considera-se que o material genético é um cromossomo, e este, é um indivíduo propriamente dito.

Essa codificação dos indivíduos visa possibilitar a aplicação de operadores genéticos e do conceito de seleção natural sobre as soluções existentes, podendo assim combinar seus

[1] Projeto independente na área da Computação Evolutiva. Objetiva fornecer informações sobre conceitos, aplicações e técnicas de implementação dos Algoritmos Genéticos. (www.geocities.com/igoryepes/)

materiais genéticos buscando ao longo de várias gerações indivíduos cada vez mais adaptados, ou seja, soluções com um menor grau de erro para o problema tratado.

Percebe-se então que os Algoritmos Genéticos trabalham com uma população de soluções para um problema, submetendo esta população a um processo de evolução inspirada nas teorias evolucionárias de Darwin e fazendo com que essas soluções adquiram ao longo do tempo uma carga genética que lhes atribuam características capazes de representar uma solução “ótima” para um problema. Após várias gerações, quando a população estiver formada por indivíduos muito semelhantes, diz-se que a população convergiu para uma solução e que, provavelmente, essa solução é “ótima”.

A evolução de um Algoritmo Genético ocorre de acordo com uma série de parâmetros que devem ser informados inicialmente para o algoritmo. Embora existam na literatura sugestões para tal, alguns desses parâmetros não possuem regras para sua escolha, sendo assim definidos empiricamente e sofrendo ajustes de acordo com a qualidade dos resultados obtidos. Essas escolhas implicarão diretamente na forma com que o Algoritmo Genético irá se portar diante do problema escolhido. Por exemplo: Um parâmetro relacionado às taxas de cruzamentos entre os indivíduos de uma população pode acarretar em uma convergência prematura deste conjunto de soluções ou então, em um caso totalmente oposto, fazer com que esta população nunca convirja.

A seguir seguem algumas notações de um Algoritmo Genético para uma inicial familiarização:

- **Indivíduo ou Cromossomo:** Qualquer possível solução. É o material genético obtido a partir da codificação de uma solução;
- **População:** Conjunto de indivíduos;
- **População Inicial:** É a população da primeira geração;
- **Gene:** É cada informação contida em um indivíduo (cromossomo);
- **Alelo:** São os valores possíveis que um gene pode assumir;
- ***Fitness*:** Grau de adaptação de um indivíduo.

2.8.1 Desenvolvimento do algoritmo genético

Antes de começar a desenvolver um sistema dessa natureza deve-se, primeiramente, conhecer claramente o tipo de problema que se trata e analisar o formato de uma possível solução para este. É importante ter certeza quanto às características de uma solução e também quanto as possíveis restrições que devem ser levadas em consideração. O passo seguinte é: Como se pode codificar essa solução (indivíduo)? A codificação de uma solução é a representação da mesma através de um material genético, ou seja, de uma seqüência de genes. Uma alternativa muito utilizada é a representação de uma solução através de um sistema binário. Por exemplo: uma seqüência como “100101” é um indivíduo com 6 (seis) genes, possuindo 2 (dois) alelos, “0” (zero) e “1” (um). Durante esse processo de codificação devem-se suprir alguns requisitos:

Deve ser possível avaliar um indivíduo a partir de seu código genético (cromossomo);

Deve-se visar uma representação na qual a aplicação dos operadores genéticos (ainda não descritos) seja facilitada;

A representação deve garantir que todo o domínio de soluções possa ser alcançado.

Uma vez estabelecida a forma de codificação das soluções, pode-se agora introduzir de forma mais clara e específica os conceitos de operadores genéticos e de seleção natural.

2.8.2 Operadores genéticos

Inspirado pela forma com que os seres vivos evoluem, foram desenvolvidos alguns operadores genéticos com a finalidade de manipular e transmitir as cargas genéticas dos indivíduos ao longo de várias gerações. São eles: Operador de Cruzamento e Mutação.

Operador de Cruzamento: Conhecido como *crossover*, é um operador responsável por combinar o material genético de dois indivíduos distintos, gerando dois filhos com características combinadas de seus pais. É capaz de analisar diversos locais no espaço de busca, sendo aplicado de acordo com uma taxa de *crossover*;

Operador de Mutação: Este operador é responsável por garantir a diversidade da população, realizando pequenos ajustes nas soluções encontradas, e é capaz de evitar que o Algoritmo Genético fique estagnado em um mínimo local[2], modificando alguns genes da população de acordo com uma taxa de mutação. A seguir apresenta-se em detalhes cada um dos operadores.

2.8.2.1 Operador de cruzamento

Existem algumas variações quanto à implementação do operador de *crossover*. Para aplicar este operador, realizando uma combinação de dois materiais genéticos, é necessário estabelecer um ponto chamado de ponto de corte. Assim como o nome já indica, este ponto corta o cromossomo em duas seqüências determinando que a primeira parte de um cromossomo será combinada com a segunda parte de outro, e vice-versa.

Após essa combinação os indivíduos gerados apresentam o mesmo tamanho de seus progenitores, sendo constituídos por uma parte de cada um deles. É comum realizar também dois cortes nos cromossomos para aplicar o operador de *crossover*, sendo assim, a troca do material genético entre os dois indivíduos se dará entre esses dois pontos.

Essas duas implementações citadas caracterizam respectivamente o *crossover* com um ponto de corte e o *crossover* com dois pontos de corte. Uma terceira alternativa para este operador é conhecida como *crossover* com máscara. Neste caso é gerada uma máscara do mesmo tamanho de um cromossomo. Analisa-se a máscara, os genes atribuídos ao indivíduo gerado serão obtidos a partir de uma regra. Quando o valor da máscara for verdadeiro (um), é

[2] Melhor solução dentre as conhecidas, não sendo a melhor solução possível para o problema.

utilizado o gene de um pai A, quando o valor da máscara for falso (zero) utiliza-se o gene de um pai B. A seguir, as Figuras 2, 3 e 4 ilustram cada variação do operador de *crossover*.

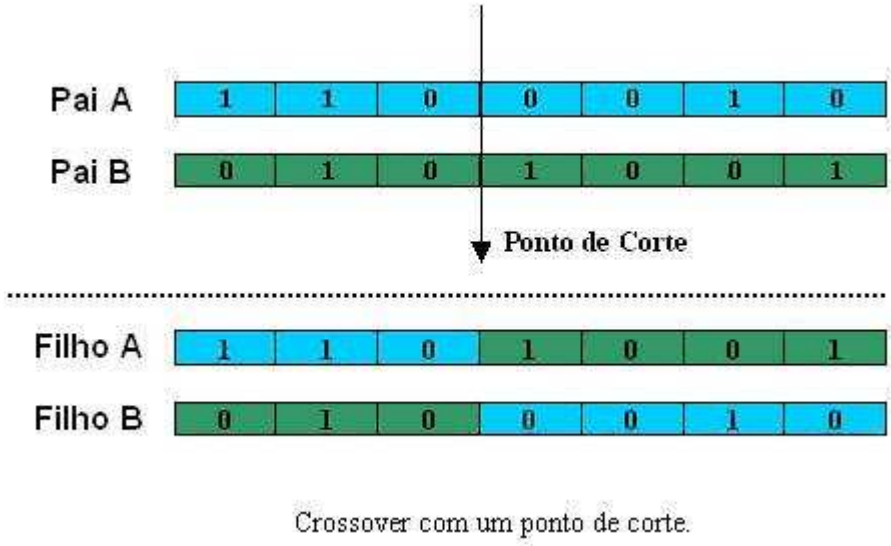


Figura 2 - Representação do *crossover* em um ponto de corte.

Fonte: SABIN e CARVALHO, 2005

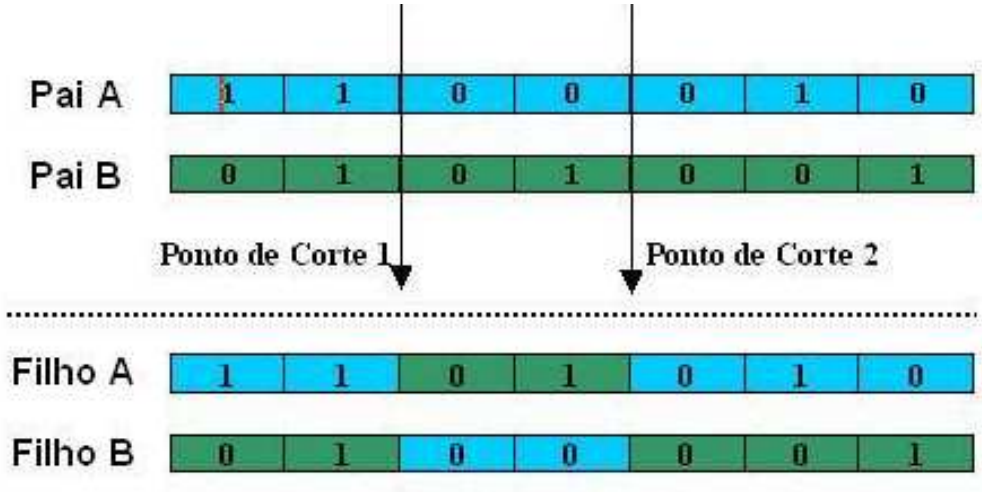


Figura 3 - Representação do *crossover* com dois pontos de corte.

Fonte: SABIN e CARVALHO, 2005

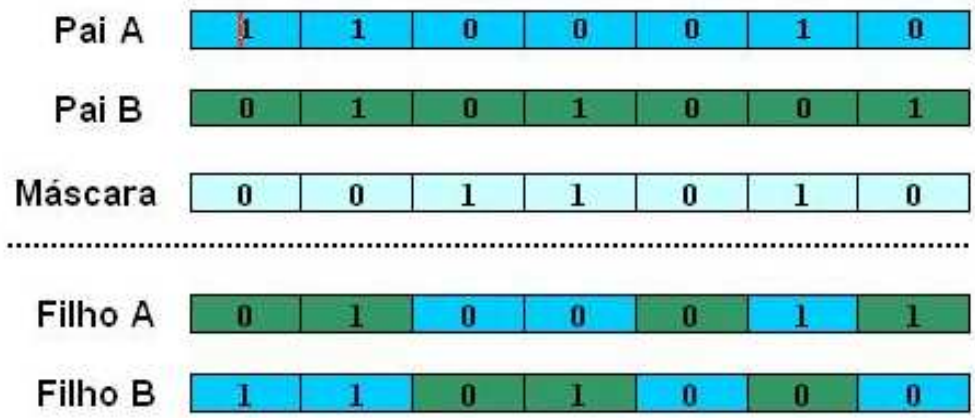


Figura 4 - Representação do *crossover* com máscara.

Fonte: SABIN e CARVALHO, 2005.

2.8.2.2 Operador de mutação

A implementação do operador de mutação é bem simples. Durante uma geração alguns genes são selecionados aleatoriamente para sofrer mutação. Quando se utiliza uma representação binária basta inverter o gene selecionado. Por exemplo: Seleciona-se um gene "n", se este possui valor 1 (um) então será trocado para 0 (zero). Em casos de representação diferente de binária e existindo mais de dois valores de alelos deve-se estabelecer uma regra para a substituição do gene encontrado.Veja ilustração na Figura 5.

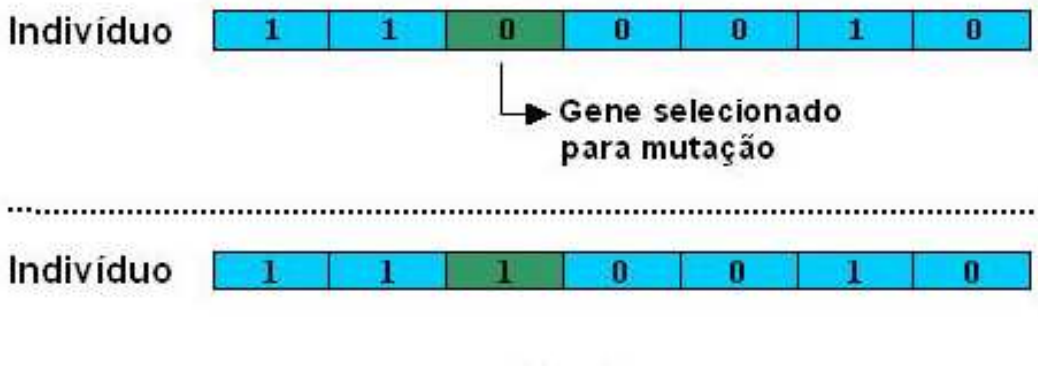


Figura 5 - Representação da mutação.

Fonte: SABIN e CARVALHO, 2005

2.8.3 Métodos de seleção

Para que seja possível aplicar o operador de *crossover* é necessário selecionar um conjunto de indivíduos para combinar seus materiais genéticos. Esse processo de seleção é conhecido como seleção natural, já citado no início desta seção como a explicação dada por Charles Darwin (YEPES, 2004) para a evolução e adaptação dos seres vivos ao longo do tempo. Para que haja uma convergência adequada da população, procura-se reproduzir da melhor maneira este processo de seleção executado pela natureza. Para que isso seja possível, é preciso levar em consideração o grau de adaptação, *fitness*, de cada indivíduo como referência para determinar quais serão submetidos ao cruzamento. Existem várias técnicas utilizadas para simular a seleção natural, dentre elas, podem ser citadas: Seleção proporcional ao desempenho, seleção por torneio, seleção por posição e seleção por truncatura. O número de indivíduos que serão selecionados depende de uma taxa de cruzamento especificada no início do sistema. Veja a seguir a descrição completa de cada uma destas técnicas.

2.8.3.1 Seleção proporcional ao desempenho

Supondo que todos os indivíduos da população já foram avaliados e a cada um destes já foi atribuído um grau de adaptação com relação ao problema, constrói-se um intervalo e se atribui a cada indivíduo uma parte deste intervalo. O detalhe é que, a porção do intervalo que será relacionada a um determinado indivíduo será sempre proporcional ao seu desempenho com relação à população atual. Assim, indivíduos muito bem adaptados receberão uma parte maior do intervalo, tendo desta forma uma maior chance de serem selecionados, enquanto que indivíduos com baixa adaptação receberão uma parte menos significativa do intervalo total, apresentando uma chance menor de serem selecionados.

Esse método é conhecido também como método da roleta, pois pode ser imaginado como uma roleta posta a girar, onde cada indivíduo é representado por uma fatia de tamanho obtido de acordo com o seu desempenho.

Isso tudo serve para que, indivíduos com alto desempenho tenham maior probabilidade de serem selecionados para combinação genética, porém isso não quer dizer necessariamente que indivíduos com baixo desempenho não serão selecionados. Uma vez formado este grupo de indivíduos, deve-se estabelecer um critério para determinar os pares de indivíduos que serão submetidos ao cruzamento.

Uma forma muito utilizada é ordenar estes indivíduos em forma crescente de adaptação e cruzar sempre o indivíduo de posição *n* com o indivíduo de posição *n+1*. Este critério pode variar muito e também está presente nos próximos métodos de seleção que serão analisados. Na Figura 6 ilustra-se o método da roleta.

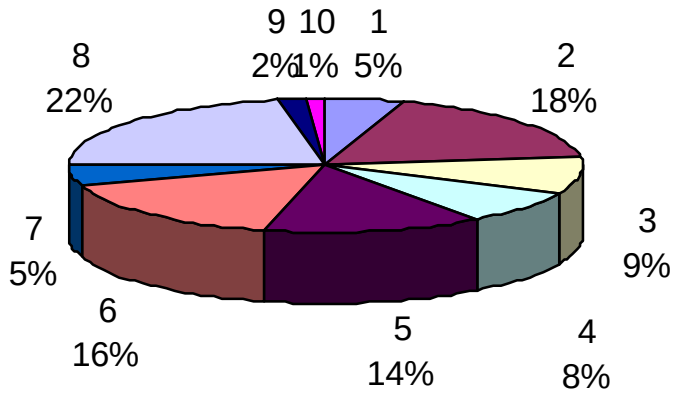


Figura 6 - Ilustração da aplicação do método da roleta.

Fonte: SABIN e CARVALHO, 2005

2.8.3.2 Seleção por torneio

Novamente, neste ponto, é necessário estar com a população totalmente avaliada. A seleção por torneio é baseada em uma competição realizada entre os indivíduos de uma população. Além da taxa de cruzamento, para este método de seleção é necessário mais um parâmetro, um número que representa o tamanho do torneio. O tamanho do torneio é um valor

que determina quantos indivíduos participarão do torneio enquanto disputam uma vaga. Esse valor pode ser fixo ou ser uma taxa com relação ao tamanho da população. Assim, serão realizados tantos torneios quanto o número de indivíduos a serem selecionados.

Esses indivíduos que participam de um torneio são selecionados aleatoriamente, independente de seu grau de adaptação. Todos os indivíduos têm a mesma chance de participarem do torneio, porém será dito campeão do mesmo sempre o melhor indivíduo. Por exemplo: Tem-se uma população com 100 indivíduos, uma taxa de cruzamento de 80% e um tamanho de torneio de 10% da população. Logo serão realizados 80 torneios de tamanho igual a 10, sendo que em cada torneio apenas um indivíduo é consagrado campeão. Os campeões dos torneios são os indivíduos selecionados para o cruzamento, que deverão ser organizados em pares de forma semelhante ao método anterior. O método de seleção por torneio está ilustrado na Figura 7.

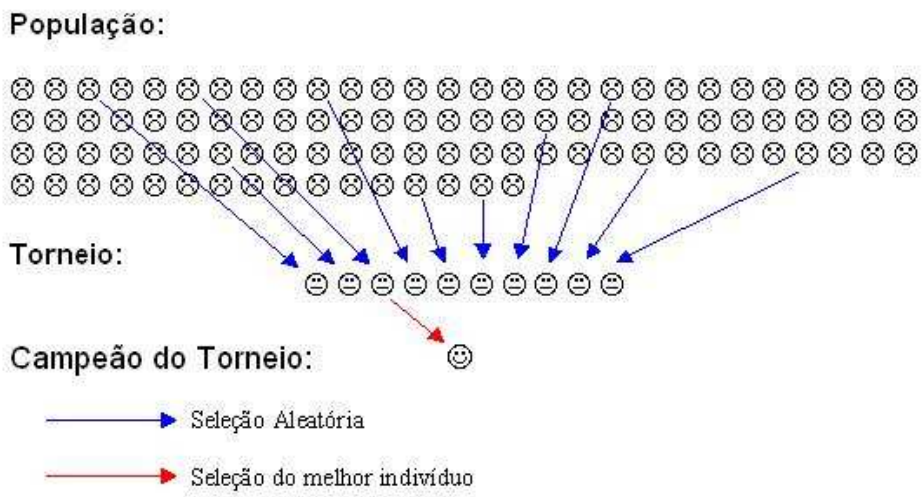


Figura 7 - Representação do método de torneio.

Fonte: SABIN e CARVALHO, 2005

2.8.3.3 Seleção por posição

Novamente este método é bem semelhante ao método da seleção proporcional ao desempenho (roleta) em que um indivíduo mais adaptado tem maior chance de ser

selecionado, mas com a diferença de que a probabilidade deste evento ocorrer está ligada, não ao seu grau de aptidão ao problema, e sim com a colocação deste indivíduo de acordo com o seu *fitness*. Assim, se um indivíduo é o terceiro melhor da população, então ele terá a terceira maior probabilidade de ser escolhido para o cruzamento entre os indivíduos da população.

Nesta implementação, ordenam-se os indivíduos em forma crescente de *fitness*, fazendo com que o melhor indivíduo ocupe a posição 1 e o pior, a posição ***n***.

A probabilidade de cada indivíduo ser selecionado pode ser dada de duas formas:

- distribuição linear: probabilidade (indivíduo i) = a x i + b, onde a>0

- distribuição exponencial: probabilidade (indivíduo i) = a x exp(b x i + c)

Em qualquer uma das formas escolhidas para a atribuição de probabilidades para os indivíduos, a soma destas deve ser igual a 1 (um).

Este método de seleção evita que um pequeno grupo de soluções com um desempenho muito melhor que o do restante da população domine as populações subseqüentes, evitando assim a convergência para um mínimo local.

No exemplo da Figura 8 é demonstrada a probabilidade de seleção de cada indivíduo de uma população de tamanho igual a dez.

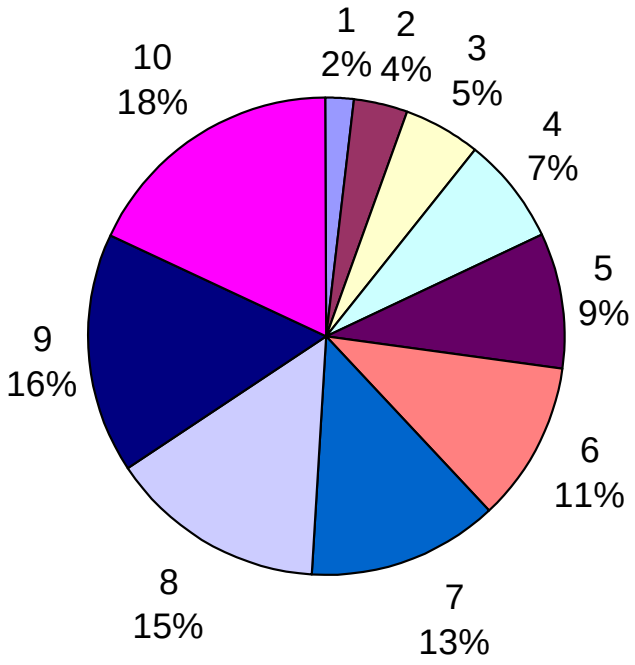


Figura 8 - Ilustração da aplicação do método da seleção por posição.

Fonte: SABIN e CARVALHO, 2005

2.8.3.4 Seleção por truncatura

Este método ordena os indivíduos de acordo com o seu desempenho, definindo um grupo contendo os “n” indivíduos mais aptos. Participarão da seleção apenas os indivíduos pertencentes a este grupo, tendo todos a mesma probabilidade de escolha. Vale ressaltar que quanto menor o tamanho (n) do grupo, maior será a pressão evolutiva.

A Figura 9 apresenta um exemplo em que o tamanho da população é igual a quinze e o número de indivíduos que pertencerão ao grupo dos selecionáveis é igual a cinco.

Indivíduos não selecionáveis

Indivíduos selecionáveis (melhores)

Figura 9 - Método de seleção por truncatura.

Fonte: SABIN e CARVALHO, 2005

Os indivíduos que poderão ser selecionados são apenas os pertencentes ao grupo de selecionáveis: 8, 5, 11, 14 e 6. Logo, a probabilidade é a mesma para todos, ou melhor, 20% para cada.

Após a seleção e a aplicação dos operadores genéticos é necessário avaliar os novos indivíduos produzidos. A partir dos resultados obtidos e juntamente com os já existentes sobre a população atual, deve-se decidir quais indivíduos que permanecerão para uma próxima geração e quais aqueles que serão substituídos por novos indivíduos que apresentaram um bom grau de adaptação. Considerando sempre que o tamanho da população deve permanecer inalterado, têm-se as seguintes opções de atualização de uma população:

2.8.4 Métodos de atualização da população

A população manipulada pelo GA é de tamanho fixo, ou seja, possui um número pré-estabelecido de indivíduos, não podendo ter este número aumentado ou diminuído. Tendo em vista esta característica, que deve ser mantida, depara-se com um novo problema: a substituição de indivíduos na população.

Quando se geram indivíduos através dos operadores genéticos, esses indivíduos devem ser colocados na população, sendo assim, para que o tamanho da população seja mantido, é necessário que outros sejam retirados dela.

Existem vários métodos para realizar a substituição de indivíduos na população depois de um cruzamento, os métodos mais utilizados são (YEPES, 2004):

Substituição imediata: os novos indivíduos gerados substituem os seus geradores;

Substituição por fator cheio: cada novo indivíduo substitui o indivíduo mais parecido com ele na população;

Substituição por inserção: são selecionados “n” indivíduos para serem eliminados nessa população (geralmente os piores), então estes indivíduos são substituídos pelos novos;

Substituição por inclusão: os novos indivíduos são incluídos na população, só então são selecionados os *t* melhores indivíduos que serão conservados nesta população, onde *t* é o tamanho (número de indivíduos) da população;

Cabe salientar que relativo à mutação, não se faz necessário reinserir indivíduos, pois esta operação não gera um novo indivíduo, apenas altera um existente. Os métodos citados acima, com algumas ressalvas, podem ser utilizados também na implementação do elitismo (que será explicado a seguir) para realizar a atualização de boas cargas genéticas.

2.8.5 Elitismo

O objetivo de um Algoritmo Genético é evoluir um conjunto de soluções a fim de obter uma solução muito otimizada para um problema. Mesmo utilizando-se métodos de seleção e aplicando-se operadores genéticos, a convergência pode não ocorrer satisfatoriamente sem a introdução de um novo conceito: Elitismo (YEPES, 2004).

O elitismo, técnica introduzida por Kenneth De Jong em 1975, garante que ao longo de várias gerações, bons indivíduos, ou melhor, boas cargas genéticas, não sejam perdidas ou deixadas para traz. Isso acontece pelo seguinte: durante o processo de seleção para o cruzamento, faz-se com que os melhores indivíduos se reproduzam mais do que os restantes,

de acordo com a teoria da seleção natural. O problema é que não se tem a garantia de que um indivíduo muito bom, ao ser cruzado com outro, terá como resultado filhos com características melhores ou no mínimo semelhantes ao pai. Neste caso, produzem-se indivíduos piores do que os existentes nas gerações anteriores.

Este problema poderia ser facilmente resolvido com os critérios de atualização de populações citados anteriormente, porém, para o caso do operador de mutação isso não funcionaria. Imagine que o operador de mutação selecione aleatoriamente um gene de um dos melhores indivíduos da população, e este gene ao ser alterado não gera um indivíduo satisfatório. Neste caso deve-se impedir essa mutação.

A solução para este problema é criar um grupo chamado elite, onde estariam salvas as cargas genéticas dos melhores indivíduos da população. Esses indivíduos participariam normalmente do processo de seleção natural e de aplicação de operadores genéticos, a diferença é que: Ao término de uma geração, se for constado que o grupo de elite foi alterado e parte dele não existe mais, suas cargas genéticas são novamente introduzidas na população. Durante este processo de introdução podem ser usados os métodos empregados para atualização de população (citado anteriormente), com a única ressalva de que o método da substituição imediata não pode ser utilizado, pois os indivíduos da elite não podem substituir os seus genitores, sendo que neste ponto não se tem mais esta informação.

A única possibilidade para que os indivíduos da elite não sejam recolocados na população é o caso de toda a população ter evoluído para um estágio no mínimo superior ao da elite, neste caso eles não são mais considerados tão essenciais e podem ser descartados. No entanto é muito improvável que tal situação venha a ocorrer. O número de indivíduos que pertencerão a esta elite pode ser obtido a partir de uma taxa de elite, que é uma porcentagem da população.

É necessário neste momento realizar uma análise nos parâmetros de entrada utilizados a fim de esclarecer o funcionamento do GA.

Tamanho do cromossomo: É o número de genes presente em um indivíduo.

Tamanho da população: É o número de indivíduos existentes em cada geração.

Número de gerações: É o número de ciclos para evoluir a população.

Taxa de cruzamento: Porcentagem da população que será cruzada a cada geração.

Taxa de mutação: Porcentagem de genes que sofrerão mutação a cada geração.

Taxa de elitismo: Porcentagem de indivíduos que serão preservados a cada geração.

Método de Seleção: Indica qual dos métodos de seleção deve ser utilizado.

2.8.6 Análise da eficiência de algoritmos genéticos

Agora que foram definidos alguns parâmetros de entrada, pode-se realizar uma análise de como a população se portará durante o processo de evolução. Este estudo será feito de forma genérica, não sendo considerado o problema em específico para o qual esta monografia se propõe, este caso será considerado mais adiante.

A idéia neste ponto é, principalmente, relacionar alguns parâmetros de entrada com a possibilidade de sucesso ou não da convergência da população. Um problema muito comum é ocorrer uma convergência prematura da população pelo fato de existir o que se chama de grande pressão evolutiva. Quanto maior for a pressão evolutiva, mais rápido ocorrerá uma convergência da população. Isso ocorre da seguinte forma: imagine que em uma dada geração exista um indivíduo ou um pequeno grupo de indivíduos que se destaque relativamente em relação aos demais, porém, este ainda não é uma solução aceitável para o problema. Caso a pressão evolutiva seja muito elevada, a evolução do restante da população se dará em função deste indivíduo, ou seja, a população convergirá para uma solução que não é a mais adequada.

Este problema pode ocorrer, por exemplo, durante um processo de seleção através do método de torneio. Considerando um tamanho de torneio muito grande, a probabilidade é que os melhores indivíduos estejam sempre participando do torneio e, sendo assim, estes serão

sempre os campeões. Não dando chance para que indivíduos intermediários possam participar do processo de cruzamento. O que acontece então é que ocorrerá uma combinação de material genético apenas entre os “melhores” indivíduos, fazendo com que a população fique estagnada neste ponto, quando na verdade este ainda não é um máximo global da função. O operador de mutação, que garante a diversidade da população, pode ser capaz de evitar esse problema, no entanto como geralmente a taxa de mutação é bem menor do que a taxa de cruzamento, essa correção pode demorar muito a acontecer, ou ainda, nunca ocorrer.

2.8.7 Passos de um algoritmo genético

Apresenta-se a seguir, de forma simplificada, os passos de um Algoritmo Genético, ilustrados na Figura 10.

Considerando os parâmetros fornecidos, cria-se a população inicial gerando aleatoriamente os genes dos indivíduos;

1. Realiza-se a avaliação do *fitness* de cada indivíduo;
2. Submete-se a população ao processo de seleção natural. Neste passo está inclusa tanto a própria seleção dos indivíduos, quando a aplicação dos operadores de cruzamento e de mutação;
3. Realiza-se uma avaliação do *fitness* dos indivíduos gerados pelo passo anterior;
4. Atualiza-se a presente população com os indivíduos gerados, mantendo o tamanho da população inalterado;
5. Caso seja satisfeito um critério de parada retorne o melhor indivíduo, caso contrário volte ao passo ”3”. O critério de parada pode ser tanto um número máximo de gerações quanto um erro mínimo desejado para uma solução.

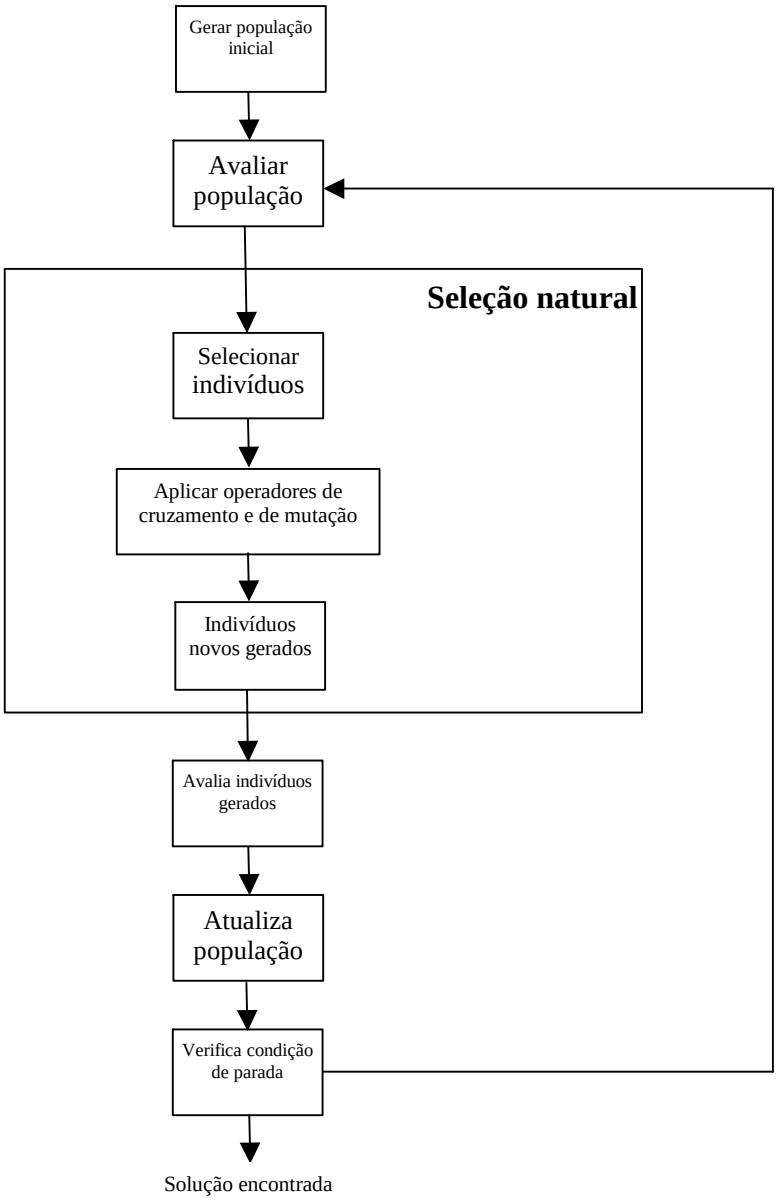


Figura 10 - Representação dos passos de um GA.

Fonte : SABIN e CARVALHO, 2005

3 METODOLOGIA

Para viabilizar a pesquisa alguns passos foram seguidos e serão descritos no decorrer deste capítulo.

Este trabalho começou com uma consistente pesquisa bibliográfica como tipo de coleta de informações sobre aplicações dos métodos multivariados PLS e iPLS na construção de métodos de calibração multivariada, o estudo dos espectros e também sobre técnicas de otimização combinatória, com a finalidade de adquirir e aprofundar os conhecimentos envolvidos nesta pesquisa.

Devido a problemas instrumentais ou relativos a natureza da amostra, alguns ajustes precisam ser feitos para corrigir o espectro e possibilitar modelos mais robustos, sendo estes ajustes chamados de pré-processamento. Existem vários tipos pré-processamentos para ajuste dos dados espectrais, mas nesta dissertação serão apresentados apenas os métodos de pré-processamento utilizados nos conjuntos de dados estudados quando estes forem abordados.

A matriz de dados obtida através de espectroscopia foi submetida ao GA-iPLS para buscar um modelo que represente melhor o problema, ou seja, alcançando um modelo mais preditivo que o iPLS. O resultado esperado é um modelo que represente da forma mais fiel possível a análise em questão, requerendo um tempo de processamento viável para aplicações industriais, visto que a análise de todas as soluções possíveis é impraticável por se tratar de um problema NP-completo[3].

Por fim, foi implementada uma fase de testes, onde o resultado de tais ferramentas foi avaliado para verificar a validade dos métodos aqui desenvolvidos, comparando-os com outros existentes.

[3] Categoria de problemas que se caracteriza por ter um espaço de busca tão amplo que é considerado impossível calcular todas as soluções possíveis

3.1 Adequação do algoritmo genético para selecionar variáveis de espectros no infravermelho.

Para que o GA consiga otimizar modelos de regressão multivariados de espectros no infravermelho é necessário entender como eles são criados e como os dados espectrais podem ser trabalhados por este algoritmo a fim de alcançar uma solução otimizada.

Com base no funcionamento do iPLS, que divide o espectro em subintervalos e encontra como solução o modelo de regressão multivariado que apresentar o menor erro, criado utilizando as variáveis de um dos intervalos, o GA deve ser capaz de identificar em quantos subintervalos o espectro deve ser dividido e indicar quais destes serão utilizados na elaboração de uma solução.

Para isso os cromossomos contêm tantos genes quanto for o número de subintervalos em que o espectro foi dividido. Desta forma, é possível informar quais variáveis de um espectro são representadas por cada gene. Para que sejam conhecidos quais subintervalos serão considerados na criação do modelo de regressão multivariado indicado por um determinado cromossomo, cada gene que o compõe apresenta os alelos **1** (indicando que o correspondente intervalo foi considerado na criação do modelo) e **0** (indicando que o correspondente intervalo foi desconsiderado na criação do modelo).

Nesta implementação, o tamanho da população indica quantos modelos de regressão multivariados serão avaliados e submetidos à evolução em cada geração do GA. As taxas de cruzamento, de mutação e de elitismo são parâmetros do algoritmo genético e não necessitam de adaptações para esta aplicação.

Para que seja possível aplicar o GA para selecionar variáveis espectrais, os conceitos a seguir apresentam algumas alterações:

- **Indivíduo ou Cromossomo:** possível solução do problema, ou seja, uma combinação de intervalos do espectro que é utilizada para criar um modelo de regressão multivariado;

- **População:** Conjunto de soluções submetidas ao processo de evolução através do GA;
- **População Inicial:** É o conjunto de soluções na primeira geração;
- **Gene:** Representa um determinado intervalo do espectro, indicando se este é ou não utilizado na criação do modelo de regressão multivariado;
- ***Fitness*:** Grau de adaptação de um indivíduo, ou seja, o erro de validação do respectivo modelo;
- **Seleção Natural:** Garante maior probabilidade se seleção às soluções que apresentarem menor erro de validação.

A implementação do GA utilizado neste trabalho foi realizada no MATLAB Version 6.5.0.180913a (R13). Para possibilitar o trabalho com o método PLS e iPLS, que foi empregado como função de avaliação do GA e como método comparativo dos resultados obtidos, também foi utilizado o pacote The iToolbox Version 1 – July (NORGAARD *et al*, 2000).

Para promover um melhor entendimento, a implementação foi dividida em duas partes: a implementação do próprio GA e a formatação da entrada/saída.

3.2 Implementação do GA-iPLS

O método de regressão multivariada PLS vem sendo utilizado para a formulação de modelos que consigam prever uma propriedade específica. O problema que pode ocorrer quando se utiliza este método é que os modelos obtidos por ele podem considerar informações irrelevantes para estimar uma determinada propriedade.

Buscando uma solução para este problema, Norgaard *et al* (2000) desenvolveu o método iPLS. Este método busca selecionar as informações mais relevantes do espectro, mas

nem sempre a sua resposta é satisfatória e, em alguns casos, pode ser pior que os resultados obtidos com o PLS aplicado sobre toda a informação instrumental (espectro). Isso se deve ao fato de que a qualidade dos modelos obtidos com o iPLS tem forte relação com a quantidade de intervalos que o espectro é dividido, tendo em vista que se os intervalos forem muito grandes, o modelo pode continuar utilizando dados irrelevantes ao problema. Porém, se os intervalos forem muito pequenos, este método pode estar desprezando dados pertinentes a propriedade que se quer prever.

Outro grande problema é que não se avalia um possível sinergismo entre diferentes intervalos do espectro quando se cria um modelo com a técnica iPLS. Para contornar estes possíveis problemas, pensou-se em utilizar uma meta-heurística baseada no paradigma da programação evolutiva, a fim de obter um modelo em que o tamanho dos intervalos do espectro não influenciassem de maneira tão drástica quanto no iPLS e que também fosse capaz de avaliar um possível sinergismo entre estes intervalos na criação de modelos de regressão multivariada.

Desta forma, implementou-se um algoritmo genético para buscar uma combinação de variáveis do espectro que proporcionasse modelos com melhor habilidade de predição. Com essa finalidade, foram desenvolvidas duas abordagens empregando-se algoritmos genéticos. O GA-iPLS *out* divide o espectro em intervalos, de forma semelhante ao iPLS, buscando combinações destes intervalos para a obtenção de modelos. Já o GA-iPLS *in* busca por variáveis dentro de intervalos indicados, ou seja, são informados quais intervalos o algoritmo irá trabalhar e este faz uma seleção dos comprimentos de onda dentro destes intervalos de forma a melhorar (refinar) a capacidade de predição do método PLS.

Para esclarecer o funcionamento dos algoritmos desenvolvidos, serão descritos todos os procedimentos adotados passo a passo.

3.2.1 Codificação

Esta etapa da implementação tem o objetivo de representar as soluções através de cromossomos para viabilizar a execução do GA. Esta codificação deve ser capaz de representar qualquer possível solução do problema.

A etapa da codificação apresenta diferenças entre o GA-iPLS *out* e o GA-iPLS *in*.

3.2.1.1 Codificação do GA-iPLS *out*

Nesta implementação, o cromossomo é representado por um vetor binário de tamanho *n*. Um cromossomo é composto por *n* genes, onde *n* é o número de intervalos que o espectro original é dividido, e cada gene tem os alelos 0 (zero) e 1 (um), ou seja, se o gene for igual a 1, o intervalo que ele representa será selecionado para a criação do modelo, se o gene for igual a 0, o intervalo não será selecionado. A codificação é feita de maneira que o primeiro gene representa o primeiro intervalo do espectro, o segundo gene representa o segundo intervalo e assim sucessivamente.

A Figura 11 apresenta um cromossomo com 20 genes, demonstrando em um espectro, quais intervalos foram selecionados.

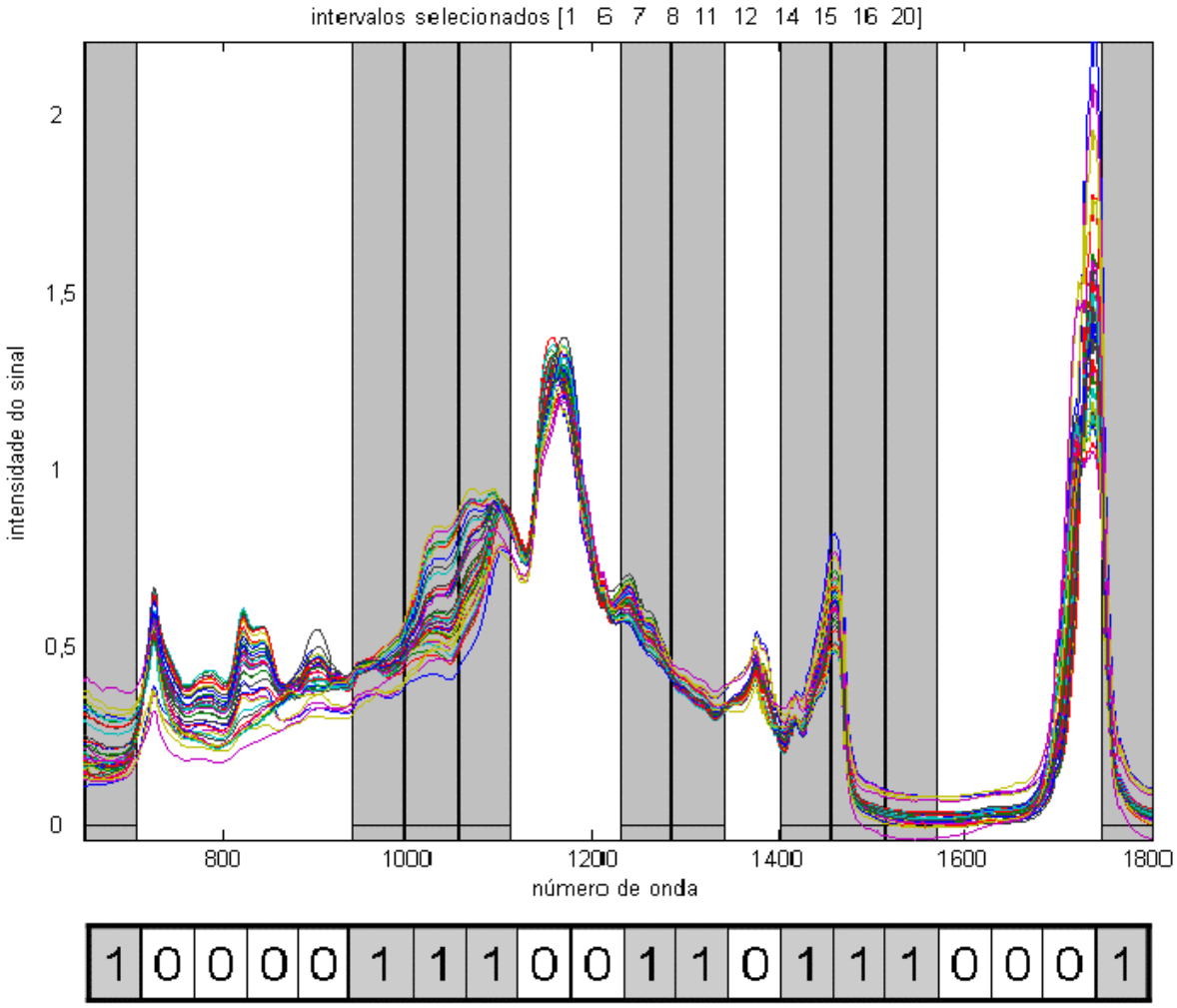


Figura 11 - Exemplo de um cromossomo e representação dos intervalos por ele selecionados.

Fonte: Elaborado pelo autor com base no espectro do infravermelho de amostras de polióis de óleo de soja.

3.2.1.2 Codificação do GA-iPLS *in*

Na codificação do GA-iPLS *in*, os genes do cromossomo não representam intervalos do espectro e sim variáveis do espectro. Estas variáveis são mapeadas dentro de determinados intervalos, que devem ser informados ao algoritmo. Desta forma, através deste método é possível refinar uma solução obtida anteriormente com outro algoritmo, como o iPLS ou o GA-iPLS *out*.

Todavia, semelhantemente ao que foi descrito no item 3.1.1.1, um cromossomo é representado por um vetor binário, onde os genes com valor igual a 1 indicam as variáveis selecionadas, já os genes com valor igual a 0 indicam as variáveis desconsideradas para a obtenção do modelo.

O exemplo do mapeamento de uma solução do GA-iPLS *out*, cujos intervalos selecionados foram o 1, 2 e 6 dos 10 intervalos existentes, servem como dados entrada do GA-iPLS *in* é ilustrado na Figura 12.

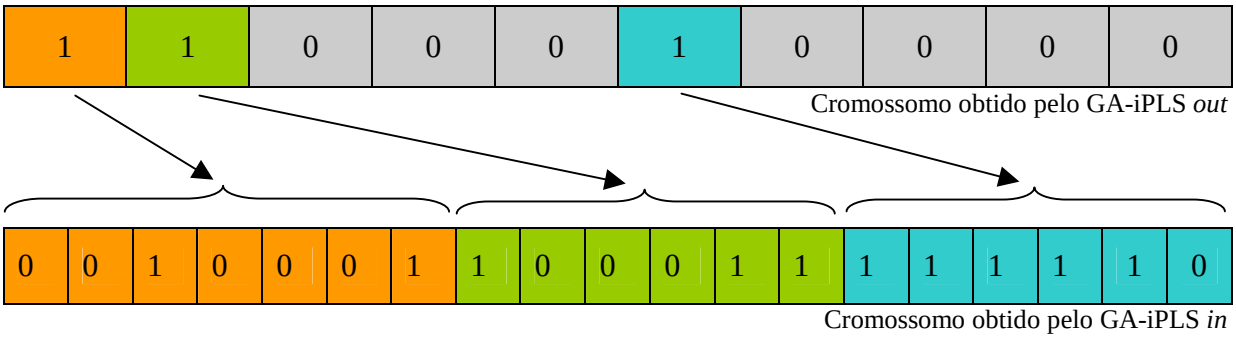


Figura 12 - Exemplo do mapeamento de um cromossomo obtido pelo GA-iPLS out em um cromossomo do GA-iPLS in.

Fonte: elaborado pelo autor.

Na Figura 13 é demonstrado um exemplo de um espectro com as variáveis selecionadas pelo GA-iPLS *in* a partir dos intervalos selecionados pelo GA-iPLS *out*.

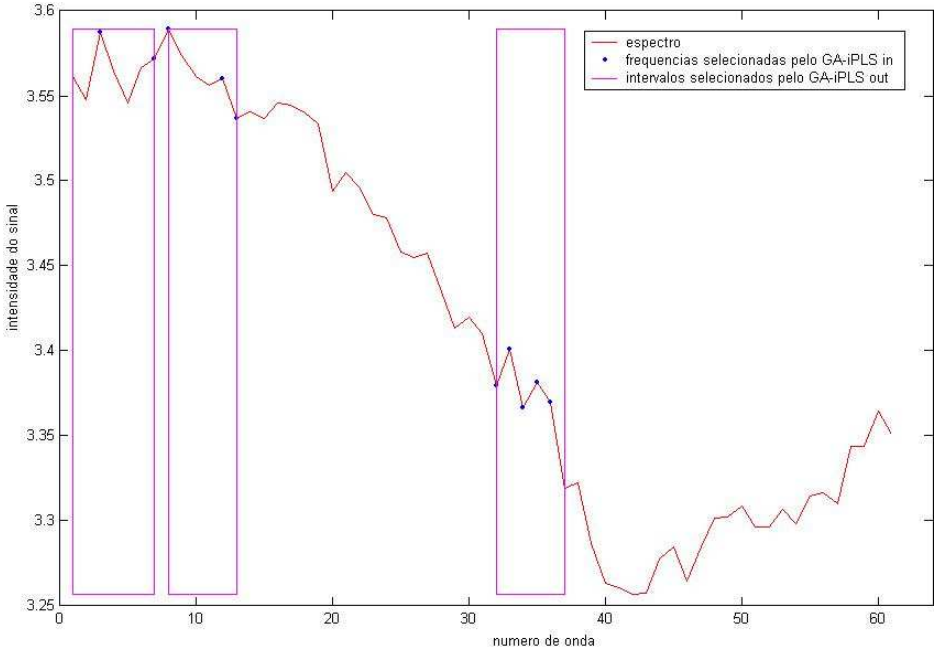


Figura 13 - Freqüências selecionadas pelo GA-iPLS *in* a partir de uma solução do GA-iPLS *out*.

Fonte: elaborado pelo autor.

3.2.2 Avaliação

A avaliação dos indivíduos gerados pelo GA-iPLS *out* e pelo GA-iPLS *in* é muito semelhante e baseada no método iPLS e tem como objetivo atribuir um valor de *fitness* a cada um deles. Este valor será utilizado para verificar o grau de adaptação dos indivíduos.

Os indivíduos são avaliados pelo GA-iPLS *out* criando-se um modelo onde apenas os intervalos indicados por este cromossomo são utilizados. Isso é feito através de um procedimento que verifica quais os intervalos que devem ser considerados na criação do modelo. Um vetor é criado somente com estes intervalos e é informado à função *plsmodel* do pacote *iToolBox*, juntamente com um modelo iPLS com o mesmo número de intervalos do indivíduo avaliado, que cria um modelo utilizando apenas estes intervalos.

A avaliação de indivíduos pelo GA-iPLS *in* é feita de forma análoga, mas os genes devem ser mapeados dentro dos intervalos informados como parâmetro de entrada deste

algoritmo. O procedimento realiza este mapeamento cria um vetor com o número das variáveis selecionadas pelo cromossomo em questão.

Um modelo iPLS com o número de intervalos igual ao número total de variáveis do espectro[4] é informado à função *plsmodel*, juntamente com o vetor das variáveis selecionadas. Então esta função retorna um modelo que considera apenas as variáveis indicadas pelo cromossomo avaliado.

O grau de adaptação de cada indivíduo pode ser calculado considerando-se ou não um conjunto de amostras de validação, ou seja, amostras que não foram levadas em consideração na criação do modelo de regressão.

3.2.2.1 Avaliação sem conjunto de validação

No caso de não existir conjunto de validação, as amostras de calibração são submetidas à função *plsmodel*, realizando uma validação cruzada para a criação de um modelo de regressão. Este modelo possui um vetor de erros (RMSECV), onde cada valor deste vetor é referente à utilização de um número diferente de variáveis latentes. Desta forma, usa-se como *fitness* o menor RMSECV encontrado no vetor de erros do modelo, obtendo-se também o respectivo número de variáveis latentes.

3.2.2.2 Avaliação com conjunto de validação

Foram propostas três diferentes maneiras de elaborar a *fitness* quando existir um conjunto de amostras de validação, onde cada uma foi implementada e testada executando o algoritmo genético e verificando a evolução e os resultados alcançados.

[4] O que significa dizer que cada intervalo deste modelo iPLS é uma única variável do espectro.

A primeira implementação, chamada de *fitness* RMSEV, é feita criando um modelo utilizando as amostras de calibração, onde os erros obtidos por este modelo são referentes aos erros das amostras de validação sobre este modelo. Neste caso, a *fitness* é o menor valor encontrado no vetor de erros do modelo.

A segunda implementação, chamada de *fitness* composta por RMSEC e RMSEV, cria o modelo de calibração da mesma forma da primeira implementação, encontrando o erro das amostras de validação (RMSEV). Porém, também é calculado o erro das amostras de calibração (RMSEC) com o mesmo número de variáveis latentes utilizados no RMSEV. A diferença é que a *fitness* é formulada com base nos erros de calibração e de validação, o que é denominado de cálculo de *fitness* composta.

Esse cálculo é feito escolhendo o menor valor encontrado no vetor de erros de validação e em seguida é escolhido o valor do vetor de erros de calibração que corresponde ao mesmo número de variáveis latentes do erro de validação escolhido.

Conhecendo-se esse dois erros (RMSEC e RMSEV), pode-se efetuar o cálculo da *fitness* composta da seguinte forma:

Calcula-se a média entre os erros de calibração e de validação

Verificar se a diferença entre o RMSEC e o RMSEV é maior que 70%

Se for menor que 70%, a *fitness* é a própria média dos erros

Se for maior que 70%, a *fitness* será a média dos erros somada a porcentagem que exceder os 70% da média dos erros

Por exemplo, se o RMSEC for igual a 100 e o RMSEV for igual a 173, isso significa que o valor de RMSEV é 73% maior que o RMSEC. Sendo assim, o resultado da *fitness* seria de 140,595, que é a média dos erros somada a penalização de 3% desta média.

A última implementação é muito parecida com a segunda, mas o modelo de calibração é criado utilizando o método de validação cruzada e depois, sobre este modelo, é encontrado o

erro das amostras de validação. O RMSECV é o menor valor encontrado no vetor de erros de calibração e o RMSEV é o erro das amostras de validação calculado com o mesmo número de variáveis latentes usado para encontrar o erro de validação cruzada. A seguir é calculada a *fitness* através dos erros RMSECV e RMSEV. Esta implementação é chamada de *fitness* composta por RMSECV e RMSEV.

3.2.3 Seleção natural

Como foi descrita no item 2.7.3, a seleção natural é responsável pelo processo de escolha dos indivíduos que serão submetidos ao processo de cruzamento. Este processo deve ser feito de forma a proporcionar uma maior probabilidade de escolha aos indivíduos mais adaptados.

Neste trabalho foi desenvolvida uma função baseada no método da seleção por posição descrita no item 2.7.3.3, devido à facilidade de implementação.

Para implementar este método, os indivíduos da população devem estar ordenados de acordo com a *fitness* de cada um, onde o primeiro indivíduo deve ser o mais adaptado e o último indivíduo, o menos adaptado. Então é criado um vetor de tamanho igual ao somatório dos índices de posições dos indivíduos da população. As primeiras t posições do vetor são preenchidos com o índice do indivíduo 1, as t-1 posições seguintes são preenchidas com o índice do indivíduo 2 e sucessivamente até que a última posição do vetor é preenchida com o índice do último indivíduo da população.

Desta forma o indivíduo mais adaptado ocupa mais posições do vetor do que um indivíduo menos adaptado. Após a criação do vetor, é sorteada uma posição e o índice nela contido é o referente ao indivíduo selecionado.

3.2.4 Cruzamento

Neste trabalho, o cruzamento foi implementado com um ponto de corte, de forma análoga ao demonstrado no item 2.8.2.1, sendo que este ponto de corte é escolhido de forma aleatória para cada cruzamento realizado.

O cruzamento realizado no GA-iPLS *in* funciona da mesma forma, tendo em vista que os cromossomos deste método apresentam somente as variáveis que se deseja selecionar, tendo seus indivíduos mapeados somente na avaliação.

A Figura 14 demonstra quatro cromossomos, dois pais e dois filhos, com os respectivos intervalos selecionados no espectro.

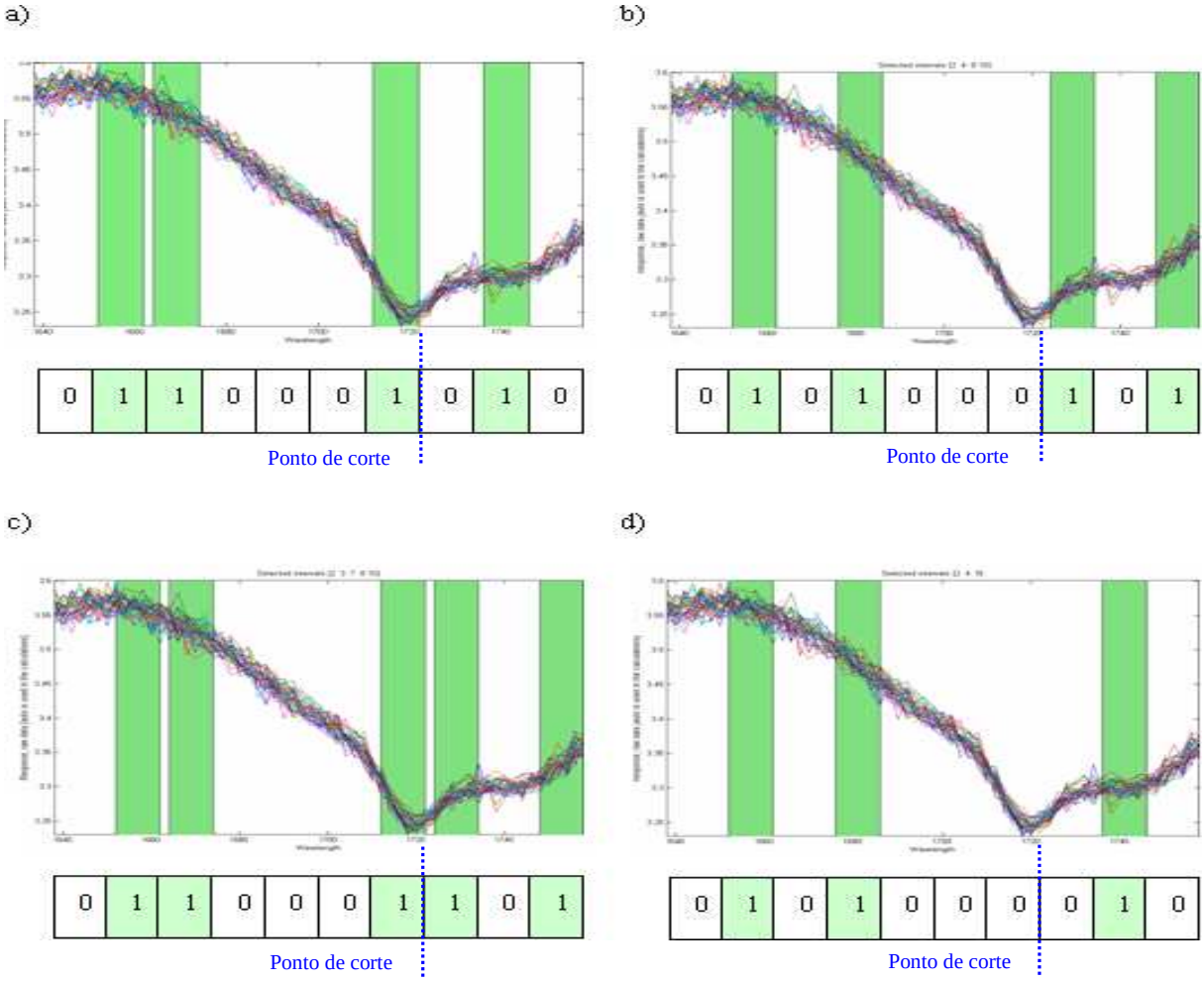


Figura 14 - Cruzamento de *a* e *b* gerando os filhos *c* e *d*, com ponto de corte igual a 7.

Fonte. Elaborado pelo autor

3.2.5 Mutação

A escolha dos indivíduos que sofrem mutação em cada geração é realizada aleatoriamente de acordo com a taxa de mutação, como descrito no item 2.8.2.2. O que difere nesta implementação é que, depois de escolhidos os indivíduos que sofrerão a mutação, para sabermos quantos e quais genes serão mutados utiliza-se uma probabilidade de mutação.

Essa implementação cria um vetor auxiliar de tamanho igual ao do cromossomo, porém este é preenchido de valores aleatórios que vão de 0 a 1. Os genes a serem mutados serão aqueles cuja posição corresponde ao índice no vetor auxiliar onde os valores forem iguais ou menores que a probabilidade de mutação. Se ocorrer de nenhum valor ser igual ou menor a esta probabilidade, a mutação ocorrerá em um único gene, o que estiver na posição referente ao índice do menor valor do vetor auxiliar.

A Figura 15 mostra um cromossomo que foi selecionado para a mutação, o vetor auxiliar e o cromossomo logo após a aplicação deste operador genético, obedecendo a uma probabilidade de mutação de 0,1.

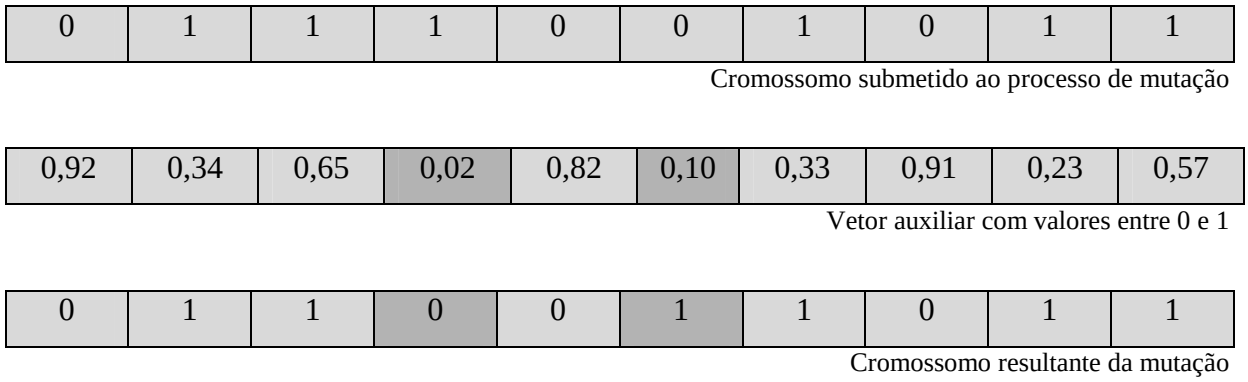


Figura 15 - Exemplo da aplicação da mutação em um cromossomo.

Fonte: elaborado pelo autor

3.2.6 Elitismo e atualização da população

O elitismo foi implementado de acordo com o explicado no item 2.8.5, onde uma porcentagem de indivíduos da população é salva, possibilitando uma reintrodução na população a fim de não serem perdidos. Salvando os melhores indivíduos em cada geração, o algoritmo garante que nunca haverá uma involução.

Tanto na recolocação dos indivíduos da elite na população, quanto na criação de indivíduos através do cruzamento, os indivíduos são acrescentados à população, salvo se esta já possuir um determinado cromossomo da elite. Porém, todas as gerações devem apresentar uma população de tamanho fixo, sendo assim, as piores soluções são retiradas da população para que esta permaneça de tamanho fixo. Este procedimento é denominado atualização por inclusão e foi descrito no item 2.8.4.

3.3 Formatação da entrada/saída

Para assegurar uma maior organização na entrada/saída do GA, foi adotado um padrão no formato de registro. Utilizando este formato, é necessário apenas um arquivo como entrada de dados para o GA, bem como para saída, ou seja, um único arquivo contendo o registro de todos os dados necessários, com exceção do número de gerações que o GA deverá realizar, é passado como entrada e, depois de executar o GA, este mesmo arquivo é atualizado, agora contendo também a saída do GA.

Este registro pode ser dividido em três partes: dados gerais, dados referentes ao GA *out* e dados referentes ao GA *in*. Veja a seguir um maior detalhamento destas partes.

- **Dados gerais:** dados que são utilizados para a criação de modelos de regressão pelo método iPLS, ou seja, são utilizados tanto na execução do GA *out* como quanto na execução do GA *in*, por serem relativos ao espectro estudado. Estes dados são compostos pelas matrizes X e Y, separadas em conjunto de calibração

(xc e yc), validação (xv e yv) e predição (xp e yp), um vetor com o rótulo das variáveis (*wave*), número máximo de variáveis latentes utilizada (no_of_lv), tipo e pré-processamento utilizado (prepro_method), método de validação cruzada utilizado (val_method), um modelo iPLS usando o número de intervalos que se deseja (Model_ext) e um modelo iPLS usando o número de intervalos igual ao número de variáveis da matriz X, o que significa obter um modelo com o valor máximo de intervalos possível (Model_in), utilizando desta forma todas as variáveis do espectro.

- **Dados referentes ao GA *out*:** dados utilizados como entrada/saída do GA *out*. Estes dados são compostos pelo tamanho do cromossomo (tam_ind), tamanho da população (tam_pop), taxa de cruzamento (t_cross), taxa de mutação (t_mut), probabilidade de mutação (p_mut), taxa de elitismo (t_elite), número de gerações que o GA executou (n_geracoes), população corrente do GA (pop), histórico da evolução, contendo o valor da *fitness* - o número de variáveis latentes utilizado e o número de intervalos selecionados pelo melhor indivíduo de cada geração - (historico), o cromossomo do melhor indivíduo de cada geração (historico_selecionados), o indivíduo com melhor *fitness* encontrado em toda a evolução (ind_otimizado) e um vetor contendo o número dos intervalos selecionados pelo ind_otimizado.

- **Dados referentes ao GA *in*:** dados utilizados como entrada/saída do GA *in*. Estes dados são compostos pelo tamanho do cromossomo (tam_ind), tamanho da população (tam_pop), taxa de cruzamento (t_cross), taxa de mutação (t_mut), probabilidade de mutação (p_mut), taxa de elitismo (t_elite), número de gerações que o GA executou (n_geracoes), população corrente do GA (pop), histórico da evolução - contendo o valor da *fitness*, o número de variáveis latentes utilizado e o número de variáveis selecionadas pelo melhor indivíduo de cada geração - (historico), o cromossomo do melhor indivíduo de cada geração (historico_selecionados), o indivíduo com melhor *fitness* encontrado em toda a evolução (ind_otimizado), um vetor contendo o número dos intervalos selecionados pelo ind_otimizado e um cromossomo que indique dentro de quais intervalos da matriz de dados original o GA *in* deve atuar selecionando variáveis.

Esta formatação dos dados foi denominada de pacote. Para auxiliar o entendimento, segue um exemplo de um pacote demonstrado nos Quadros 1, 2 e 3.

pacote =
- xc: [informações espectrias das amostras de calibração]
- yc: [valor da concentração/propriedade que se deseja prever das amostras de calibração]
- xv: [informações espectrias das amostras de validação]
- yv: [valor da concentração/propriedade que se deseja prever das amostras de validação]
- xp: [informações espectrias das amostras de predição]
- yp: [valor da concentração/propriedade que se deseja prever das amostras de predição]
- wave: [todas as variáveis do espectro]
- no_of_lv: número máximo de variáveis latentes das soluções
- prepro_method: tipo de pré-processamento
- val_method: método de validação'
- GAout: [estrutura que contém todas as informações utilizadas na execução do GA-iPLS out]
- GAin: [estrutura que contém todas as informações utilizadas na execução do GA-iPLS in]
- Model_ext: [estrutura que contém todas as informações utilizadas na criação e execução de um modelo gerado através do método iPLS]
- Model_in: [estrutura que contém todas as informações utilizadas na criação e execução do modelo auxiliar que possui o número de intervalos igual ao numero de variáveis do espectro]

Quadro 1 - Dados referentes ao espectro e são utilizados como parâmetro do iPLS.

Fonte: elaborado pelo autor

pacote.GAout =
- pop: [matriz que contém todas as soluções de uma geração para o problema]
- historico: [matriz que armazena o erro de validação e o número de variáveis latentes utilizadas na melhor solução de cada geração]
- historico_selecionados: [matriz que armazena a melhor solução encontrada em cada geração durante a evolução]
- selecionados: [intervalos utilizados pela melhor solução encontrada]
- ind_otimizado: [melhor solução encontrada]
- n_geracoes: [número de gerações executadas]
- tam_ind: [número de intervalos em que o espectro foi dividido (tamanho do cromossomo)]
- tam_pop: [número de soluções em cada geração]
- t_cross: [porcentagem da população que será combinada através do cruzamento]
- t_mut: [porcentagem da população que será mutada]
- p_mut: [probabilidade referente ao número de genes que serão mutados em uma solução]
- t_elite: [porcentagem da população que fará parte da elite]

Quadro 2 - Dados utilizados pelo GA-iPLS *out*

Fonte: elaborado pelo autor

pacote.GAin =
ind_ext: [intervalos do espectro selecionados por uma solução inicial que será refinada]
pop: [matriz que contém todas as soluções de uma geração para o problema]
historico: [matriz que armazena o erro de validação e o número de variáveis latentes utilizadas na melhor solução de cada geração]
historico_selecionados: [matriz que armazena a melhor solução encontrada em cada geração durante a evolução]
selecionados: [intervalos utilizados pela melhor solução encontrada]
ind_otimizado: [melhor solução encontrada]
n_geracoes: [número de gerações executadas]
tam_ind: [número de intervalos em que o espectro foi dividido (tamanho do cromossomo)]
tam_pop: [número de soluções em cada geração]
t_cross: [porcentagem da população que será combinada através do cruzamento]
t_mut: [porcentagem da população que será mutada]
p_mut: [probabilidade referente ao número de genes que serão mutados em uma solução]
t_elite: [porcentagem da população que fará parte da elite]

Quadro 3 - Dados utilizados pelo GA-iPLS *in*.

Fonte: elaborado pelo autor

4 RESULTADOS

Os resultados obtidos pelos algoritmos desenvolvidos nesta pesquisa são comparados com os resultados de alguns métodos já existentes no qual este estudo foi baseado, como o PLS e o iPLS.

Como não foi abordado um único problema, cada distinto conjunto de dados submetido à otimização através do algoritmo genético será tratado separadamente, discorrendo sobre a importância para a indústria, os motivos que levaram a pesquisa e a natureza dos dados de cada problema, mostrando a forma de aquisição do espectro e explanando também sobre os resultados obtidos através de diferentes métodos, comparando-os.

4.1 Determinação do índice de OH de polióis de óleo de soja

A indústria de poliuretano tem interesse neste tipo de análise, pois utilizam polióis na fabricação de vários materiais, como revestimento de assoalhos, adesivos, poliuretanos termoplásticos e materiais livres de compostos orgânicos voláteis. A determinação do teor de hidroxilas de polióis de óleo de soja é muito importante na preparação de poliuretano (SABIN *et al*, 2006).

As amostras de polióis foram preparadas no Instituto de Química da Universidade Federal do Rio Grande do Sul (UFRGS) a partir de óleo de soja refinado obtido da CBM Ind. Com. Distrib. Ltda , éter etílico provindo da Synth e solução de peróxido de hidrogênio a 30%, cloreto de sódio, bicarbonato de sódio, bissulfito de sódio e sulfato de sódio obtido da Nuclear. Utilizaram-se métodos de titulação para a determinação do valor de hidroxilas como recomendado pela *American Oil Chemists' Society* (AOCS), cujo valor é expresso em miligramas de hidróxido de potássio (KOH) por grama de amostra. A Figura 16 demonstra a estrutura de um poliol que pode ser formado a partir do óleo de soja.

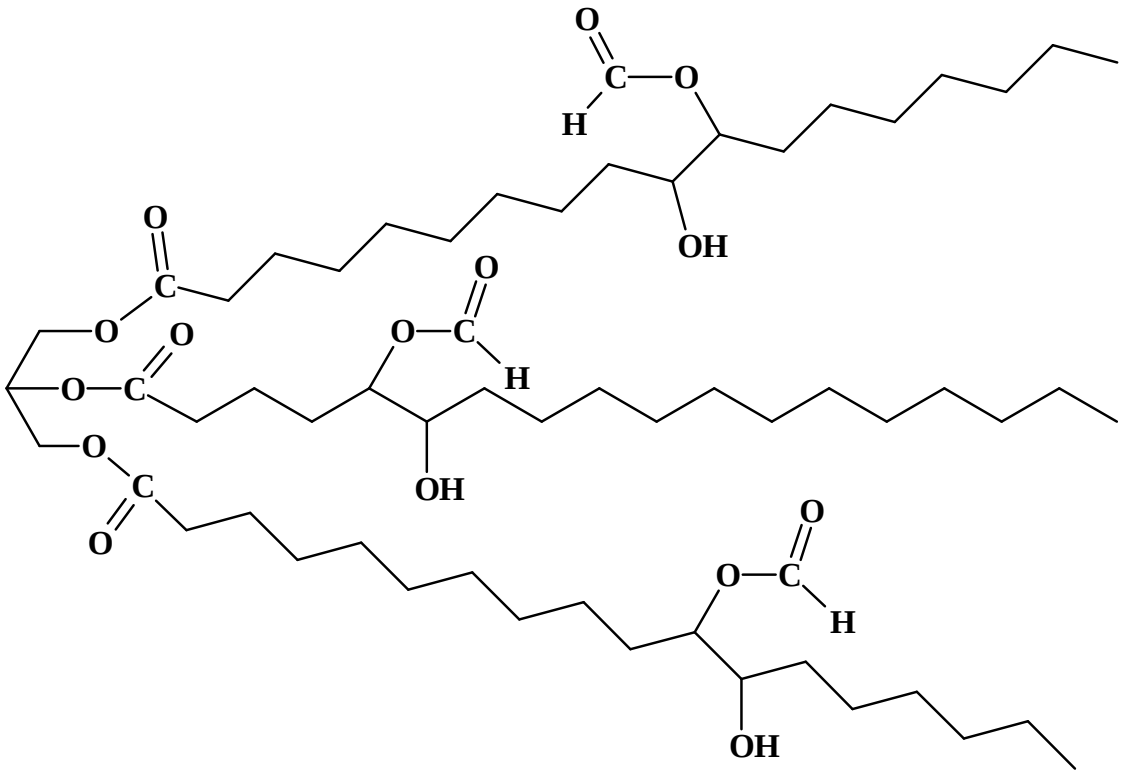


Figura 16 – Fórmula estruturada de um poliol

Fonte: elaborado pelo autor

Para a obtenção dos espectros, foi utilizado um espectrofotômetro Nicolet Magma 550 FT-IR com um acessório de reflectância total atenuada horizontal (HATR) equipado com cristal de seleneto de zinco. Estes dados estão disponíveis na base de dados do grupo de pesquisa em Quimiometria da Universidade de Santa Cruz do Sul (UNISC), de onde foram adquiridos para a realização deste estudo.

Os dados foram separados em dois distintos conjuntos: calibração e predição. O conjunto de calibração é composto por 42 amostras e o conjunto de predição é composto por 20 amostras, optando-se por utilizar validação cruzada. A faixa espectral compreende os números de onda que vão de 649 até 1805 cm^{-1} e os valores de concentração de hidroxilas de polióis de óleo de soja das amostras analisadas ficaram entre 23,66 e 195,04 miligramas por grama de amostra.

Para submeter a matriz de espectros aos métodos de regressão desejados neste estudo, foi necessária a aplicação de um pré-processamento onde os dados foram auto-escalados para que os resultados obtidos apresentassem maior precisão.

4.1.1 Resultados obtidos aplicando o PLS

Os resultados obtidos através da aplicação do método PLS, que utiliza todo o espectro para criar um modelo de regressão multivariada, é apresentado na Tabela 1.

Tabela 1 – Resultados do modelo de regressão obtido com o método PLS para a determinação de OH em polióis de óleo de soja

Nº de freqüências selecionadas	**VL**	**Calibração**		**Predição**	
		R^2_{cal}	**RMSECV** (mg de KOH/g de amostra)	**R^2_{pred}**	**RMSEP** (mg de KOH/g de amostra)
600	3	0,9894	7,23	0,9915	6,8

Fonte: elaborado pelo autor

4.1.2 Resultados obtidos aplicando o iPLS

Os resultados obtidos através da aplicação do método iPLS, que avalia cada um dos intervalos e retorna um modelo de regressão multivariada feito sobre o intervalo que apresentar melhor a resposta, é apresentado na Tabela 2.

Tabela 2 – Resultados dos modelos de regressão obtidos através do método iPLS para a determinação de OH em polióis de óleo de soja, dividindo o espectro em 20, 30 e 60 intervalos

	Nº de freqüências selecionadas	**VL**	**Calibração**		**Predição**	
			R^2_{cal}	**RMSECV** (mg de KOH/g de amostra)	**R^2_{pred}**	**RMSEP** (mg de KOH/g de amostra)
iPLS 20	**30**	**3**	**0,9896**	**7,15**	**0,9892**	**6,28**
iPLS 30	20	2	0,9901	6,98	0,9890	6,46
iPLS 60	10	4	0,9898	7,08	0,9851	7,52

Fonte: elaborado pelo autor

Figura 17 demonstra os erros de validação cruzada utilizando o método iPLS com o espectro dividido em 20 intervalos, pois esta foi a configuração que apresentou o melhor resultado conforme a Tabela 2, onde a linha pontilhada representa o erro do modelo que utiliza todo o espectro e as barras representam os erros dos modelos construídos para cada intervalo individualmente.

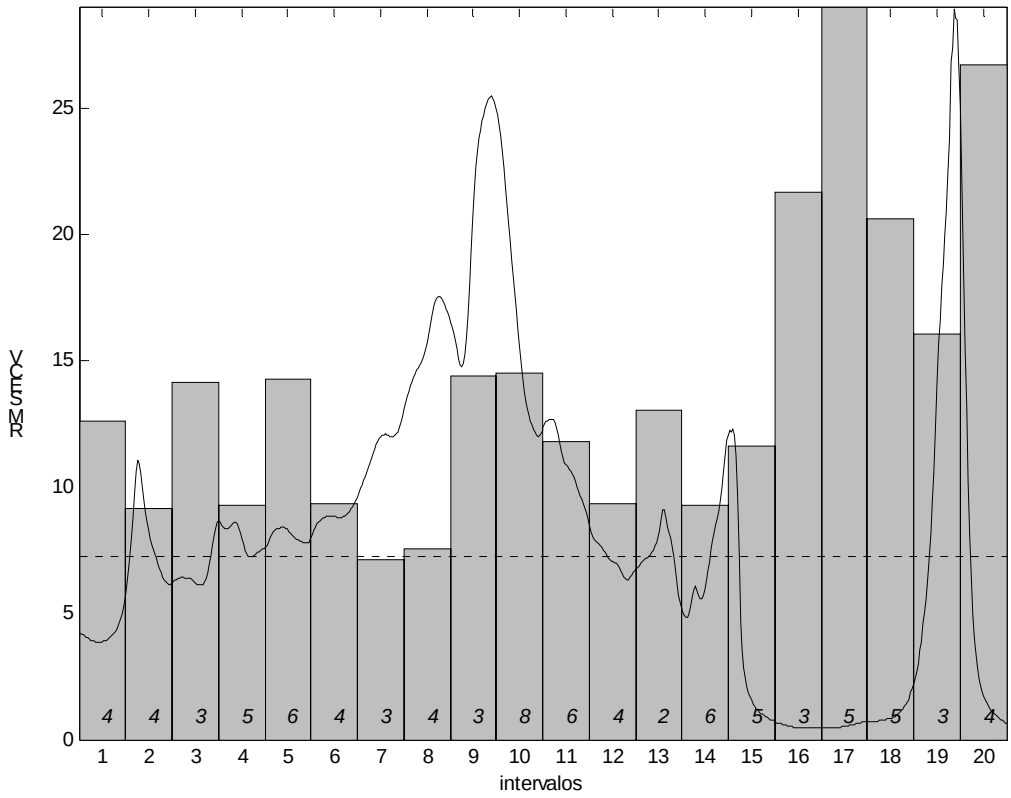


Figura 17 – Gráfico dos erros do modelo iPLS, dividindo o espectro de polióis de óleo de soja em 20 intervalos

Fonte: elaborado pelo autor

Com base na Figura 17 observa-se que o modelo com melhor desempenho já alcançado com a subdivisão do espectro em 20 intervalos foi gerado utilizando o 7º intervalo e 3 variáveis latentes, sendo na Figura 18 apresentado o espectro, ressaltando a área selecionada pela aplicação do método iPLS com esta configuração.

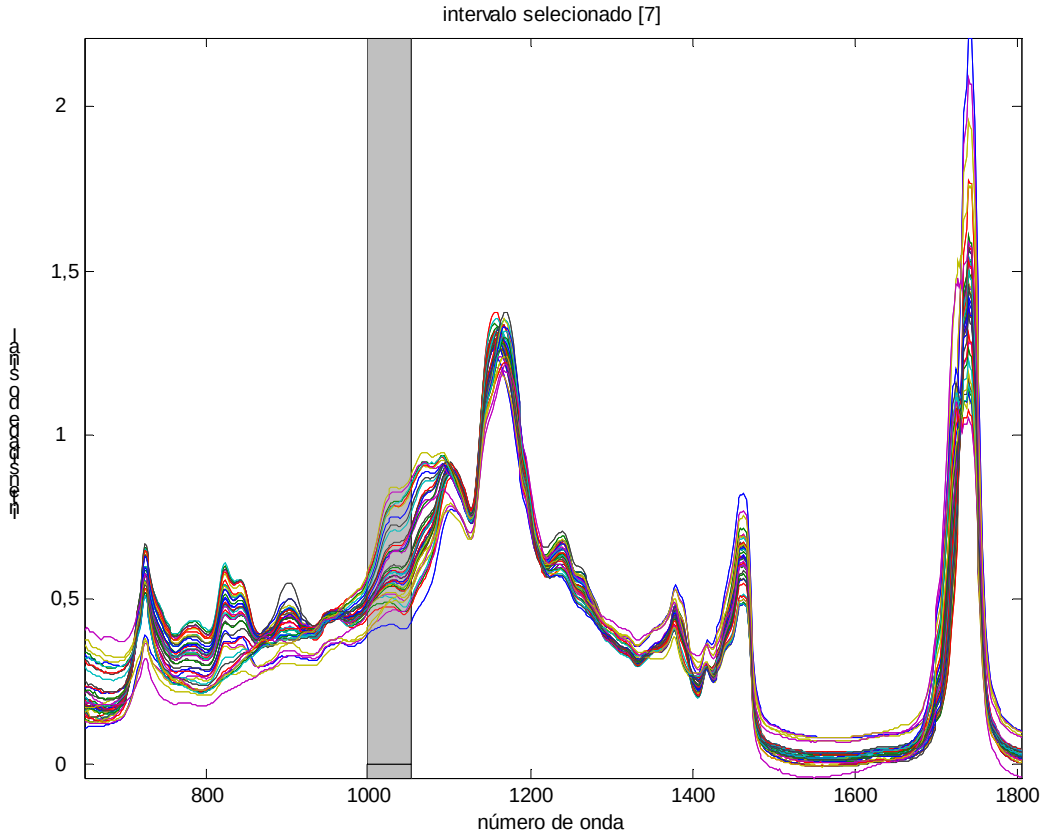


Figura 18 – Espectro de polióis de óleo de soja, ressaltando a região selecionada pelo método iPLS com o espectro dividido em 20 intervalos

Fonte: elaborado pelo autor

A regressão, o seu coeficiente e o RMSEP das amostras de predição sobre o modelo gerado através do método iPLS para o intervalo 7 utilizando as amostras de calibração, são apresentados na Figura 19.

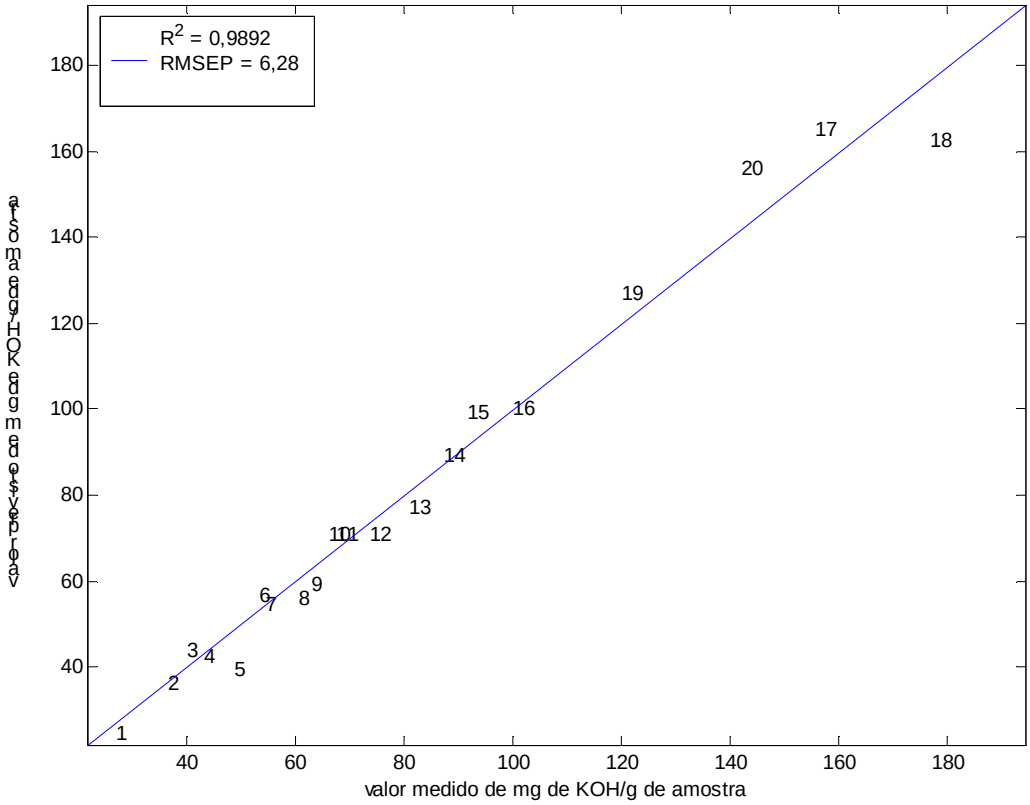


Figura 19 – Predição de OH de polióis de óleo de soja para o modelo gerado utilizando o 7º intervalo do método iPLS com o espectro dividido em 20 intervalos

Fonte: elaborado pelo autor

4.1.3 Resultados obtidos aplicando o GA-iPLS *out*

Os resultados obtidos através da aplicação do método GA-iPLS *out*, que retorna um modelo de regressão multivariada obtido pela combinação dos intervalos selecionados pela otimização do GA, é apresentado na Tabela 3.

Tabela 3 – Resultados da aplicação do GA-iPLS *out* para a determinação de OH em polióis de óleo de soja, dividindo o espectro em 20, 30 e 60 intervalos

		Nº de freqüências selecionadas	**VL**	**Calibração**		**Predição**	
				R^2_{cal}	**RMSECV** (mg de KOH/g de amostra)	**R^2_{pred}**	**RMSEP** (mg de KOH/g de amostra)
GA-iPLS *out* 20 intervalos	1ª execução	210	4	0,9917	6,41	0,9925	6,32
	2ª execução	300	3	0,9926	6,04	0,9909	6,54
	3ª execução	**210**	**5**	**0,9926**	**6,03**	**0,9933**	**5,74**
GA-iPLS *out* 30 intervalos	1ª execução	140	7	0,9930	5,86	0,9922	5,76
	2ª execução	260	4	0,9915	6,48	0,9944	5,88
	3ª execução	**220**	**3**	**0,9928**	**5,96**	**0,9942**	**5,70**
GA-iPLS *out* 60 intervalos	1ª execução	250	3	0,9929	5,93	0,9936	5,90
	2ª execução	270	3	0,9921	6,22	0,9941	5,88
	3ª execução	**270**	**3**	**0,9920**	**6,28**	**0,9943**	**5,74**

Fonte: elaborado pelo autor

As Figuras 20, 21 e 22 mostram um comparativo entre as evoluções das três diferentes execuções, para cada configuração do GA-iPLS *out*, com 20, 30 e 60 intervalos respectivamente. Observa-se que todas as execuções para 30 e 60 intervalos apresentam resultados muito próximos, quando são realizadas 500 iterações. Já as execuções para 20 intervalos foram bastante distintas, evidenciando uma maior dificuldade em combinar os intervalos (sinais analíticos) mais representativos. Apesar disto, a melhor execução em cada caso resultou em RMSEP equivalentes.

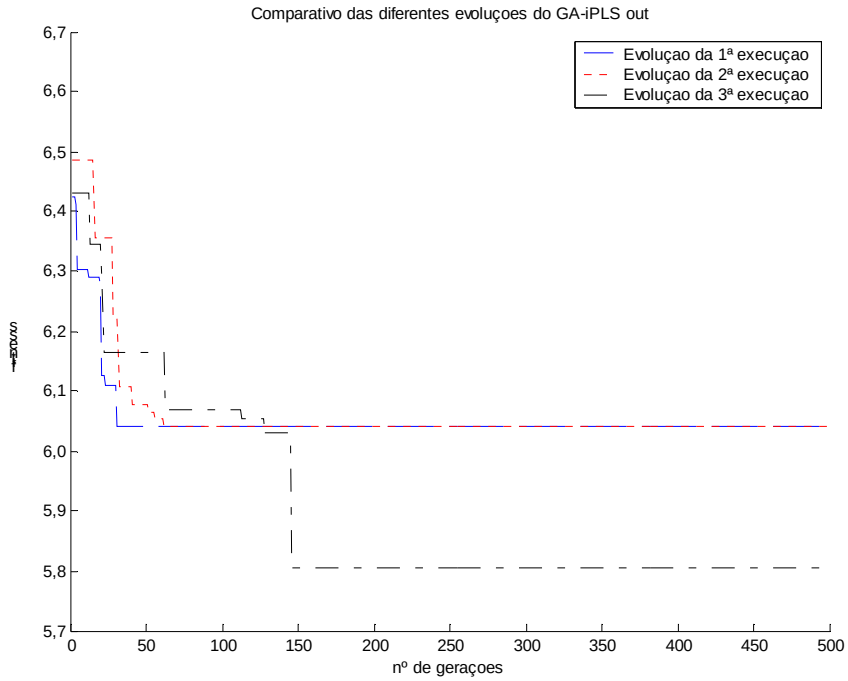


Figura 20 - Evoluções das três execuções do GA-iPLS out para a determinação de OH em polióis de óleo de soja, dividindo o espectro em 20 intervalos

Fonte: elaborado pelo autor

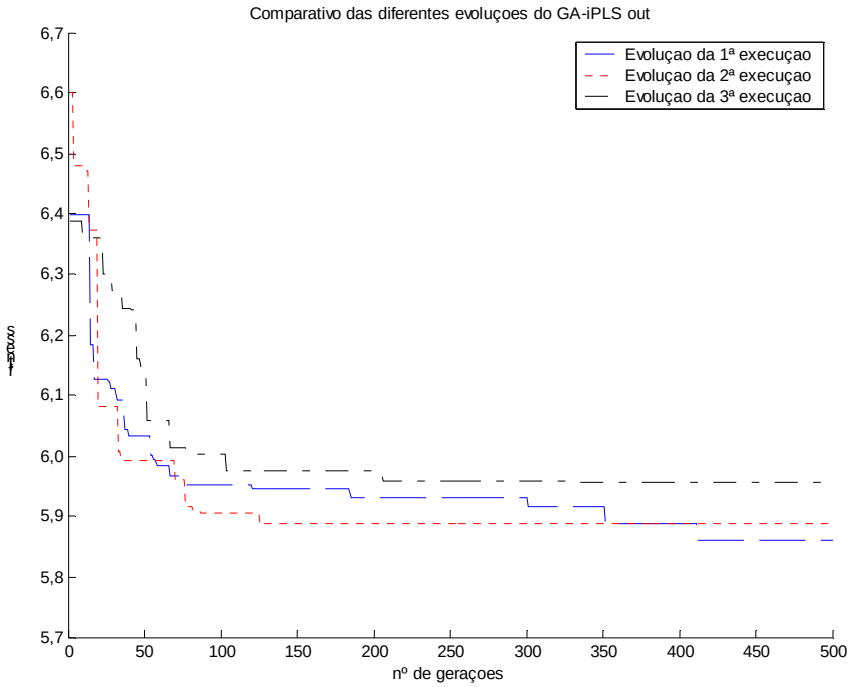


Figura 21 - Evoluções das três execuções do GA-iPLS out para a determinação de OH em polióis de óleo de soja, dividindo o espectro em 30 intervalos

Fonte: elaborado pelo autor

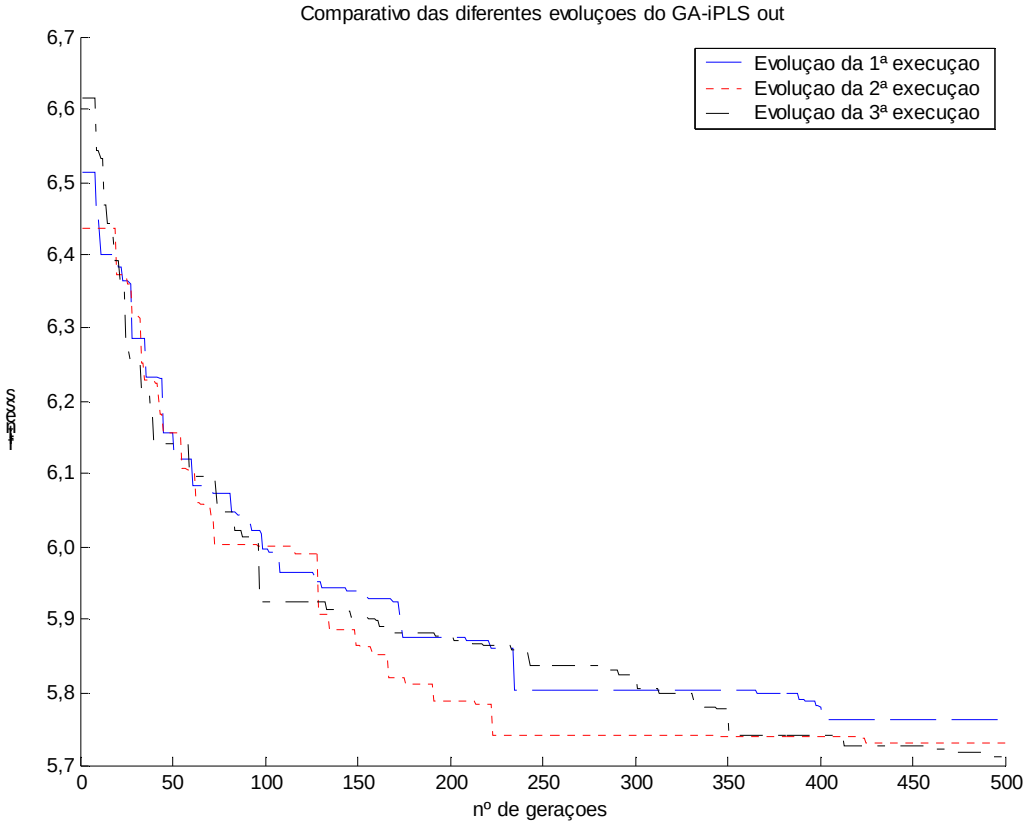


Figura 22 - Evoluções das três execuções do GA-iPLS out para a determinação de OH em polióis de óleo de soja, dividindo o espectro em 60 intervalos

Fonte: elaborado pelo autor

Com base nestes resultados, foi selecionada a 3ª execução utilizando 30 intervalos, conforme Figura 23 que apresenta o espectro de polióis de óleo de soja, onde as barras verdes ressaltam as regiões selecionadas pelo algoritmo GA-iPLS *out*. Dentre os sinais selecionados pode-se destacar a região próxima a 1720 - 1725 cm^{-1} referente ao estiramento da carbonila de formato (BARBOSA, 2007).

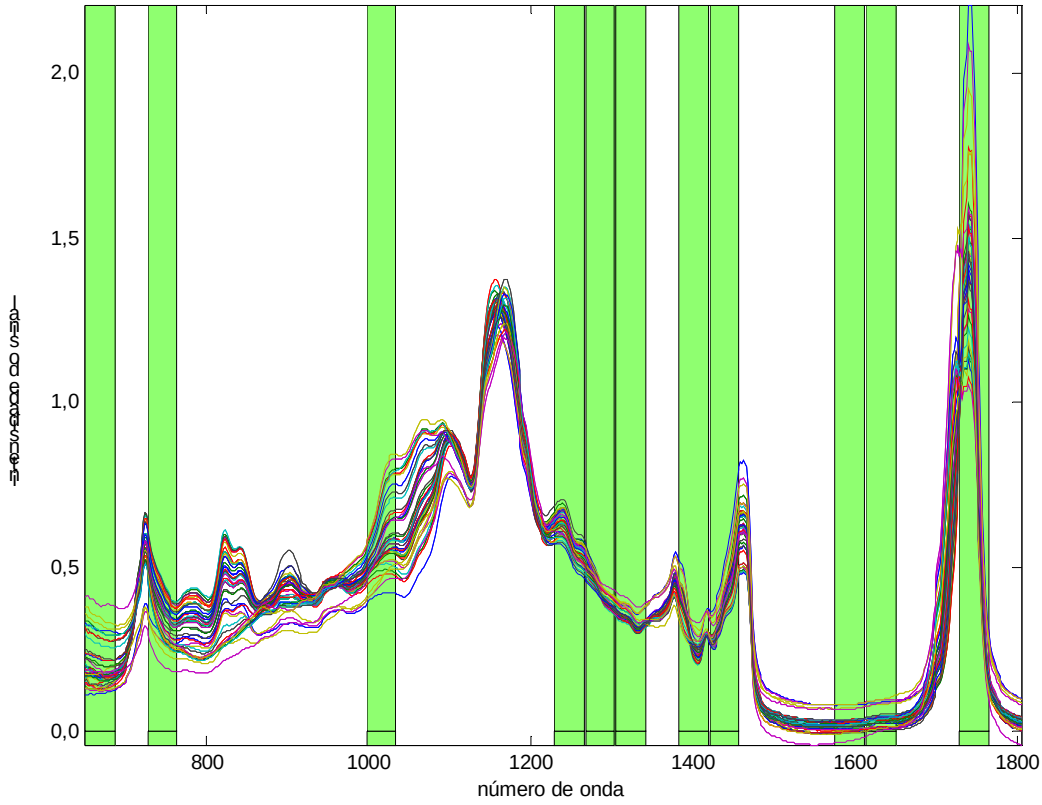


Figura 23 – Espectro de polióis de óleo de soja, ressaltando as regiões selecionadas pelo método GA-iPLS out, para o espectro dividido em 30 intervalos

Fonte: elaborado pelo autor

O bom comportamento para a predição das amostras externas para o referido modelo pode ser visualizado na Figura 24, onde a regressão, o seu coeficiente e o RMSEP das amostras de predição sobre o modelo gerado através do método GA-iPLS *out* dividindo o espectro em 30 intervalos.

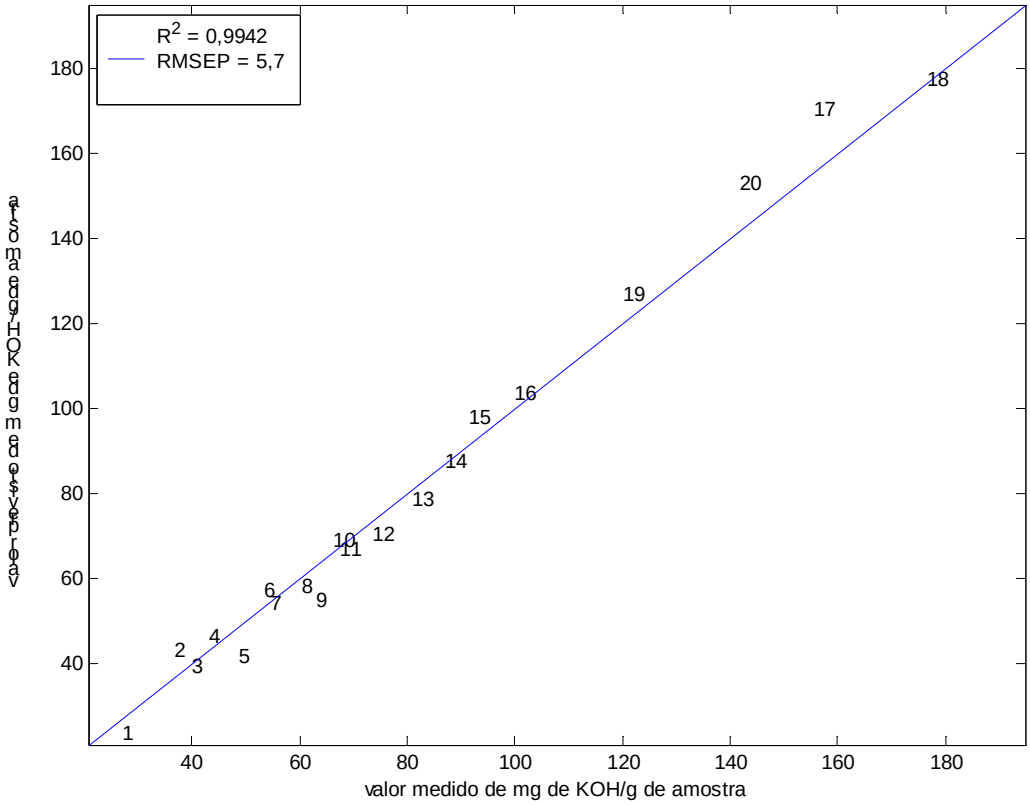


Figura 24 – Predição de OH de polióis de óleo de soja sobre o modelo gerado pelo método GA-iPLS *out* dividindo o espectro em 30 intervalos

Fonte: elaborado pelo autor

4.1.4 Resultados Obtidos aplicando o GA-iPLS *in*

As variáveis que fazem parte deste processo de seleção são aquelas contidas nos intervalos indicados a este algoritmo como resposta inicial, que neste caso são as respostas obtidas pelo GA-iPLS *out*. Os resultados obtidos através da aplicação do método GA-iPLS *in* são os apresentados na Tabela 4, para cada um dos melhores resultados encontrados para 20, 30 e 60 intervalos.

Tabela 4 – Resultados da aplicação do GA-iPLS *in* para a determinação de OH em polióis de óleo de soja, refinando as melhores soluções encontradas pelo GA-iPLS *out*

	Nº de freqüências selecionadas	**VL**	**Calibração**		**Predição**	
			R^2_{cal}	**RMSECV** (mg de KOH/g de amostra)	**R^2_{pred}**	**RMSEP** (mg de KOH/g de amostra)
Solução GA-iPLS out 20 intervalos	210	5	0,9926	6,03	0,9933	5,74
1ª execução	107	5	0,9943	5,31	0,9921	5,76
2ª execução	106	5	0,9936	5,63	0,9927	5,69
3ª execução	**98**	**5**	**0,9931**	**5,82**	**0,9935**	**5,67**
Solução GA-iPLS out 30 intervalos	220	3	0,9928	5,96	0,9942	5,70
1ª execução	**113**	**3**	**0,9931**	**5,84**	**0,9947**	**5,50**
2ª execução	119	3	0,9929	5,92	0,9946	5,55
3ª execução	81	3	0,9934	5,69	0,9946	5,52
Solução GA-iPLS out 60 intervalos	270	3	0,9920	6,28	0,9943	5,74
1ª execução	119	3	0,9933	5,74	0,9943	5,57
2ª execução	141	3	0,9926	6,03	0,9948	5,59
3ª execução	**124**	**3**	**0,9924**	**6,13**	**0,9950**	**5,34**

Fonte: elaborado pelo autor

As Figuras 25, 26 e 27 mostram um comparativo entre as evoluções das três diferentes execuções, para cada configuração do GA-iPLS *in*, sobre a melhor resposta do GA-iPLS *out* com o espectro dividido em 20, 30 e 60 intervalos respectivamente. Neste Caso todas as repetições evoluíram de forma semelhante, porém para 20 intervalos não foi observada melhora significativa.

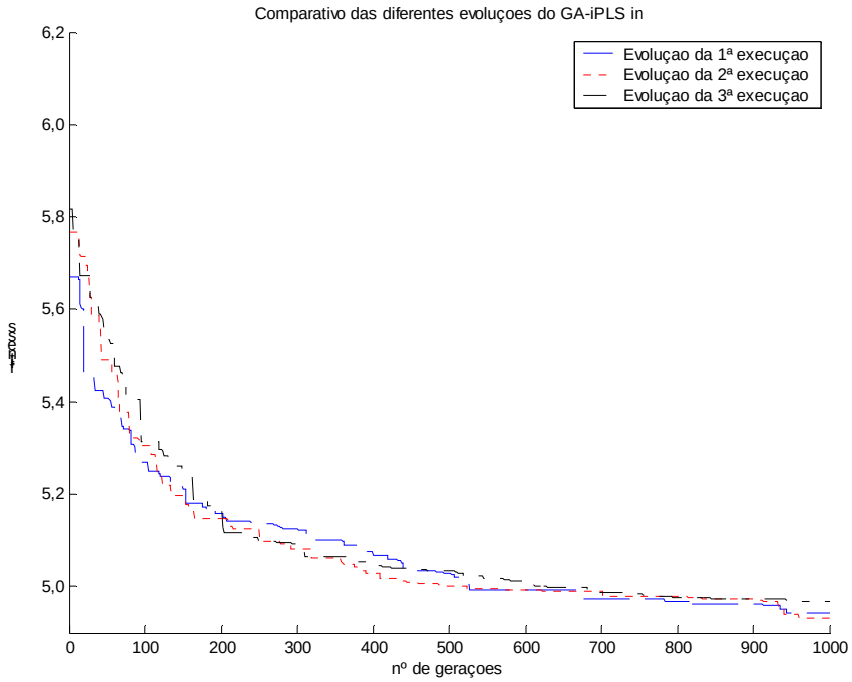


Figura 25 – Evoluções das três execuções do GA-iPLS *in* para a determinação de OH em polióis de óleo de soja, sobre a melhor resposta do GA-iPLS *out* com o espectro dividido em 20 intervalos

Fonte: elaborado pelo autor

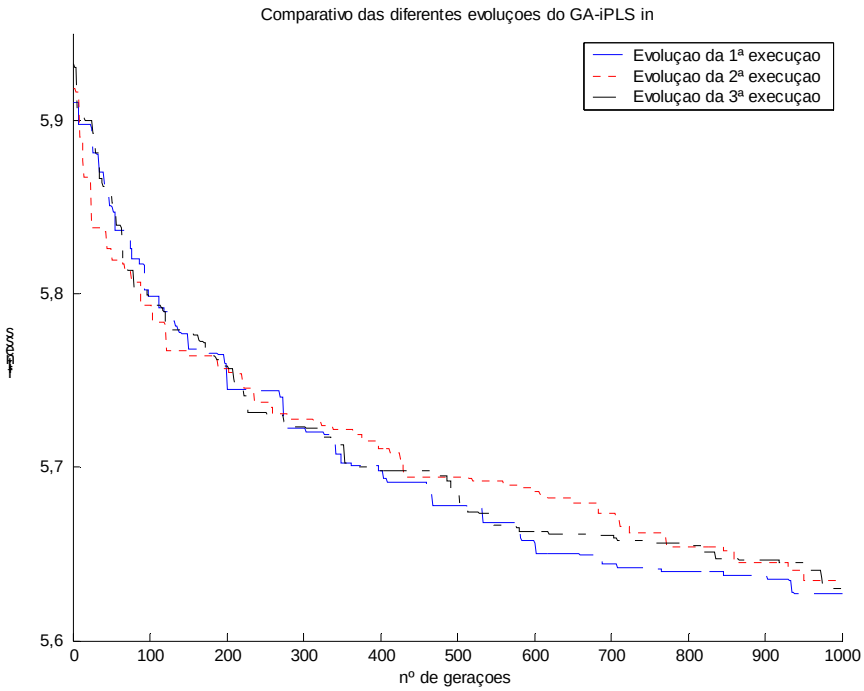
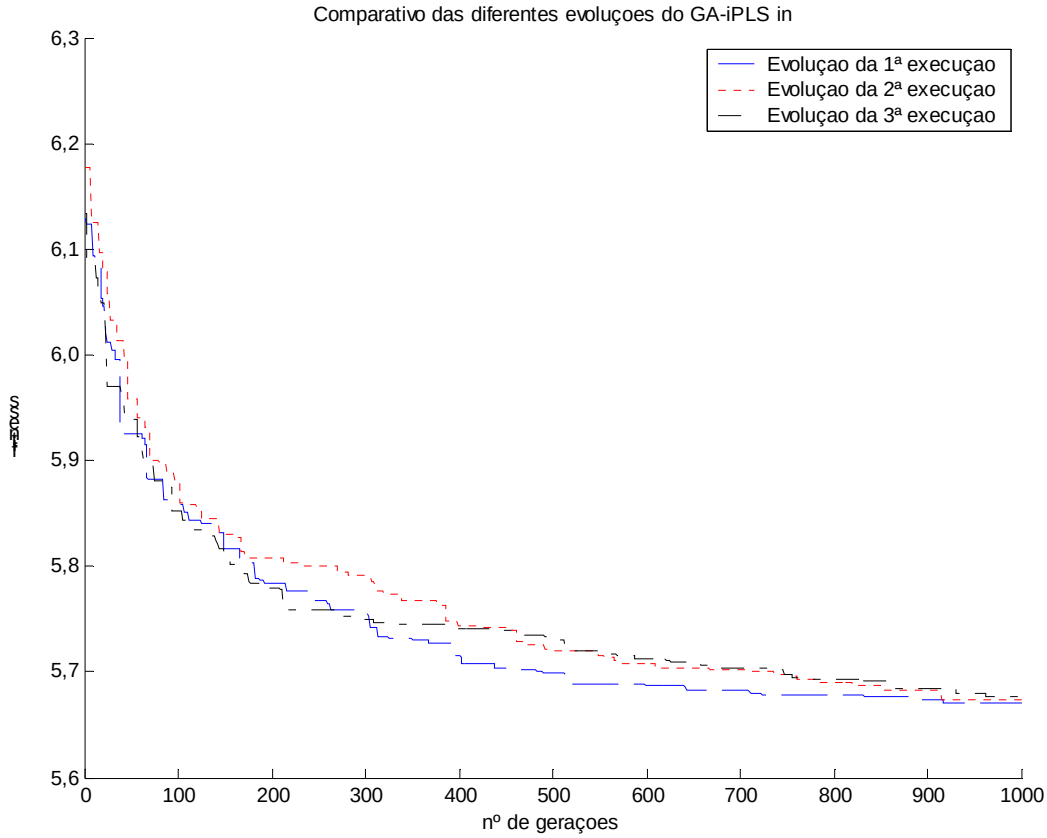


Figura 26 - Evoluções das três execuções do GA-iPLS *in* para a determinação de OH em polióis de óleo de soja, sobre a melhor resposta do GA-iPLS *out* com o espectro dividido em 30 intervalos

Fonte: elaborado pelo autor



Figusra 27 - Evoluções das três execuções do GA-iPLS *in* para a determinação de OH em polióis de óleo de soja, sobre a melhor resposta do GA-iPLS *out* com o espectro dividido em 60 intervalos

Fonte: elaborado pelo autor

Com base nestes resultados selecionou-se a terceira execução do algoritmo genético sobre a solução do GA-iPLS *out* com 60 intervalos, que resultou em um RMSEP igual a 5,34 mg de KOH/g de amostra. As variáveis selecionadas pelo GA-iPLS *in* mapeadas em uma solução alcançada pelo GA-iPLS *out* são demonstradas na Figura 28 e a regressão das amostras de predição para o modelo obtido por este método é apresentado na Figuras 29. Dentre os sinais selecionados pode-se destacar a região próxima a 1720 - 1725 cm^{-1} referente ao estiramento da carbonila de formato e a região próxima de 1190 cm^{-1} referente ao estiramento O-C-C do grupo éster alifático saturado (BARBOSA, 2007). Neste caso houve uma redução de 79,33% no número de variáveis espectrais utilizadas na criação do modelo de regressão multivariado encontrado como resposta pelo GA-iPLS.

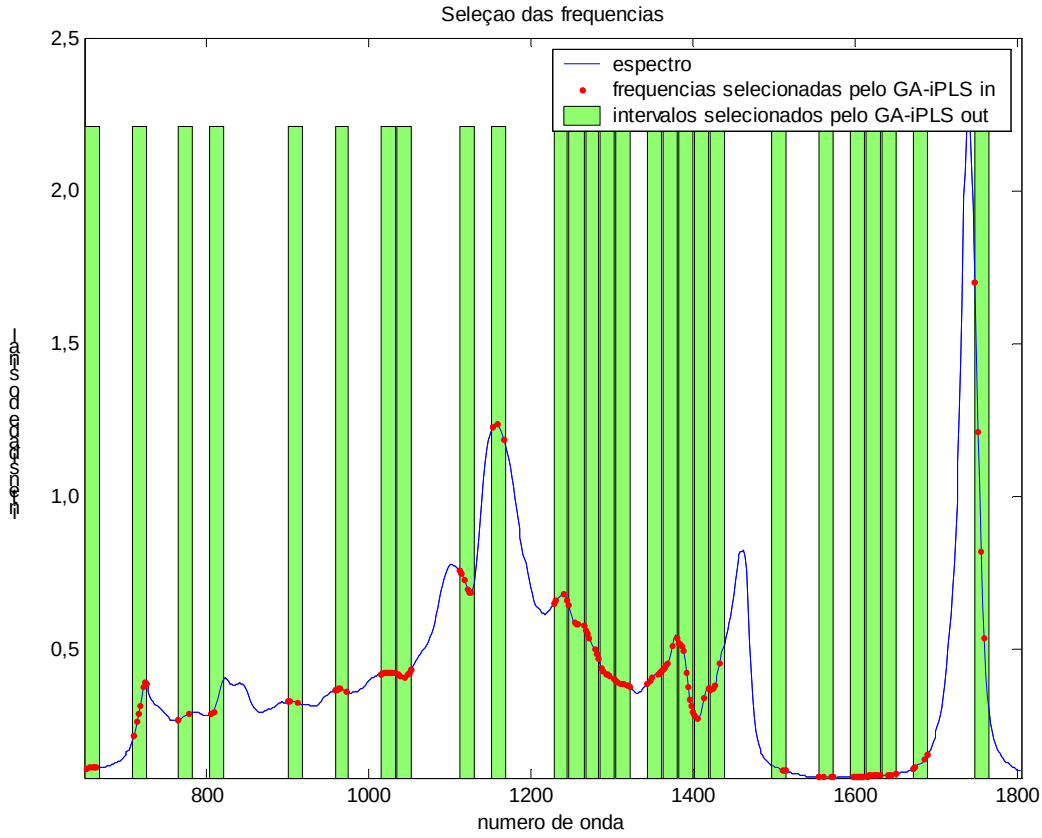


Figura 28 – Espectro de polióis de óleo de soja, ressaltando as regiões selecionados pelo método GA-iPLS in, sobre a solução encontrada pelo GA-iPLS out com o espectro dividido em 60 intervalos

Fonte: elaborado pelo autor

A regressão, o seu coeficiente e o RMSEP das amostras de predição sobre o modelo gerado através do método GA-iPLS *in*, que refina a solução do GA-iPLS *out* utilizando as amostras de calibração e dividindo o espectro em 30 intervalos, são apresentados na Figura 29.

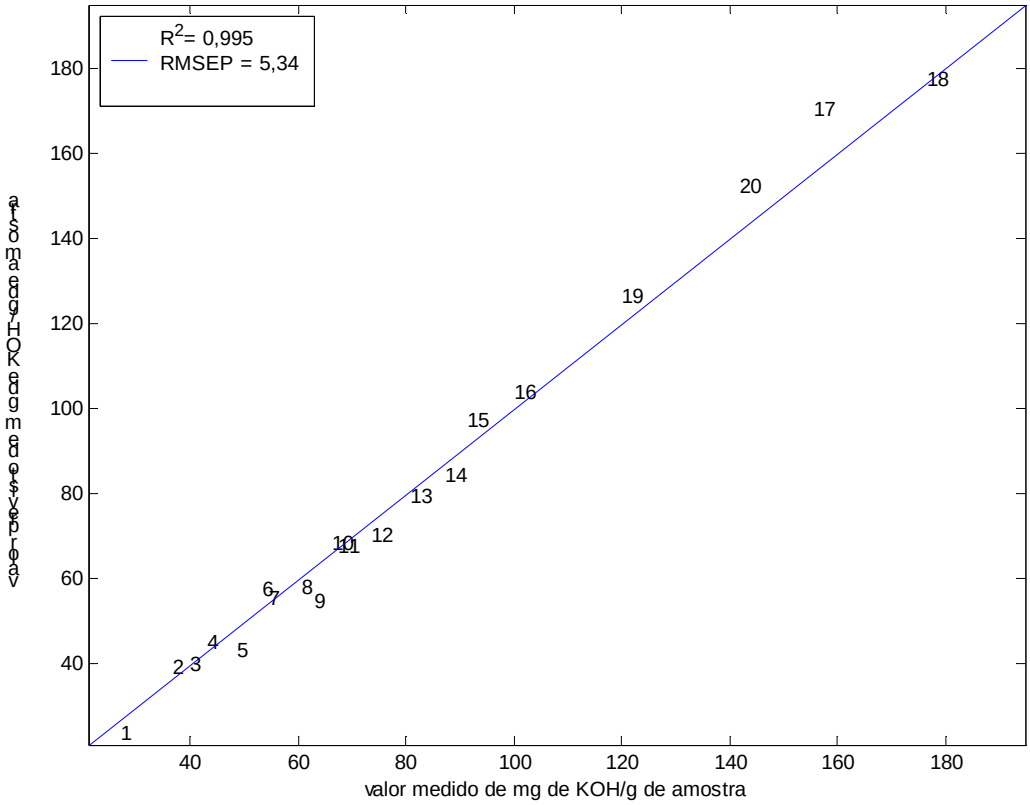


Figura 29 – Predição de OH de polióis de óleo de soja sobre o modelo gerado pelo método GA-iPLS *in* gerado a partir da solução obtida pelo GA-iPLS *out* dividindo o espectro em 60 intervalos

Fonte: elaborado pelo autor

A Tabela 5 apresenta o valor de miligramas de KOH por grama de cada amostra de predição ($\mathbf{y_{ref}}$) e o valor previsto ($\mathbf{y_{pred}}$), o erro percentual de cada amostra (**%**) e a média dos erros percentuais para cada um dos métodos empregados neste estudo.

Tabela 5 - Valores medidos e previstos de OH de polióis de óleo de soja e os erros percentuais para as amostras externas

	y_{ref} (mg de KOH/g de amostra)	PLS		iPLS		GA-iPLS out		GA-iPLS in	
		y_{pred} (mg de KOH/g de amostra)	%	y_{pred} (mg de KOH/g de amostra)	%	y_{pred} (mg de KOH/g de amostra)	%	y_{pred} (mg de KOH/g de amostra)	%
1	25,9000	24,3222	6,09	24,3438	6,01	22,7813	12,04	23,4346	9,52
2	35,4300	47,7201	34,69	35,9125	1,36	42,3242	19,46	38,9117	9,83
3	38,8600	38,7018	0,41	43,7889	12,68	39,7155	2,20	39,7885	2,39
4	41,9400	50,1729	19,63	42,1219	0,43	46,8169	11,63	44,7842	6,78
5	47,6400	41,5995	12,68	39,2019	17,71	42,2930	11,22	43,1655	9,39
6	52,3100	55,6046	6,30	56,6260	8,25	57,1593	9,27	57,6170	10,15
7	53,4600	53,1678	0,55	54,3462	1,66	54,6071	2,15	55,3647	3,56
8	59,3700	63,6655	7,24	55,9862	5,70	59,4136	0,07	57,9052	2,47
9	61,7500	55,0983	10,77	59,1307	4,24	54,6641	11,48	54,4044	11,90
10	66,1900	67,8343	2,48	71,0062	7,28	68,0977	2,88	68,2005	3,04
11	67,4500	67,7215	0,40	70,8298	5,01	67,5270	0,11	67,5325	0,12
12	73,6200	72,8051	1,11	70,9304	3,65	71,1082	3,41	70,0501	4,85
13	80,8600	79,3609	1,85	76,9280	4,86	78,9267	2,39	79,3537	1,86
14	87,0800	82,0913	5,73	89,1778	2,41	84,8074	2,61	84,0234	3,51
15	91,4100	95,5254	4,50	99,0138	8,32	96,9604	6,07	97,1794	6,31
16	100,0500	105,7229	5,67	100,1378	0,09	104,1176	4,07	103,5745	3,52
17	155,4300	174,4022	12,21	165,0937	6,22	172,1655	10,77	170,2513	9,54
18	176,5400	178,6489	1,19	162,6467	7,87	177,9329	0,79	177,0965	0,32
19	120,0600	126,7781	5,60	126,8878	5,69	126,5904	5,44	126,2780	5,18
20	141,8700	152,1354	7,24	155,9923	9,95	150,9770	6,42	152,0434	7,17
Erro percentual médio:			**7,32**		**5,97**		**6,22**		**5,57**

Fonte: elaborado pelo autor com base nos resultados obtidos

Conforme evidenciado, bons resultados para a determinação de OH de polióis de óleo de soja foram encontrados utilizando-se o GA-iPLS *out* e o GA-iPLS *in* em comparação com os resultados alcançados pelo PLS e iPLS, demonstrando que houve uma otimização dos modelos de regressão multivariados para este problema. Embora o RMSEP obtido utilizando-se o método GA-iPLS *in* não tenha diminuído de forma tão expressiva em comparação ao que foi alcançado através do GA-iPLS *out*, o erro médio percentual das amostras de predição foi reduzido e um decremento considerável no número de variáveis espectrais envolvidas, próximo a 43%, indicando que o GA-iPLS *in* é capaz de refinar as soluções e encontrar modelos mais robustos.

4.2 Determinação de cloridrato de propranolol em fármacos anti-hipertensivos

A preparação das amostras e a aquisição dos dados espectrais foram realizadas pelo Instituto de Química da Universidade Federal de Santa Maria (UFSM). Para a moagem e homogeneização das amostras, foi utilizado um moinho criogênico Spex Certiprep 6750 Freezer Mill com argônio líquido. O cloridrato de propranolol utilizado como referência, no teor de 100%, foi fornecido pela Farmacopéia Brasileira[5] e as amostras analisadas foram obtidas de medicamentos legalmente comercializados e o intervalo de concentração foi de 0,10 a 0,46 miligramas de cloridrato de propranolol por miligrama de amostra (ZENI, 2005). Para complementar as informações sobre esta substância, a Figura 30 apresenta a estrutura do cloridrato de propranolol.

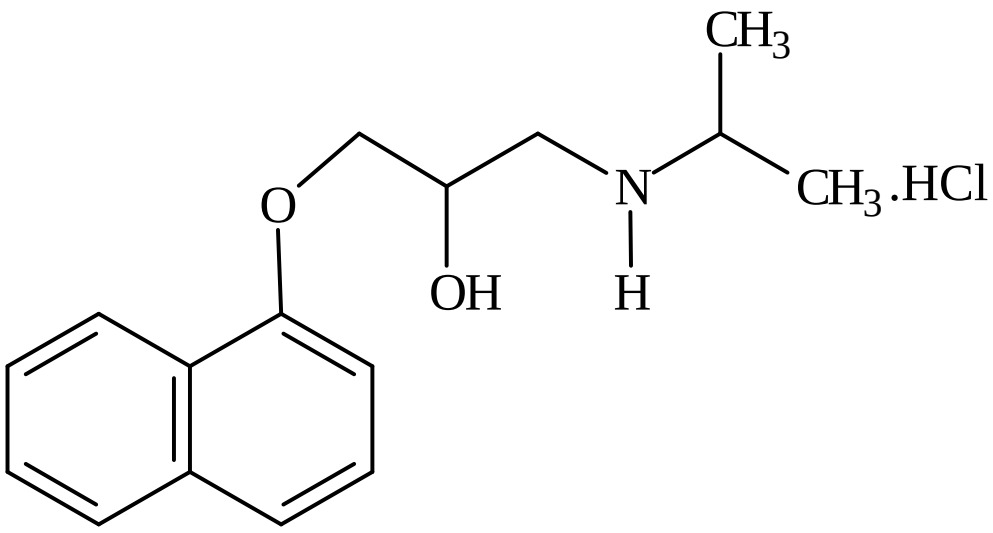


Figura 30 – Fórmula estruturada do cloridrato de propranolol

Fonte: Elaborado pelo autor.

Para a determinação da concentração de cloridrato de propranolol, utilizou-se um espectômetro de absorção no ultravioleta visível Shimadzu Multispec-1501, de acordo com a monografia número 143.1 da Farmacopéia Brasileira. Para a obtenção do espectro utilizou-se

[5] Lote 1005

um espectômetro com transformada de Fourier Perkin Elmer Spectrum One com dispositivo de ATR com cristal de seleneto de zinco (ZENI, 2005).

Os dados foram separados em três distintos conjuntos: calibração, validação e predição. O conjunto de calibração é composto por 15 amostras, o conjunto de validação é composto por 6 amostras e o conjunto de predição é composto por 5 amostras. A faixa espectral compreende aos números de onda que vão de 650 até 4000 cm^{-1} e os valores de concentração de cloridrato de propranolol das amostras analisadas ficaram entre 0,1042 e 0,4679 miligramas por miligrama de amostra.

Para submeter este espectro aos métodos de regressão desejados neste estudo, foi necessária a aplicação de um pré-processamento onde os dados foram passados pelo método de correção do espalhamento da luz (MSC), devido a amostra apresentar diferentes tamanhos de partículas, e auto-escalados para que os resultados da regressão não sejam intensamente afetados pela magnitude dos sinais e sim pela variação destes entre as diferentes amostras.

A primeira tentativa de otimização dos modelos de regressão multivariados foi utilizando a *fitness* RMSEV, explicada no item 3.2.2.2, onde o grau de adaptação das soluções é o próprio erro das amostras de validação.

Como as soluções encontradas utilizando outra técnica de *fitness* foram melhores que estas, a Tabela 6 apresenta somente os melhores resultados obtidos com RMSEV como *fitness*, juntamente com os resultados do PLS e do iPLS para fins de comparação. Os algoritmos foram executados dividindo o espectro em 25, 50 e 60 intervalos.

Tabela 6 - Comparação entre as melhores respostas obtidas através do PLS, iPLS, GA-iPLS *out* e GA-iPLS *in*

	Nº freqüências selecionadas	VL	Calibração		Validação		Predição	
			R^2_{cal}	RMSECV (mg de cloridrato de propranolol por mg de amostra)	R^2_{val}	RMSEV (mg de cloridrato de propranolol por mg de amostra)	R^2_{pred}	RMSEP (mg de cloridrato de propranolol por mg de amostra)
PLS	3351	9	0,9980	0,0055	0,9916	0,0095	0,9901	0,0223
iPLS 50	67	2	0,7996	0,0531	0,8990	0,0296	0,9058	0,0327
GA-iPLS out 25	1742	11	0,9993	0,0032	0,9999	0,0009	0,9962	0,0221
GA-iPLS in	876	11	0,9993	0,0032	0,9999	0,0012	0,9967	0,0193

Fonte: elaborado pelo autor

Tentou-se também, atingindo melhores resultados, utilizar a técnica da *fitness* composta pelo RMSECV e o RMSEV. Este método será abordado de forma mais aprofundada a seguir.

4.2.1 Resultados obtidos aplicando o PLS

Todos os resultados que serão demonstrados foram obtidos utilizando-se a técnica de *fitness* composta, calculada com base no RMSECV e RMSEV. A Tabela 7 apresenta os resultados obtidos através do método PLS, ou seja, utilizando todas as informações do espectro.

Tabela 7 - Resultados do modelo de regressão obtido com o método PLS para a determinação de concentração de cloridrato de propranolol

	Nº freqüências selecionadas	VL	Cabilbração		Validação		Predição	
			R^2_{cal}	RMSECV (mg de cloridrato de propranolol por mg de amostra)	R^2_{val}	RMSEV (mg de cloridrato de propranolol por mg de amostra)	R^2_{pred}	RMSEP (mg de cloridrato de propranolol por mg de amostra)
PLS	3351	5	0,9371	0,0317	0,7571	0,0554	0,9191	0,0745

Fonte: elaborado pelo autor

4.2.2 Resultados obtidos aplicando o iPLS

Foi executado o iPLS configurando as divisões do espectro em 25, 50 e 100 intervalos. Os resultados alcançados com essas diferentes configurações são demonstrados e comparados na Tabela 8.

Tabela 8 - Resultados dos modelos de regressão obtidos através do método iPLS para a determinação de concentração de cloridrato de propranolol, dividindo o espectro em 25, 50 e 100 intervalos

	Nº freqüências selecionadas	**VL**	**Cabilbração**		**Validação**		**Predição**	
			R^2_{cal}	**RMSECV** (mg de cloridrato de propranolol por mg de amostra)	R^2_{val}	**RMSEV** (mg de cloridrato de propranolol por mg de amostra)	R^2_{pred}	**RMSEP** (mg de cloridrato de propranolol por mg de amostra)
iPLS 25	134	3	0,9110	0,0367	-0,2281	0,5190	0,8821	0,1863
iPLS 50	**67**	**2**	**0,7942**	**0,0544**	**0,4145**	**0,0747**	**0,7240**	**0,0544**
iPLS 100	34	1	0,7897	0,0545	-0,1271	0,0701	0,9637	0,0618

Fonte: elaborado pelo autor

Observa-se na Tabela 8 que somente os iPLS com os espectros divididos em 50 e 100 intervalos resultam em bons modelos de calibração. A Figura 31 demonstra os erros de validação cruzada utilizando o método iPLS dividindo o espectro em 50 intervalos, onde a linha pontilhada representa o erro do modelo que utiliza todo o espectro e as barras representam os erros dos modelos construídos para cada intervalo individualmente.

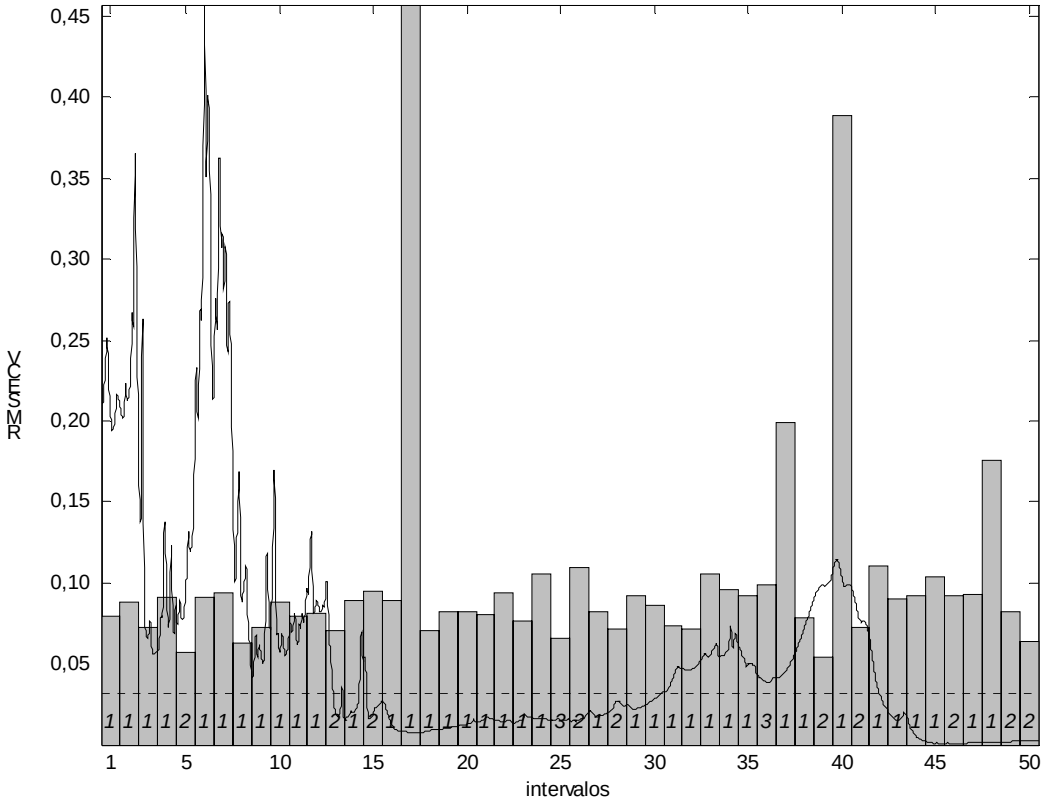


Figura 31 – Gráfico dos erros do modelo iPLS, dividindo o espectro de amostras de cloridrato de propranolol em 50 intervalos

Fonte: elaborado pelo autor

A Figura 32 apresenta o espectro, ressaltando a área selecionada pela aplicação do método iPLS dividindo o espectro em 50 intervalos, referente ao intervalo 39. A região selecionada é característica do estiramento N-H de amina secundária alifática.

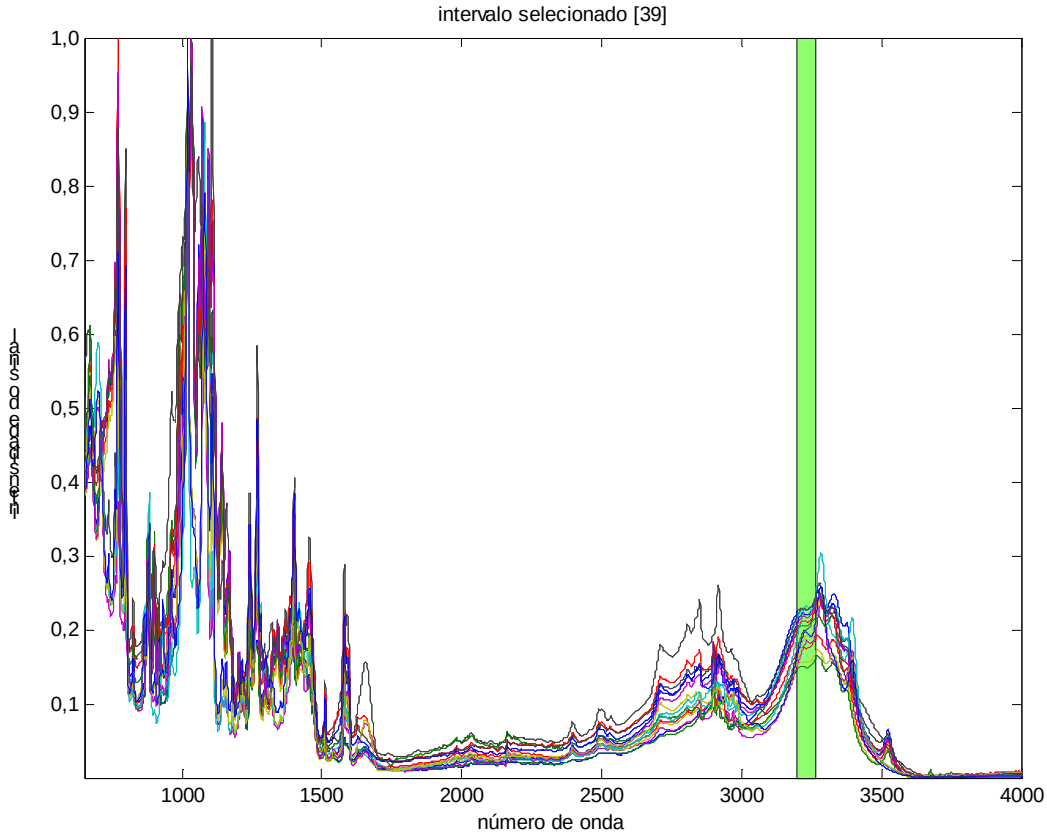


Figura 32 – Espectro de amostras de cloridrato de propranolol, ressaltando a região selecionada pelo método iPLS

Fonte: elaborado pelo autor

O gráfico da regressão das amostras de predição sobre este modelo iPLS, juntamente com o coeficiente de correlação e o RMSEP é ilustrado na Figura 33. Observa-se que o modelo não prevê de forma adequada as amostras 3, 4 e 5.

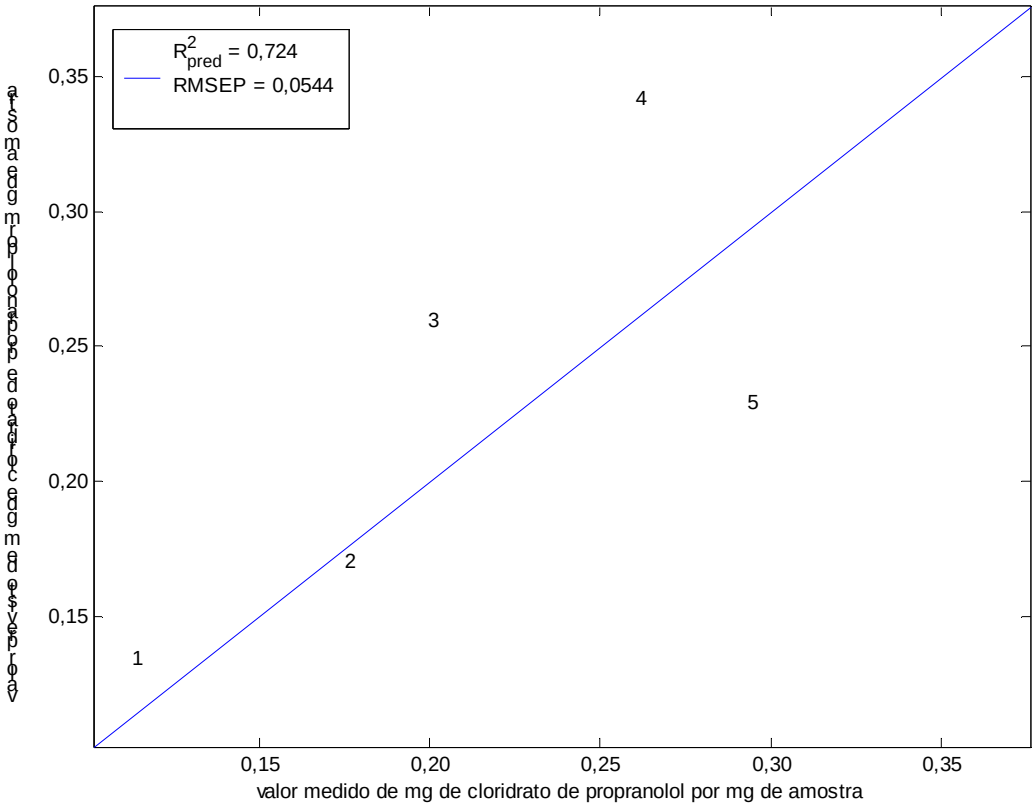


Figura 33 – Predição de amostras de cloridrato de propranolol sobre o modelo gerado pelo método iPLS

Fonte: elaborado pelo autor

4.2.3 Resultados obtidos aplicando o GA-iPLS *out*

Executou-se o GA-iPLS *out*, dividindo o espectro em 25, 50 e 100 intervalos, utilizando a *fitness* composta do erro de validação cruzada e de validação, como referido no item 3.2.2.2. A Tabela 9 apresenta os resultados obtidos através deste processo.

Tabela 9 - Resultados da aplicação do GA-iPLS *out* para a determinação de concentração de cloridrato de propranolol, dividindo o espectro em 25, 50 e 100 intervalos

		Nº freqüências selecionadas	**VL**	**Cabilbração**		**Validação**		**Predição**	
				R^2_{cal}	**RMSECV** (mg de cloridrato de propranolol por mg de amostra)	**R^2_{val}**	**RMSEV** (mg de cloridrato de propranolol por mg de amostra)	**R^2_{pred}**	**RMSEP** (mg de cloridrato de propranolol por mg de amostra)
25 intervalos	1ª execução	670	4	0,9668	0,0227	0,9934	0,0151	0,9888	0,0172
	2ª execução	**670**	**4**	**0,9673**	**0,0226**	**0,9996**	**0,0155**	**0,9919**	**0,0167**
	3ª execução	939	4	0,9668	0,0228	0,9962	0,0179	0,9962	0,0165
50 intervalos	1ª execução	1474	4	0,9719	0,0209	0,9954	0,0180	0,9961	0,0198
	2ª execução	**1608**	**5**	**0,9630**	**0,0239**	**0,9970**	**0,0165**	**0,9868**	**0,0192**
	3ª execução	1072	5	0,9677	0,0224	0,9793	0,0174	0,9952	0,0199
100 intervalos	**1ª execução**	**1505**	**5**	**0,9734**	**0,0207**	**0,9960**	**0,0152**	**0,9944**	**0,0168**
	2ª execução	1437	5	0,9716	0,0213	0,9985	0,0127	0,9825	0,0202
	3ª execução	1439	5	0,9691	0,0219	0,9935	0,0195	0,9885	0,0178

Fonte: elaborado pelo autor

As figuras 34, 35 e 36 apresentam um comparativo entre as evoluções das diferentes execuções do GA-iPLS *out* quando o espectro é dividido em 25, 50 e 100 intervalos, respectivamente. De uma forma geral a evolução da *fitness* para as replicatas em cada caso foi equivalente, excetuando-se a 2ª execução do GA-iPLS *out* para 50 intervalos.

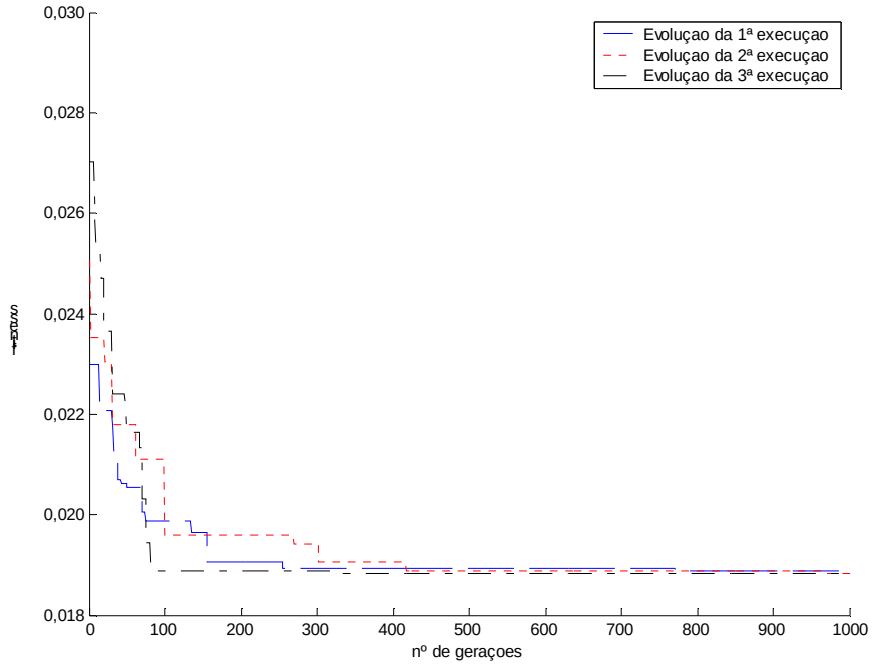


Figura 34 – Evoluções das três execuções do GA-iPLS *out* para a determinação de concentração de cloridrato de propranolol, dividindo o espectro em 25 intervalos

Fonte: elaborado pelo autor

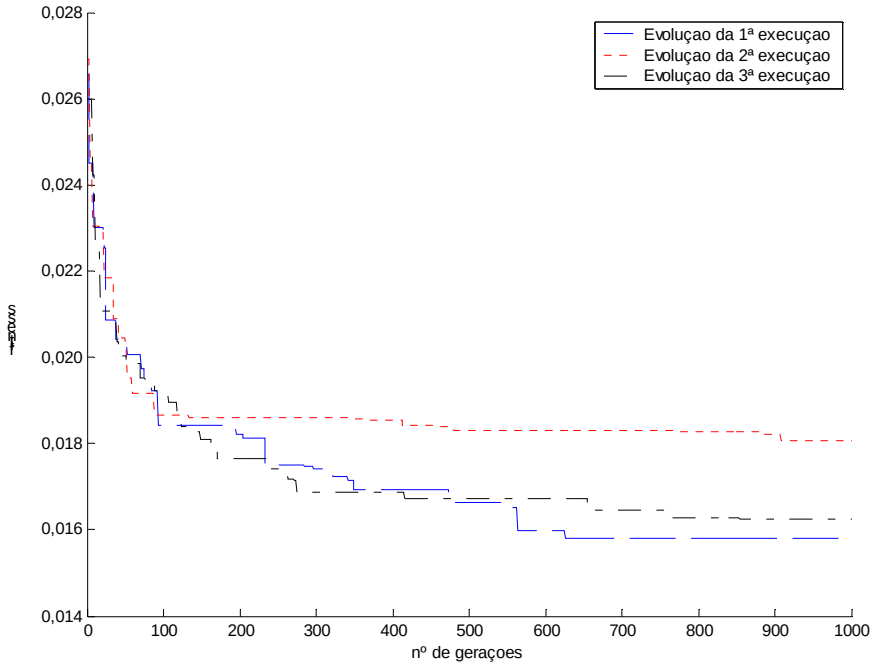


Figura 35 – Evoluções das três execuções do GA-iPLS *out* para a determinação de concentração de cloridrato de propranolol, dividindo o espectro em 50 intervalos

Fonte: elaborado pelo autor

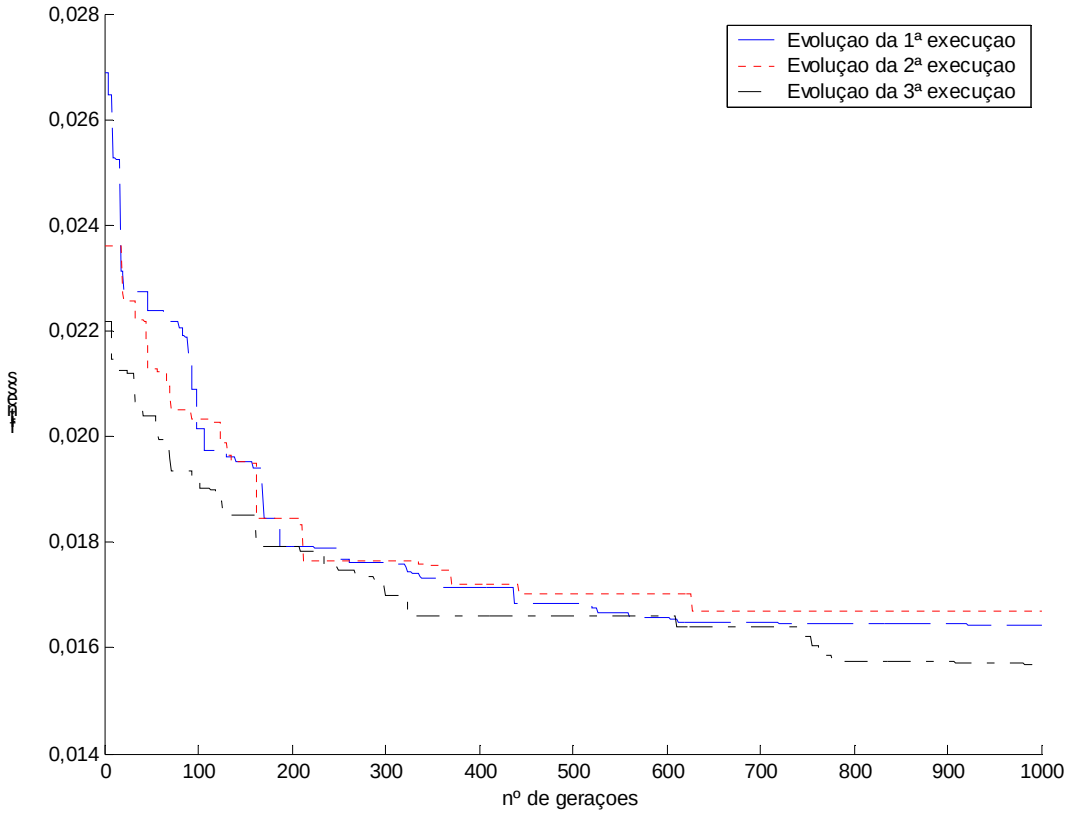


Figura 36 – Evoluções das três execuções do GA-iPLS *out* para a determinação de concentração de cloridrato de propranolol, dividindo o espectro em 100 intervalos

Fonte: elaborado pelo autor

A Figura 37 ilustra as regiões selecionadas pelo GA-iPLS *out* que alcançou o melhor resultado e a Figura 38 apresenta a regressão das amostras de predição para esta mesma solução. Neste caso observa-se uma sensível melhora na habilidade de predição das amostras. Dentre os sinais selecionados pode-se destacar a região compreendida entre 3300 - 3500 cm^{-1} referente ao estiramento do grupo NH de amina secundária alifática e região compreendida entre 1230 – 1270 cm^{-1} referente à deformação assimétrica =C-O-C de alquil aril éter (BARBOSA, 2007).

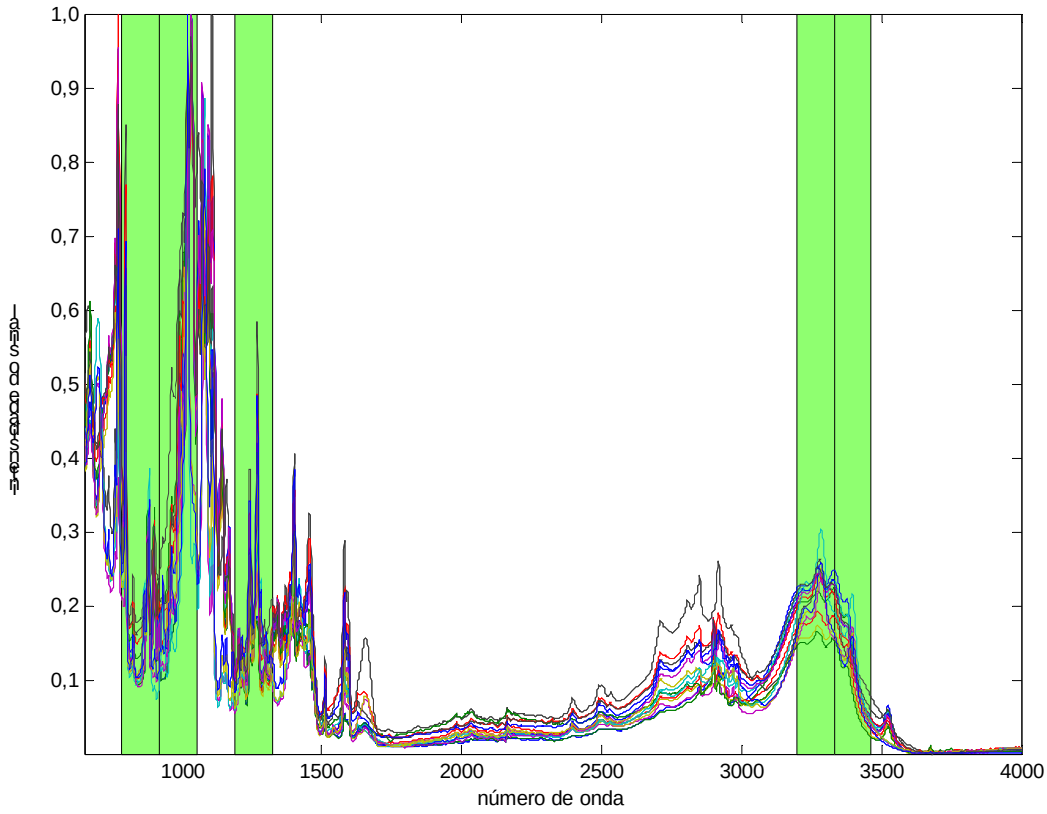


Figura 37 – Espectro de amostras de cloridrato de propranolol, ressaltando as regiões selecionadas pelo método GA-iPLS *out*, para o espectro dividido em 25 intervalos

Fonte: elaborado pelo autor

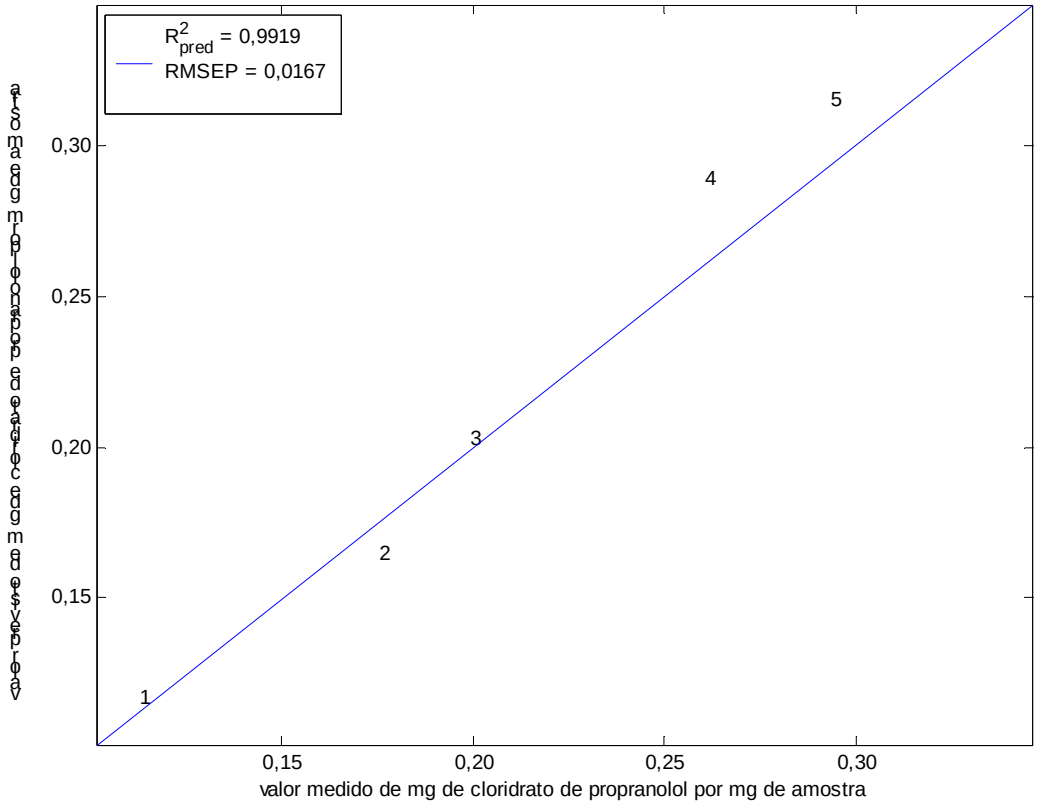


Figura 38 – Predição de amostras de cloridrato de propranolol sobre o modelo gerado pelo método GA-iPLS *out* dividindo o espectro em 25 intervalos

Fonte: elaborado pelo autor

4.2.4 Resultados obtidos pelo GA-iPLS *in*

A partir das soluções alcançadas anteriormente com o GA-iPLS *out*, é aplicado o GA-iPLS *in* para otimizar ainda mais o modelo de regressão multivariado, refinando e encontrando modelos ainda mais robustos. A Tabela 10 apresenta as soluções alcançadas utilizando esta implementação.

Tabela 10 - Resultados da aplicação do GA-iPLS *in* para a determinação de concentração de cloridrato de propranolol, refinando as melhores soluções encontradas pelo GA-iPLS *out*

	Nº freqüências selecionadas	**VL**	**Calibração**		**Validação**		**Predição**	
			R^2_{cal}	**RMSECV** (mg de cloridrato de propranolol por mg de amostra)	**R^2_{val}**	**RMSEV** (mg de cloridrato de propranolol por mg de amostra)	**R^2_{pred}**	**RMSEP** (mg de cloridrato de propranolol por mg de amostra)
Solução out 25 intervalos	670	4	0,9673	0,0226	0,9996	0,0155	0,9919	0,0167
1ª execução	324	4	0,9803	0,0175	0,9957	0,0134	0,9942	0,0151
2ª execução	**334**	**4**	**0,9773**	**0,0189**	**0,9966**	**0,0143**	**0,9930**	**0,0130**
3ª execução	344	4	0,9778	0,0187	0,9972	0,0158	0,9932	0,0144
Solução out 50 intervalos	1608	5	0,9630	0,0239	0,9970	0,0165	0,9868	0,0192
1ª execução	791	5	0,9660	0,0229	0,9952	0,0157	0,9744	0,0210
2ª execução	832	5	0,9634	0,0238	0,9935	0,0147	0,9852	0,0192
3ª execução	**443**	**5**	**0,9686**	**0,0223**	**0,9735**	**0,0218**	**0,9724**	**0,0171**
Solução out 100 intervalos	1505	5	0,9734	0,0207	0,9960	0,0152	0,9944	0,0168
1ª execução	733	5	0,9748	0,0198	0,9965	0,0123	0,9936	0,0166
2ª execução	731	6	0,9745	0,0199	0,9955	0,0122	0,9904	0,0154
3ª execução	**717**	**5**	**0,9750**	**0,0199**	**0,9908**	**0,0137**	**0,9937**	**0,0153**

Fonte: elaborado pelo autor

As figuras 39, 40 e 41 mostram um comparativo entre as evoluções das três diferentes execuções, para cada configuração do GA-iPLS *in*, sobre a melhor resposta do GA-iPLS *out* com o espectro dividido em 25, 50 e 100 intervalos respectivamente. Todas as repetições apresentam evolução semelhante para 1000 iterações, sendo que para 25 intervalos são obtidos modelos com melhor capacidade de predição das amostras externas.

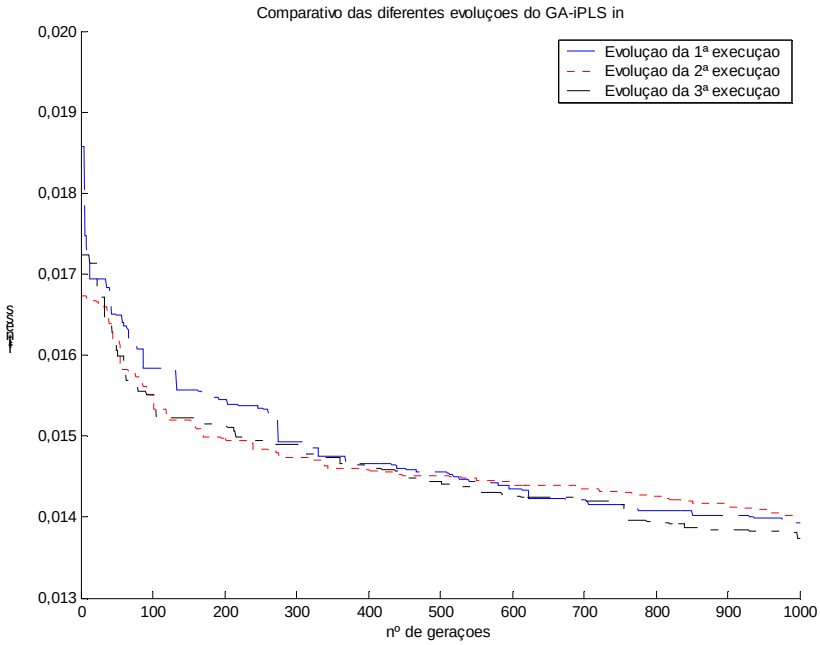


Figura 39 – Evoluções das três execuções do GA-iPLS *in* para a determinação de concentração de cloridrato de propranolol, sobre a melhor resposta do GA-iPLS out com o espectro dividido em 25 intervalos

Fonte: elaborado pelo autor

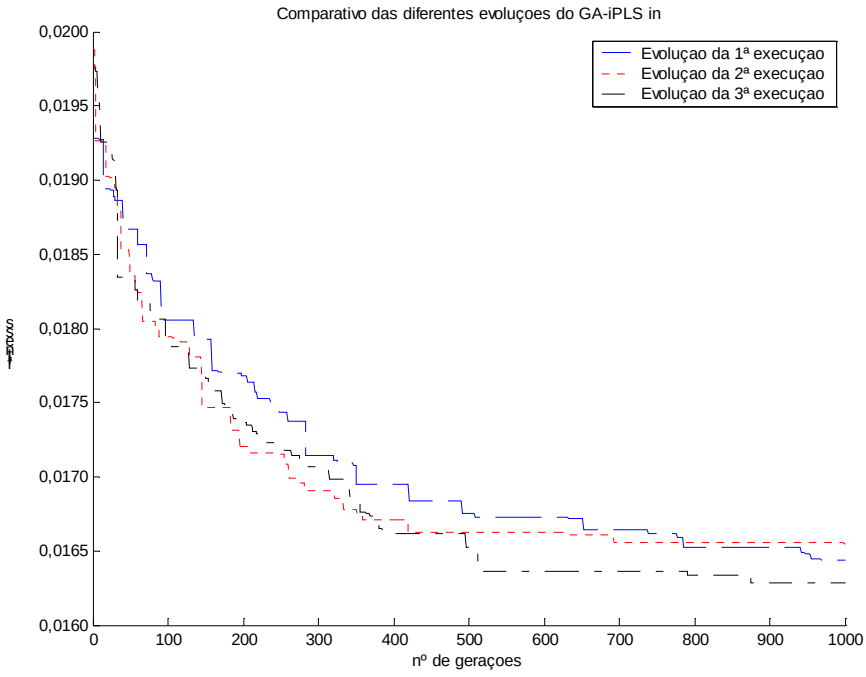


Figura 40 – Evoluções das três execuções do GA-iPLS *in* para a determinação de concentração de cloridrato de propranolol, sobre a melhor resposta do GA-iPLS out com o espectro dividido em 50 intervalos

Fonte: elaborado pelo autor

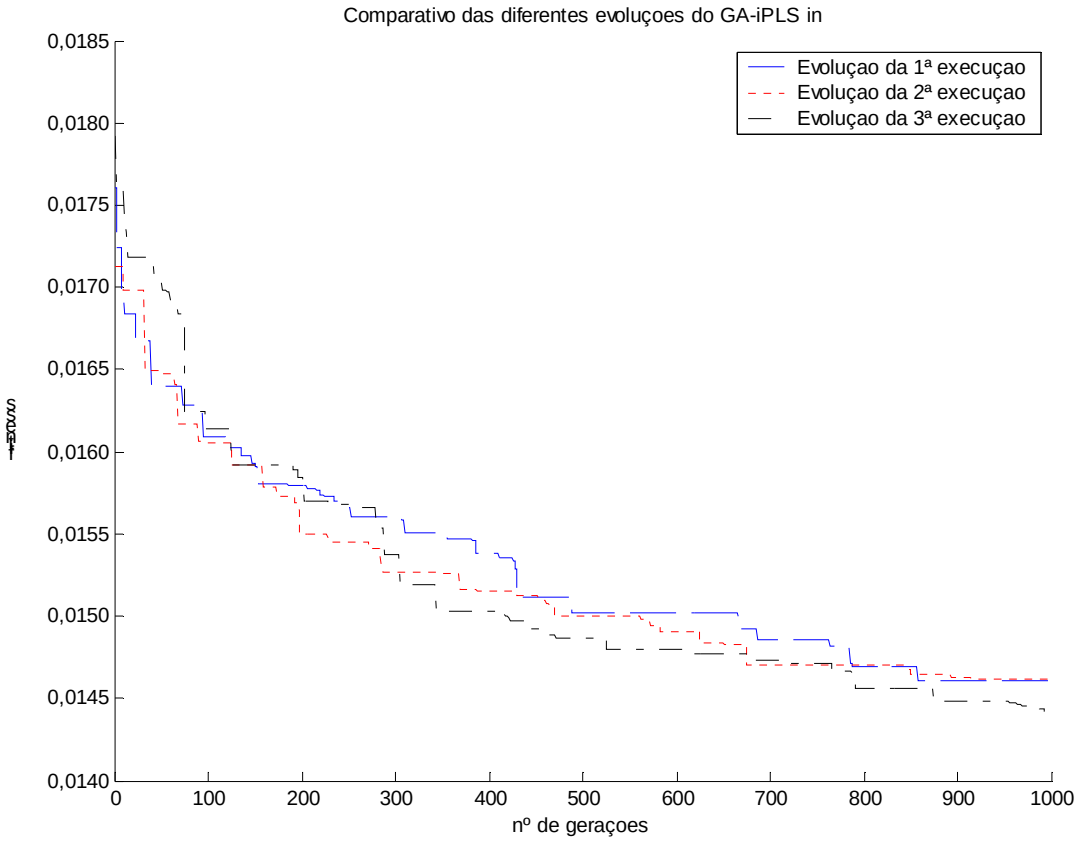


Figura 41 – Evoluções das três execuções do GA-iPLS *in* para a determinação de concentração de cloridrato de propranolol, sobre a melhor resposta do GA-iPLS out com o espectro dividido em 100 intervalos

Fonte: elaborado pelo autor

A Figura 42 ilustra o espectro das amostras de cloridrato de propranolol, ressaltando os comprimentos de onda selecionados pelo método GA-iPLS *in* dentro das regiões selecionadas pelo GA-iPLS *out* com o espectro dividido em 25 intervalos e com RMSEP de 0,0130 miligramas de cloridrato de propranolol por miligrama de amostra. A Figura 43 demonstra a regressão das amostras de predição sobre este modelo, informando também o coeficiente de regressão e o RMSEP das amostras de predição. Neste caso houve uma redução de aproximadamente 90% no número de variáveis espectrais utilizadas na criação do modelo de regressão multivariado encontrado como resposta pelo GA-iPLS.

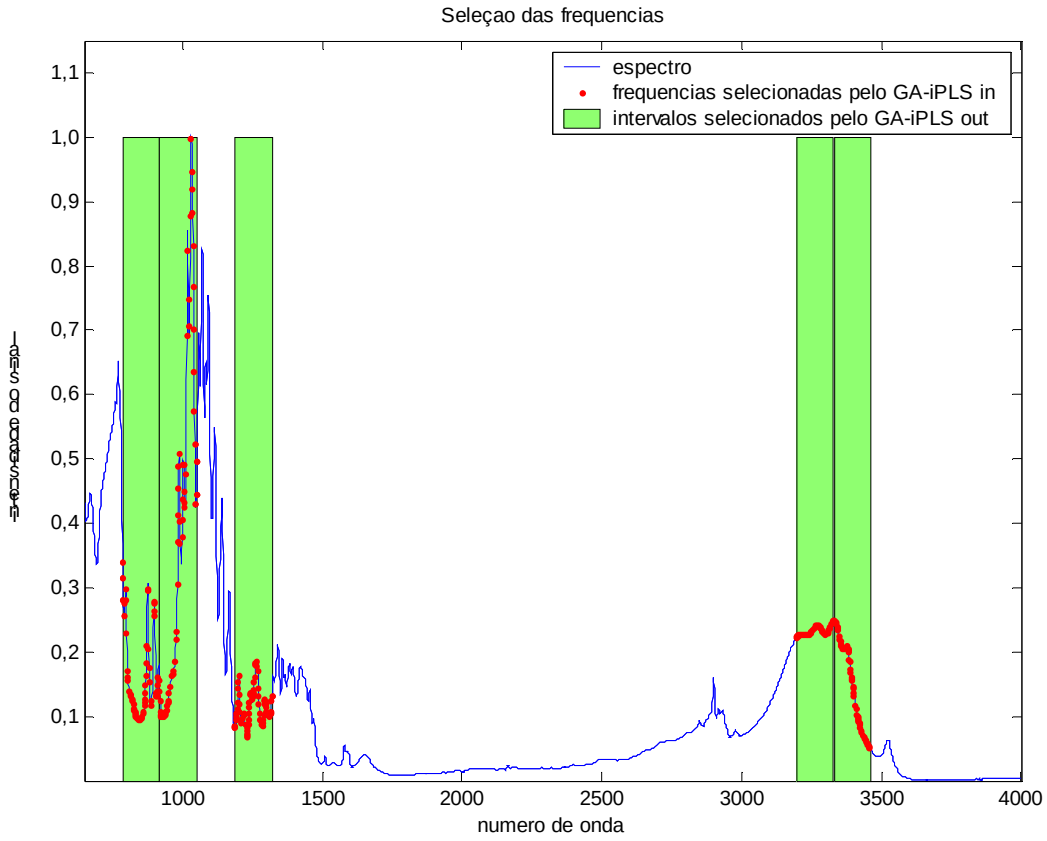


Figura 42 – Espectro de amostras de cloridrato de propranolol, ressaltando as regiões selecionadas pelo método GA-iPLS *in*, sobre a solução encontrada pelo GA-iPLS *out* com o espectro dividido em 25 intervalos

Fonte: elaborado pelo autor

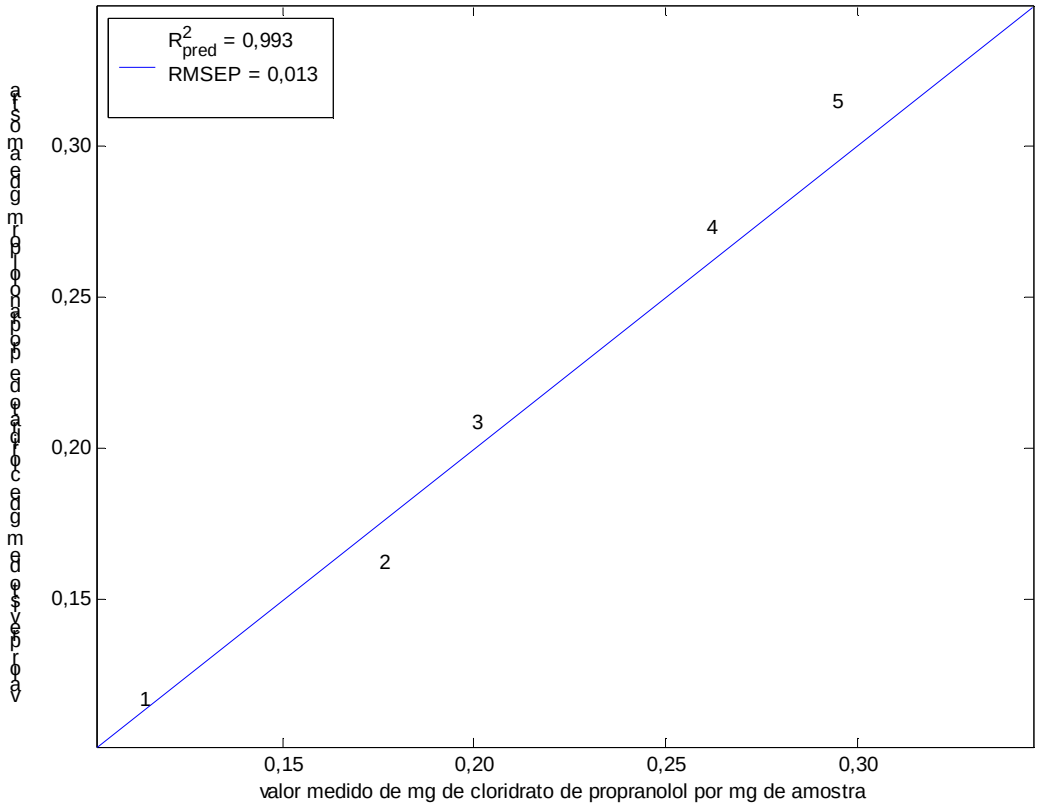


Figura 43 – Predição de amostras de cloridrato de propranolol sobre o modelo gerado pelo método GA-iPLS *in* gerado a partir da solução obtida pelo GA-iPLS *out*, dividindo o espectro em 25 intervalos

Fonte: elaborado pelo autor

A Tabela 11 apresenta o valor de miligramas de cloridrato de propranolol por miligrama de cada amostra de predição ($\mathbf{y_{ref}}$) e o valor previsto ($\mathbf{y_{pred}}$), o erro percentual de cada amostra (**%**) e a média dos erros percentuais para cada um dos métodos empregados neste estudo.

Tabela 11 - Valores medidos e previstos e os erros percentuais das amostras de cloridrato de propranolol

		PLS		iPLS		GA-iPLS out		GA-iPLS in	
	y_{ref} (mg de cloridrato de propranolol por mg de amostra)	y_{pred} (mg de cloridrato de propranolol por mg de amostra)	%	y_{pred} (mg de cloridrato de propranolol por mg de amostra)	%	y_{pred} (mg de cloridrato de propranolol por mg de amostra)	%	y_{pred} (mg de cloridrato de propranolol por mg de amostra)	%
1	0,1128	0,1388	23,05	0,1339	18,71	0,1167	3,46	0,1168	3,55
2	0,1753	0,1974	12,61	0,1700	3,02	0,1646	6,10	0,1622	7,47
3	0,1994	0,1905	4,46	0,2594	30,09	0,2030	1,81	0,2085	4,56
4	0,2605	0,4054	55,62	0,3419	31,25	0,2886	10,79	0,2725	4,61
5	0,2935	0,3674	25,18	0,2293	21,87	0,3149	7,29	0,3143	7,09
Erro percentual médio:			**24,18**		**20,99**		**5,89**		**5,46**

Fonte: elaborado pelo autor com base nos resultados obtidos

Conforme evidenciado, excelentes resultados para a determinação de cloridrato de propranolol foram encontrados utilizando-se o GA-iPLS *out* e o GA-iPLS *in* em comparação com os resultados alcançados pelo PLS e iPLS, demonstrando que houve uma otimização dos modelos de regressão multivariados para este problema. Embora o erro médio percentual das amostras de predição obtido utilizando-se o método GA-iPLS *in* não tenha diminuído de forma tão expressiva em comparação ao que foi alcançado através do GA-iPLS *out*, o RMSEP foi reduzido e um decremento considerável no número de variáveis espectrais envolvidas, próxima a 50%, indicando que o GA-iPLS *in* é capaz de refinar as soluções e encontrar modelos mais robustos.

5 CONCLUSÃO

Este trabalho teve o propósito de pesquisar e desenvolver métodos de otimização de modelos de regressão multivariados, sendo este objetivo alcançado através da construção de um modelo híbrido utilizando algoritmos genéticos (método heurístico), juntamente com o método de regressão de mínimos quadrados parciais por intervalo (iPLS) (método determinístico), para realizar a seleção das variáveis utilizadas na produção de modelos mais preditivos.

A metodologia utilizada nesta dissertação - a utilização de algoritmos genéticos para seleção de variáveis combinado ao método iPLS para a otimização de modelos de regressão multivariados - demonstrou-se eficiente no cumprimento do seu objetivo, obtendo bons resultados tanto para a determinação de OH em polióis de óleo de soja quanto na determinação da concentração de cloridrato de propranolol. Observou-se que em ambos os casos que aplicando somente o método iPLS os modelos obtidos são sempre inferiores aos resultados dos modelos híbridos.

Através de uma análise dos resultados obtidos pela otimização utilizando o GA-iPLS *out*, podemos concluir que este algoritmo auxiliou no processo de seleção de variáveis, encontrando modelos mais robustos, reduzindo sua complexidade através da redução do número de variáveis e apresentando menores erros de predição. Estes melhores resultados foram assim obtidos em função do GA-iPLS permitir a combinação de diferentes regiões do espectro resultando num sinergismo.

Observando os resultados obtidos pelo refinamento proporcionado através do GA-iPLS *in*, podemos perceber uma diminuição significativa do número de variáveis envolvidas na construção do modelo de regressão multivariado, diminuindo também os erros deste modelo. Para este caso destacamos que a finalidade do GA-iPLS *in* é minimizar o antagonismo, isto é, eliminar variáveis não significativas que se encontram dentro dos intervalos selecionados pelo GA-iPLS *out*.

Em todas as aplicações apresentadas, foi observada a seleção de regiões referentes aos sinais presentes nas estruturas químicas dos constituintes das amostras estudadas em cada problema. Este fato é importante uma vez que ratifica a utilização dessas ferramentas de otimização na construção dos modelos de calibração multivariados.

Esta conclusão pode ser confirmada atentando-se a queda de percentual dos erros em ambos os problemas estudados:

Determinação de OH de polióis de óleo de soja: em comparação com os resultados encontrados usando-se o método iPLS, o GA-iPLS *out* apontou uma queda de 8,6% no RMSEP e um acréscimo de 4,18% no erro médio percentual das amostras de predição. Já o GA-iPLS *in* alcançou uma redução de 14,97% no RMSEP e de 15,63% no erro médio percentual das amostras de predição, em comparação ao iPLS.

Determinação de cloridrato de propranolol: em comparação com os resultados encontrados usando-se o método iPLS, o GA-iPLS *out* apontou uma queda de 69,3% no RMSEP e de 71,94% no erro médio percentual das amostras de predição. Já o GA-iPLS *in* alcançou um diminuição de 76,1% no RMSEP e de 73,99% no erro médio percentual das amostras de predição, em comparação ao iPLS

Com base neste estudo, podemos concluir que um método que seja capaz de selecionar as variáveis espectrais de forma eficiente pode auxiliar na redução da complexidade dos modelos e torná-los mais precisos em relação à propriedade que se almeja prever.

Com uma boa capacidade preditiva destas técnicas e aliado ao baixo custo e rapidez e possibilidade de análises não destrutivas quando da utilização de espectroscopia no infravermelho para este fim, vislumbra-se um ganho em termos de custo e de tempo de análise em indústrias que utilizam desses meios para controle de qualidade de seus produtos ou que necessitam de algum tipo de análise química passível de ser realizada com estes instrumentos.

Uma característica desejável e de destaque na ferramenta desenvolvida nesta pesquisa é a automaticidade do processo de otimização dos modelos de regressão multivariados, juntamente com um conjunto de funções que auxiliam na visualização dos resultados alcançados e na comparação entre estes resultados e os obtidos pela metodologia clássica,

indicando as principais características dos modelos encontrados e gerando gráficos que facilitam a análise dos resultados.

Desta forma o algoritmo desenvolvido é capaz de ser executado em qualquer conjunto de dados espectrais sem a necessidade de alteração do programa, bastando apenas formatar os dados de maneira com que o programa os reconheça.

Para evitar possíveis erros nesta formatação, foi criado também um algoritmo para auxiliar na correta estruturação dos dados, proporcionando maior organização e facilitando a alteração de todos os parâmetros do GA e de criação dos modelos de regressão multivariados, que se encontram em um único arquivo com estruturação própria.

No decorrer do trabalho, embora bons resultados tenham sido obtidos para ambos os conjuntos de dados, estes resultados podem variar dependendo do problema estudado. Tendo isto em vista, é desejável a aplicação dos métodos aqui estudados para a otimização de modelos de regressão multivariados em outros problemas. Também pode ser interessante o estudo de outros métodos para realizar a seleção de variáveis e verificar se pode ser mais efetivo para este fim, como a busca tabu e o enxame de partículas.

REFERÊNCIAS

BARBOSA, L. C. A. **Espectroscopia no Infravermelho na caracterização de compostos orgânicos**. 1 ed. Viçosa: Editora UFV. 2007.

BORIN, A. POPPI, R. J. **Application of Mid Infrared Spectroscopy and iPLS for the Quantification of Contaminants in Lubricating Oil**. In: Vibrational Spectroscopy, n.37, p.27-32, 2005.

BORIN, A. POPPI, R. J. **Multivariate Quality Control of Lubricating Oils Using Fourier Transform Infrared Spectroscopy**. In: Journal of the Brazilian Chemical Society, vol. 15, n.4, p.570-576, 2004.

CARVALHO, C. W. et al. **Determinação de Fármacos Anti-Hipertensivos por Reflexão no Infravermelho, Regressão Multivariada e Algoritmos Genéticos**. In: Tecno-Lógica. Santa Cruz do Sul, v.6, n.1, p.9-27, jan./jun. 2002.

COSTA FILHO, P. A. POPI, J. **Aplicação de Algoritmos Genéticos na Seleção de Variáveis em Espectroscopia no Infravermelho Médio. Determinação Simultânea de Glicose, Maltose e Frutose**. In: Química Nova, v.25, n.1, p. 46-52, 2002.

CHRISTY, A. A. EGEBERG, P. K. **Quantitative Determination of Saturated and Unsaturated Fatty Acids in Edible Oils by Infrared Spectroscopy and Chemometrics**. In: Chemometrics and Intelligent Laboratory Systems, vol. 82, n.1-2, p.130-136, 2006.

FERRÃO, M. F. **Técnicas de Reflexão no Infravermelho Aplicadas na Análise de Alimentos**. In: Tecno-Lógica. Santa Cruz do Sul, v.5, n.1, p.63-85, jan./jun. 2001.

FERRÃO, M. F. et al. **Determinação Simultânea dos Teores de Cinza e Proteína em Farinha de Trigo Empregando NIRR-PLS e DRIFT-PLS**. In: Ciência e Tecnologia de Alimentos, Campinas, v.24, n.3, p.333-340, jul./set. 2004.

FERREIRA, M. M. C. et al. **Quimiometria I: calibração multivariada, um tutorial**. In: Química Nova, v.22, n.5, São Paulo, set./out. 1999. ISSN 0100-4042

FERREIRA, M. M. C. MONTANARI, C. A. GLAUDIO, A. C. **Seleção de Variáveis em QSAR**. In: Química Nova, v.25, n.3, p.439-448, 2002.

FURTADO, J. C. et al.**Otimização Via Algoritmo Genético e Busca Tabu na Determinação de Proteína em Farinha de Trigo por Reflexão Difusa no Infravermelho**. In: Tecno-Lógica. Santa Cruz do Sul, v.6, n.2, p.41-71, jul./dez. 2002.

GOICOECHEA, H. C. OLIVIERI, A. C. **A New Family of Genetic Algorithms for Wavelength Interval Selection in Multivariate Analytical**. In: Journal of Chemometrics, n.17, p.338-345, 2003.

KONZEN, P. H. A. *et al*. **Otimização de Métodos de Controle de Qualidade de Fármacos Usando Algoritmo Genético e Busca Tabu**. In: Pesquisa Operacional, Vol. 23, n.1, p.189-207, 2003.

LEARDI, R. **Application of Genetic Algorithm-PLS for Feature Selection in Spectral Data Sets**. In: Journal of Chemometrics, n.14, p.643-655, 2000.

LEARDI, R. NØRGAARD, L. **Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions**. In: Journal of Chemometrics, n.18, p. 486-497, 2004.

MITCHELL, M. **An Introduction to Genetic Algorithms**. Massachusets. MIT Press, 1996.

MORGANO, M. A. et al. **Determinação Simultânea dos Teores de Cafeína, Trigonelina e Ácido Clorogênico em Amostras de Café Cru por Análise Multivariada (PLS) em Dados de Espectroscopia Difusa no Infravermelho Próximo**. In: II SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, p.1502-1510, 2001.

NØRGAARD, L. et al. **Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy**. In: Applied Spectroscopy, v.54, n.3, p. 413-419, 2000.

ÖJELUND, H. MADSEN, H. THYREGOD, P. **Calibration with Empirically Mean Subset**. In: Applied Spectroscopy, v.56, n.7, p. 887-896, 2002.

OLIVEIRA, F. C. C. et al. **Escolha da Faixa Espectral no Uso Combinado de Métodos Espectroscópicos e Quimiométricos**. In: Química Nova, v.27, n.7, p.218-225, 2004.

YEPES, I. Projeto ISIS: Sistemas Inteligentes. Uma incursão aos Algoritmos Genéticos. Disponível em: <http://www.geocities.com/igoryepes/index.htm>. Acesso em: 20 de setembro de 2006.

SABIN, G. P. CARVALHO, C. S. **Dimensionamento de Redes de Abastecimento de Água Utilizando Algoritmos Genéticos – Projeto AGua**. 2005. Monografia (Curso de Engenharia de Computação) - Fundação Universidade Federal do Rio Grande, Rio Grande.

SABIN, J. G. **Determinação de Princípios Ativos Presentes em Fármacos de Ação Antidepressiva Utilizando Espectroscopia no Infravermelho com Transformada de Fourier**. 2002. Monografia (Curso de Química Industrial) - Universidade de Santa Cruz do Sul, Santa Cruz do Sul.

SKOOG, D. A. HOLLER, F. J. NIEMAN, T. A. **Princípios de Análise Instrumental**. 5 ed. Porto Alegre: Bookman. 2002.

SMIDERLE, A. **Técnicas da Pesquisa Operacional Aplicadas - um Problema de Cobertura de Arcos**. 2001. 131 f. Dissertação (Programa de Pós-Graduação em Métodos Numéricos em Engenharia - Mestrado) - Universidade Federal do Paraná, Curitiba.

WETZEL, D. L. **Near-infrared reflectance analysis. Sleeper among spectroscopic techniques**. In: Anal. Chem., n 55, p1165-1176, 1983.

ZENI, D. **Determinação de Cloridrato de Propranolol em Medicamentos por Espectroscopia no Infravermelho com Calibração Multivariada (PLS)**. 2005. Dissertação (Programa de Pós-Graduação em Química) - Universidade Federal de Santa Maria, Sana Maria.

ANEXOS

ANEXO A – artigo aprovado no XII ICIEOM e publicado em sua íntegra nos anais deste evento

Genetic algorithm for optimization of the interval partial least-squares regression using attenuated total reflectance infrared data

Gustavo Post Sabin (UNISC) gustavosabin@yahoo.com.br
Marco Flôres Ferrão (UNISC) ferrao@unisc.br
João Carlos Furtado (UNISC) jcarlosf@unisc.br
Simone da Câmara Godoy (UFRGS) godoysc@iq.ufrgs.br
Annelise Engel Gerbase (UFRGS) agerbase@ufrgs.br
Cesar Liberato Petzhold (UFRGS) petzhold@iq.ufrgs.br

Summary

This paper presents the use of genetic algorithm (GA) for optimization partial least-squares regression models for quantitative determination of hydroxyl value (OHV) of hydroxylated soybean oils by Fourier transform infrared spectroscopy with horizontal attenuated total reflection accessory (FT-IR/HATR). A full-spectrum partial least-squares (PLS) calibration model for the prediction of OHV was developed using the range 1805.1 to 649.9 cm-1, covering an analytical range of 23.66-195.04 mg of KOH/g per sample . Sixty interval partial least-squares (iPLS) models were developed in the same spectral region were compared using root mean squares error of cross-validation (RMSECV). It is possible verify the superior ability of prediction of the combine intervals when GA-iPLS was used. It was obtained a determination coefficient of 0.9932 and RMSEP of 5.65 mg of KOH/g. The model using some intervals presents better predictions with lower errors in relation to the full model. This study shows that it is possible to optimized OHV determination of hydroxylated soybean oil using GA-iPLS by FT-IR/HATR spectra.
Keywords: genetic algorithm, attenuated total reflectance, ipls.

1. Introduction

The environmental and sustainability aspects of using oleochemical polyols are of great importance to the polyurethane industry considering the broad range of applications of these materials. Oleochemical polyols can be used in the production of VOC-free, two-components polyurethane coatings and floorings, adhesives and thermoplastic polyurethanes (GUO, JAVNI & PETROVIC, 2000; PETROVIC, GUO & ZHANG, 2000; HU *et al*., 2002; GUO *et al*., 2002). For polyurethane preparation, it is important to know the final hydroxyl value of the soybean polyol. Usually, the OHV is determined by titration methods such as the American Oil Chemists' Society (AOCS, 1997) hydroxyl value determination (AOCS Cd 13-60) used in this work. The hydroxyl value is expressed in mg of KOH per g of oil. This method is reliable and reproducible if carried out under standardized conditions, but it is time-consuming, labor-intensive, reasonably sensitive, largely dependent on the skills of the analyst, uses large amounts of sample and reagents, and some of them (pyridine, acetic anhydride) are hazardous and difficult to dispose off.

Similar problems were also observed in other chemical analyses of fats and oils based on titration methods. Therefore, spectroscopic methods are being increasingly used to replace wet chemical procedures. Infrared spectroscopy is one that has found increasing use due to its low cost, shorter time of analysis, non-destructiveness, small quantities of sample, in addition to accuracy and reliability when associated with chemometric methods (PARREIRA et al.,

2002; AL-ALAWI & VAN DE VOORT, 2004). Moreover, FT-IR coupled with horizontal attenuated total reflectance (HATR) accessory simplifies many of the sample handling problems commonly associated with infrared analysis and is readily amenable to routine quality control applications (BORIN & POPPI, 2004).

Partial least-squares (PLS) regression (GELADI & KOWALSKI, 1986) is the most popular multivariate calibration technique to build prediction models using spectroscopic signal. In 1998, Spiegelman *et al.*, demonstrated that spectral region selection can significantly improve the performance of these full-spectrum calibration techniques. Specific regions (or infrared signals) are selected where colinearity is not so important, generating more stable models with superior interpretability. In practice, the optimization of the multivariate regression models is based on the identification of a subset of the complete data that will produce the lowest prediction error. For this propose have been used the genetic algorithm (GA) procedures (LEARDI, 2000; KONZEN *et al.*, 2003) and interval partial least-squares regression (NORGAARD *et al.*, 2000).

In this work were compared the performances of the interval partial least-squares (iPLS) and GA-iPLS with the method of the PLS to determine OHV of hydroxylated soybean oil using FT-IR/HATR spectra.

2. Horizontal Attenuated Total Reflectance

Basic the theoretical principles of the spectroscopy of internal reflection (IRS), had been published in the beginning of years 60 for Harrick (1960) and Fahrenfort (1961), that they describe details of the theory and they present experimental results. The phenomenon of the internal reflection is observed under certain conditions. When the radiation enters in a prism made with a high refractive index in relation to the external way (ATR crystal) the radiation will be reflected total internally. Figure 1 show the representation of the infrared radiation propagates through the internal reflectance element (IRE) (MIRABELLA, 1985).

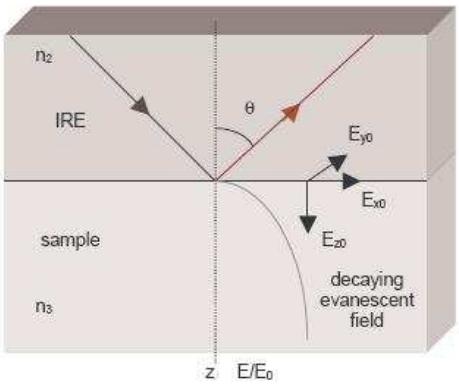


Figure 1 - Representation of the infrared radiation propagates through the IRE.

For a refractive index such that $n_2 > n_3$, applying it law of Snell, the refraction angle is imaginary for values of the angle of incidence such that satisfied equation 1.

$$\theta_i > \text{sen}^{-1}(n_3/n_2) \quad (1)$$

The angle above of which this refracted wave leaves of being real is called of critical angle (θ_c) and is express in agreement with equation 2.

$$\theta_c = \text{sen}^{-1}\left(n_3 / n_2\right) \quad (2)$$

In these conditions the beam total is reflected in the interface n_3/n_2.

The properties of the evanescent field can be observed in figure 1, where the propagation of the infrared radiation of medium 2, with refractive index n_2, it undergoes internal reflection in the interface with medium 3, refractive index n_3, when the angle of incidence exceeds the critical angle. This phenomenon of total internal reflection is most easily described for an infinite plane wave at an interface between a semi-infinite nonabsorbent media (MIRABELLA, 1985).

The depth of penetration (d_p) in medium 2 is defined as that at which the amplitude has fallen to e^{-1} of its value at the interface and is given by equation 3.

$$d_p = \frac{\lambda}{2\pi n_1 \left(\text{sen}^2\theta_i - \left(\frac{n_3}{n_2}\right)^2 \right)^{1/2}} \quad (3)$$

Where λ is the wavelength of the radiation, θ_i the angle of incidence between the radiation beam and the normal to the surface, n_2 the refractive index of the ATR element and n_3 the refractive index of the sample. It can be seen from equation 3 that the depth of penetration is directly proportional to wavelength (COMPTON & COMPTON, 1993).

The radiation happens in the sample penetrates few microns of depth. Therefore, any material that is in contact with the ATR crystal can absorb the incident radiation attenuating its intensity, resulting in the infrared spectrum. The depth of penetration also changes with the refractive index of the sample. This effect is less when the refractive index of the ATR crystal is great. The choice of the crystal can result in distortions of the band in the spectrum. When the angle of incidence will be very large that the critical angle, these resultant distortions in the spectrum are minimized.

The IRE is composed of a material with a high index of refraction, such as zinc selenide (ZnSe), thalium iodide-thalium bromide (KRS-5), germanium (Ge) or silicon (Si). A problem generally found, is the difficulty to get good reproducibility in the contact of the sample with the ATR element. This effect is observed in the variation of the intensity of the bands with the applied pressure. Increasing the pressure, the contact efficiency is increased and, consequently, the intensities of the bands increase. The area of contact between the crystal and the sample is a factor that also influences in the intensity of the bands, for a good reproducibility. For quantitative measures, all must be placed the area of the crystal in contact with the sample. Irregularities in the surface of the sample go to make it difficult a more effective contact (MIRABELLA, 1985).

Many of the published papers were devoted to the measurement of reflected light from the samples using accessories for attenuated total reflection (ATR) (COSTA FILHO & POPPI, 2002; CHRISTY, EGEBERG & OSTENSEN, 2003; BORIN & POPPI, 2005; CHRISTY & EGEBERG, 2006). Horizontal attenuated total reflectance (HATR) is a well-known technique that produces good quality and highly reproducible spectra, if good contact can be established between the sample and the internal reflectance element (IRE)(FERRÃO & DAVANZO, 2005).

3. Interval Partial Least Squares

The iPLS algorithm was been used for interval selection (wavenumber set) in spectral multivariate regression problems (NORGAARD et al., 2000). The principle of this algorithm is to split the spectra into smaller equidistant regions and, afterwards, developed PLS regression models for each of sub-intervals. Thereafter, an averange error is calculated for every sub-interval and the full-spectrum model. The region with the lowest error is chosen. An optimized region can be found by reducing or increasing it by subtracting or adding new variables, symmetrically or asymmetrically. One of the main advantages of this method is to possibility to represent a local regression model in a graphical display, focusing on a choice of better intervals and permiting a comparison among interval models and the full-spectrum model.

4. Experimental

4.1. Chemicals:

Refined soybean oil was purchase from was supplied by CBM Ind. Com. Distrib. Ltda (Cachoeirinha, RS, Brazil). Formic acid and ethyl ether were purchased from Synth. Hydrogen peroxide solution 30%, sodium chloride, sodium carbonate, sodium hydrogensulfite, sodium sulfate anhydrous were purchase from Nuclear. All chemicals are analytical grade and were used without further purification.

4.2. Calibration Standards:

Soybean polyols were synthesized following the method described below and were used as calibration standards. Depending on the time of reaction, soybean polyols with different OH functionality were obtained. The acid number (AN) and the OHV of the soybean polyols were determined by the AOCS standard method Cd 3a-63 (AOCS, 1980) and Cd 13–60 (AOCS, 1997), respectively. The OHV cover a range of 23.66-195,04 mg. KOH per g of sample.

4.3. Instrumentation:

A Nicolet Magna 550 FT-IR spectrophotometer with a 4 cm-1 resolution and 16 scans was used for the measurement of soybean polyols. The duplicate spectra were recorded by applying the soybean polyol sample on the surface of a Pike horizontal attenuated total reflectance (HATR) sample-handling accessory with ZnSe crystal.

4.4. Multivariate modeling using Interval Partial Least Squares (IPLS):

The average specters for each samples had been gotten using the two replicates obtained. Data were treated with multiplicative scattered correction (MSC) technique before further multivariate analysis. The iToolbox for MATLAB (NORGAARD et al., 2000) was used for developed iPLS multivariate models. The program was run on an IBM-compatible Intel Pentium 4 CPU 3.00 GHz and 1 Gbytes RAM microcomputer.

Calibration: FT-IR/HATR spectra of the 42 soybean polyol samples were used. The samples presenting extreme OHV values were included in the calibration set. Cross-validation following the leave-one-out procedure was performed during the validation step in order to define the optimum number of factors that should be kept in the model and to detect any outliers.

Validation: Spectra of twenty soybean polyol samples were used for the validation of the multivariate regression models. To evaluate the error of calibration model, the root mean

square error was used, calculated by equation 4:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (4)$$

where n is the number of spectra, y_i and $\hat{y}_i$ are the values determined by AOCS standard method and those predicted by PLS (or iPLS) models, respectively, in the calibration set (RMSEC) or external validation set (RMSEP).

4.5. Optimization Using Genetic Algorithm

Genetic algorithm was implemented to select the intervals of the spectrum that will be used for to get a model by iPLS method. This algorithm is called GA-iPLS. The GA-iPLS has as parameters, the number of intervals that this spectrum will be divided (α), the number of cycles that will be executed (β), the size of the population (χ), crossover rate (δ), mutation rate (ε), elite rate (ϕ) and the spectra data.

The GA-iPLS works as the steps following:

Step 1: To generate initial population

An initial population is generated randomly in accordance with the parameters α (corresponds to the size of the chromosomes) and χ (number of chromosomes in the population). Each chromosome is a binary vector where the positions that have value 1 indicate which intervals are selected and the positions that have value 0 indicate the intervals no considered for the solution, being the first component of the vector corresponding to the first interval of the spectrum, the second component of the vector corresponding to the second interval of the spectrum and thus for ahead.

Step 2: To evaluate the solutions

The solutions are evaluated in accordance with the RMSECV (fitness) gotten by PLS model using the selected intervals. How much smaller is the RMSECV, better is the solution.

Step 3: Elite

The ϕ represents the percentage of the population that is part of the elite. The more adapted solutions belong the elite and are saves to guarantee that they had continued to be part of the next generation.

Step 4: Crossover

The genes are combinations of two selected chromosomes (parents), generating two new chromosomes (descendant). The selection of the parents is made in accordance with the level of adaptation of the solutions. The more adapted solution, have more possibility of being selected for the crossover. This process is repeated some times until δ had been satisfied, what it indicates the percentage of the population that will be submitted to the crossover.

Step 5: Mutation

A chromosome is selected randomly. After that one of genes is modified randomly too. This process is repeated some times until ε had been satisfied, indicating the percentage of the population that will be submitted to the mutation process.

Step 6: To actualize population

In this stage the elite is introduced in the population, having substituted the less adapted solutions.

These steps are executed the number of times informed by β variable.

5. Results and discussion

Low-molecular-weight liquid epoxy polyol esters or ethers from vegetable oils can be employed as polyols in polyurethane formulation. Usually, hydroxyl groups have been introduced through a two-step synthesis involving firstly the epoxidation of the unsaturated sites with formic acid and hydrogen peroxide, followed by epoxy ring opening with mono or polyfunctional alcohols, amino alcohols, or acids. Depending on the reaction conditions, polyols with high OH functionality (complete reaction) or epoxy polyol esters with remaining epoxy groups (partial conversion) are obtained.

In this work, epoxy polyol esters were prepared by "one-step" synthesis using the formic acid/H_2O_2 system. The hydroxylation reaction was carried out at constant temperature, 65°C, and by increasing the reaction time it was possible to prepare soy polyols with different OH functionality, which were used as calibration standards and for external validation samples.

The spectra of the 62 samples were pre-processed by multiplicative scatter correction (MSC), aiming at correcting the baseline deviation between the spectra. The results for full-spectrum and iPLS models are presented in Figure 2 and Table 1. The number in the inferior part of each bar represents the number of latent variable for the iPLS model. The local model using 19th interval is similar to the model using all spectral range (649.9-1805.1 cm^{-1}).

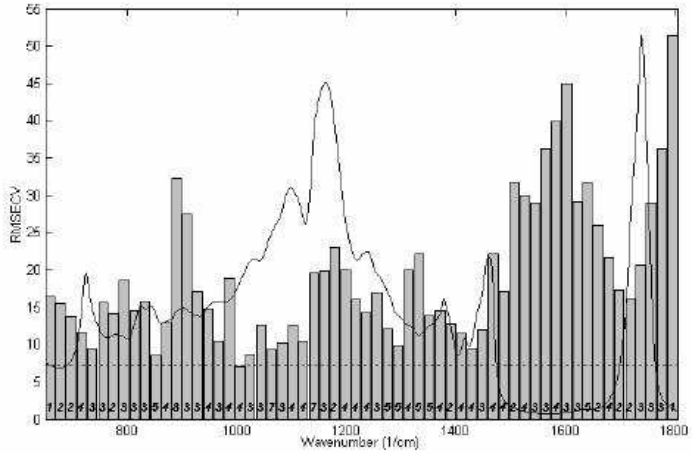


Figure 2 – Cross-validated prediction errors (RMSECV) for full-spectrum model (line) and 60 interval models (bars) for OHV determination using PLS and iPLS algorithms.

It is possible verify the superior ability of prediction of the combine intervals when GA-iPLS was used (Table 1 and Figure 3). The model using some intervals presents better predictions with lower errors in relation to the full model.

Model	Intervals	LVs*	RMSECV[a] (mg of KOH/g)	R^2_{cal}	RMSEP[b] (mg of KOH/g)
PLS	All	3	7.23	0.9894	6.81
iPLS	19	4	7.08	0.9898	7.52

GA-iPLS	1 16 17 19 20 31 32 33 35 37 41 42 43 48 49 50 51 57	3	5.78	0.9932	5.65

* LVs: Latent Variables

a RMSECV: Root Mean Square Error of Cross Validation

b RMSEP: Root Mean Square Error of Prediction (using external validation set)

Table 1 – Performance comparison results for PLS, iPLS and GA-iPLS regression models.

This results indicate that is possible to refine initial full-spectrum PLS model by eliminate spectral region that no contain information for the specific properties, in this case hydroxyl value (OHV) of hydroxylated soybean oils. The other hand, when the combine certain regions (spectroscopic signal) it is possible obtain synergic models, selected only most important spectral information for the specific properties, decreasing of the risk of over-fit.

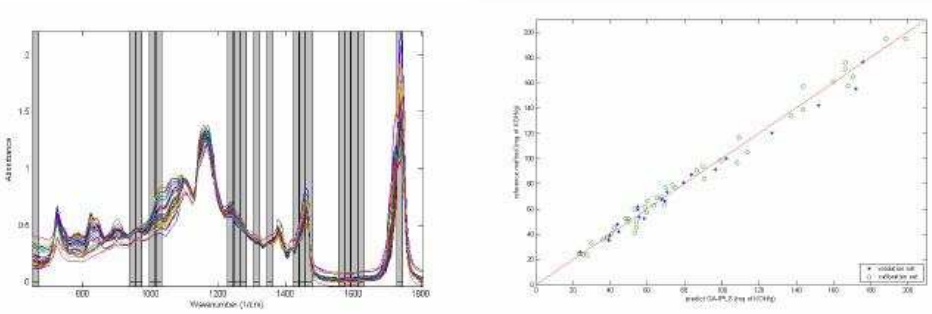


Figure 3 – Intervals selected by GA-iPLS model for OHV determination (left). Graph reference versus predict OHV values for GA-iPLS model (right).

5. Conclusions

This paper presents the use of genetic algorithm (GA) for optimization interval partial least-squares regression models (iPLS) for quantitative determination of hydroxyl value (OHV) of hydroxylated soybean oils using horizontal attenuated total reflection FT-IR-HATR data. The results show the superior ability of prediction of the combine intervals when GA-iPLS was used. All models using some intervals present better predictions with lower errors in relation to the full model (using range 1805.1 to 649.9 cm^{-1}). This study shows that it is possible to optimized OHV determination of hydroxylated soybean oil using GA-iPLS by FT-IR/HATR spectra.

Acknowledgements

This work was supported by the Brazilian National Research Council (CNPq) and the Coordination for Improvement of Higher Education Personnel (CAPES).

References

AL-ALAWI, A. & VAN DE VOORT, F.R. *New method for the quantitative determination of free fatty acids in oil by FTIR spectroscopy*. Journal of the American Oil Chemists' Society. Vol.81, p. 441-446, 2004.

AMERICAN OIL CHEMISTS' SOCIETY. *Official Methods and Recommended Practices of de American Oil Chemists' Society*, 4th edn, Champaing, 1980.

AMERICAN OIL CHEMISTS' SOCIETY. *Official Methods and Recommended Practices of de American Oil Chemists' Society*, 4th edn, Champaing, 1993, revised 1997.

BORIN, A. & POPPI, R.J. *Multivariate Quality Control of Lubricating Oils Using Fourier Transform Infrared*

Spectroscopy. Journal of the Brazilian Chemical Society. Vol.15 , n.4, p. 570-576, 2004.

BORIN, A. & POPPI, R.J. *Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil.* Vibrational Spectroscopy. Vol.37, p.27-32, 2005.

CHRISTY, A.A. & EGEBERG, P.K. *Quantitative determination of saturated and unsaturated fatty acids in edible oils by infrared spectroscopy and chemometrics.* Chemometrics and Intelligent Laboratory Systems. Vol.82, n.1-2, p.130-136, 2006.

CHRISTY, A.A.; EGEBERG, P.K. & OSTENSEN, E.T. *Simultaneous quantitative determination of isolated trans fatty acids and conjugated linoleic acids in oils and fats by chemometric analysis of the infrared profiles.* Vibrational Spectroscopy. Vol.33, n.1, p.37-48, 2003.

COMPTON, S.V. & COMPTON, D.A.C. *Optimization of Data by Internal Reflectance Spectroscopy* in Practical sampling techniques for infrared analysis – Ed. COLEMAN, P.B., Boca Raton: CRC Press, 1993.

COSTA FILHO, P.A. & POPPI, R.J. *Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio: determinação simultânea de glicose, maltose e frutose.* Química Nova. Vol.25, n.1, p.46-52, 2002.

FAHRENFORT, J. *Attenuated total reflection: A new principle for the production of useful infrared spectra of organic compounds.* Spectrochimica Acta. Vol.17, p.698-709, 1961.

FERRÃO, M.F. & DAVANZO, C.U. *Horizontal attenuated total reflection applied to simultaneous determination of ash and protein contents in commercial wheat flour.* Analytica Chimica Acta. Vol.540, p.411-415, 2005.

GELADI, P. & KOWALSKI, B.R. *Partial Least-Squares Regression: A Tutorial.* Analytica Chimica Acta. Vol.185, p.1-17, 1986.

GUO, A.; DEMYDOV, D.; ZHANG, W. & PETROVIC, Z.S. *Polyols and Polyurethanes from Hydroformylation of Soybean Oil.* Journal of Polymers and the Environment. Vol.10, p.49-52, 2002.

GUO, A.; JAVNI, I. & PETROVIC, Z.S. Rigid polyurethane foams based on soybean oil. Journal of Applied Polymer Science. Vol.77, n.2, p. 467-473, 2000.

HARRICK, N.J. *Surface chemistry from spectral analysis of totally internally reflected radiation.* The Journal of Physical Chemistry. Vol.64, p.1110-1114, 1960.

HU, Y.H.; GAO, Y.; WANG, D.N.; HU, C.P.; ZU, S.; VANOVERLOOP, L. & RANDALL, D. *Rigid polyurethane foam prepared from a rape seed oil based polyol.* Journal of Applied Polymer Science. Vol.84, n.3, p.591-597, 2002.

KONZEN, P.H.A.; FURTADO, J.C.; CARVALHO, C.W.; FERRÃO, M.F.; MOLZ, R.F.; BASSANI, I.A. & HÜNING, S.L. *Otimização de métodos de controle de qualidade de fármacos usando algoritmo genético e busca tabu.* Pesquisa Operacional. Vol.23, n.1, p.189-207, 2003.

LEARDI, R. *Application of genetic algorithm-PLS for feature selection in spectral data sets.* Journal of Chemometrics. Vol.14, p.543-565, 2000.

MIRABELLA, F.M. Jr. *Internal reflection spectroscopy.* Applied Spectroscopy Reviews. Vol.21, p.45-178, 1985.

NORGAARD, L.; SAUDLAND, A.; WAGNER, J.; NIELSEN J. P.; MUNCK, L. & ENGELSEN, S. B. *Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy.* Applied Spectroscopy. Vol.54, n.3, p.413-419, 2000.

PARREIRA, T.F.; FERREIRA, M.M.C.; SALES, H.J.S. & ALMEIDA, W.B. *Quantitative Determination of Epoxidized Soybean Oil Using Near Infrared Spectroscopy and Multivariate Calibration.* Applied Spectroscopy. Vol.56, p.1607-1614, 2002.

PETROVIC, Z.S.; GUO, A. & ZHANG, W. *Structure and properties of polyurethanes based on halogenated and nonhalogenated soy-polyols.* Journal of Polymer Science Part A: Polymer Chemistry. Vol. 38, n.22, p.4062-4069, 2000.

SPIEGELMAN, C. H.; MCSHANE, M. J.; COTÉ, G. L.; GOETZ, M. J.; MOTAMEDI, M. & YUE Q. L. *Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm.* Analytical

Chemistry. Vol.70, p.35-44, 1998.