UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE QUÍMICA - IQ

ALGORITMO GENÉTICO NA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA DE DADOS ESPECTROSCÓPICOS.

Dissertação de Mestrado

PAULO AUGUSTO DA COSTA FILHO

Orientador: Prof. Dr. RONEI JESUS POPPI

Campinas

1998

Aos meus pais,
Paulo e Maria,
minhas irmãs
Cátia e Márcia

Aos meus tios,
Hélvio , João,
Luisa e Terezinha

Agradecimentos

Aos Prof. Dr. Celio Pasquini, Prof. Dr. Jarbas J. R. Rohwedder, Prof. Dr. Marco Aurélio Zezzi Arruda, Prof. Dr. Ivo R. M. Júnior pelos ensinamentos e amizade.

Aos amigos do LAQQA, GIA e Química Ambiental : Eduardo, Renato, Henrique, Cesinha, Saliba, Cleidiane, Carminha, Fernanda, Dani, Raquel, Ricardo, Roberta, Jefferson, Iara, Martha, Ileana, Soraya, Randal, Paulo Fidêncio, Rosângela, Emerson, Sílvio, Giovana, Sérgio, Marcelo.

Aos amigos do dia a dia : Patrícia Tavares, Patrícia, Érico Teixeira, Edenir, Cláudio, Jordan.

Aos funcionários do Instituto de Química que contribuíram em muito com a execução deste trabalho.

À PETROBRÁS - REPLAN pelo fornecimento das gasolinas sem aditivos utilizadas neste trabalho.

Ao CNPq pela bolsa de mestrado.

Agradecimento especial :

Prof. Dr. Ronei Jesus Poppi pela amizade, orientação, apoio, dedicação e incentivo.

Quando surgem os ventos das mudanças,
alguns constroem abrigos,
outros moinhos.

Índice

Resumo.....vii

Abstract.....viii

Introdução.....1

Objetivos.....3

Capítulo 1 Conceitos Básicos

1. Introdução.....5
1.1 Arranjo de dados e notação Algébrica.....5
1.2 Posto de uma Matriz.....6
1.3 Autovetores e Autovalores.....6
1.4 Decomposição de Valores Singulares.....7
1.5 Pré-tratamento dos Dados.....7
1.6 Relação entre Matriz de Dados.....8
1.7 Regressão Linear Múltipla (RLM).....9
1.8 Mínimos Quadrados Parciais (PLS).....13
1.9 Método Generalizado de Posto (GRAM).....16

Capítulo 2 Algoritmo Genético

2. Introdução ao Algoritmo Genético.....21
2.1 Teoria de Darwin.....21
2.2 Algoritmo Genético.....22
2.2.1 Algoritmo Genético Básico.....23
2.3 Algoritmo Genético em Otimização.....30
2.4 Operações e Técnicas Avançadas.....31
2.5 Aplicação do Algoritmo Genético em Química.....35
2.6 Algoritmo Genético na Seleção de Variáveis.....37

Capitulo 3 Aplicações do Algoritmo Genético na Seleção de Variáveis

3 Aplicação Teórica.....40

3.1 Desenvolvimento dos programas.....40

3.2 Exemplo de funcionamento do Algoritmo Genético.....40

3.3 Determinação de Açúcares.....46

3.3.1 Introdução.....46

3.3.2. Determinação de Açúcares por ATR.....49

3.3.2.1 Estruturas Moleculares.....50

3.3.3 Procedimento Experimental.....52

3.3.3.1 Parâmetros de configuração do Algoritmo Genético.....56

3.3.4 Resultados e Discussão.....57

3.3.4.1 Determinação do Número ideal de Componentes Principais e de Amostras Anômalas.....57

3.3.4.2. Método dos Mínimos Quadrados Parciais na Construção do Modelo , na Faixa espectral entre 950 a 1200 cm^{-1}.....59

3.3.4.3. Algoritmo Genético na Seleção de Variáveis no Método dos Mínimos Quadrados Parciais, sem Pré-Tratamento dos Dados.....61

3.3.4.4 Algoritmo Genético na Seleção de Variáveis para a Regressão Linear Múltipla, sem Pré-Tratamento dos Dados.....63

3.3.4.5. Método do Mínimos Quadrados Parciais , com Pré-Tratamento dos Dados.....64

3.3.4.6. Algoritmo Genético na Seleção de Variáveis no Método dos Mínimos Quadrados Parciais, com Pré-Tratamento dos Dados.....65

3.3.4.7 Algoritmo Genético na Seleção de Variáveis para a Regressão Linear Múltipla, com Pré-Tratamento dos Dados.....66

3.3.4.8. Análise dos Resultados Obtidos.....67

Capítulo 3 Aplicações do Algoritmo Genético na Seleção de Variáveis

3.3.4.9 Estudo da Permuta de Variáveis Selecionadas entre os Modelos do PLS e RLM.....71

3.4 Determinação de Aditivos em Gasolina.....73

3.4.1 Introdução.....73

3.4.2 Procedimento Experimental.....76

3.4.2.1 Parâmetros de configuração do Algoritmo Genético.....79

3.4.3 Resultados e Discussão.....80

3.5 Determinação de Zinco e Cobre em Sistema de Fluxo.....89

3.5.1 Introdução.....89

3.5.2 Complexante.....91

3.5.3 Procedimento Experimental.....92

3.5.4 Preparação das Soluções para Padronização do Cu^{2+} e Zn^{2+}....92

3.5.4.1 Padronização das soluções estoque de Cu^{2+} e Zn^{2+}.....93

3.5.4.2 Geração do Gradiente de pH em Fluxo.....94

3.5.4.3 Seleção e validação das variáveis do Cu^{2+} e Zn^{2+}.....96

3.5.4.4 Parâmetros de configuração do Algoritmo Genético.....98

3.5.5 Resultados e Discussão.....99

Conclusão.....104

Perspectivas de Trabalhos Futuros.....106

Referências Bibliográficas.....108

Apêndice 118

A- Programa (em Linguagem Matlab) do Algoritmo Genético para a Seleção de Variáveis

Resumo

ALGORITMO GENÉTICO NA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO MULTIVARIADA DE DADOS ESPECTROSCÓPICOS.

Autor: Paulo Augusto da Costa Filho
Orientador: Ronei Jesus Poppi

Palavras-Chave : Quimiometria, análise por injeção em fluxo, infravermelho próximo, reflexão total atenuada.

Este trabalho descreve o desenvolvimento e aplicação do Algoritmo Genético na seleção de variáveis de dados espectroscópicos em calibração multivariada. O algoritmo genético é um algoritmo baseado na teoria da evolução de Darwin e tem despertado nos últimos anos grande interesse na comunidade cientifica. Um programa para a realização dos cálculos com o algoritmo genético foi escrito em linguagem MATLAB. Para a avaliação do programa foram realizados três experimentos: i - determinação simultânea de açúcares em solução por espectroscopia no infravermelho médio, utilizando reflexão total atenuada (ATR); ii – Determinação simultânea de Metil-terc-butil-éter e Etanol em Gasolina baseado em espectroscopia no infravermelho próximo; iii – Otimização simultânea de mais de uma dimensão de variáveis, utilizando o algoritmo genético no método generalizado de aniquilação de posto, aplicado em análise em fluxo para a determinação de Cobre e Zinco. Os dois primeiros experimentos envolveram a utilização dos métodos de calibração dos mínimos quadrados parciais (PLS) e da regressão linear múltipla (RLM), realizado-se a comparação dos resultados obtidos, com a seleção de variáveis. No terceiro experimento foi realizado a seleção de variáveis em dados bilineares, determinando os melhores conjuntos de tempo (pH's) e comprimentos de onda, de forma a otimizar a melhor condição para a determinação das espécies de interesse.

Abstract

GENETIC ALGORITHM IN VARIABLE SELECTION IN MULTIVARIATE CALIBRATION OF SPECTROSCOPIC DATA.

Author: Paulo Augusto da Costa Filho
Adviser: Ronei Jesus Poppi

Key words: Chemometric, flow injection analysis, near infrared, attenuated total reflection.

This work describes the development and application of genetic algorithm in variable selection of spectroscopic data in multivariate calibration. The genetic algorithm is based on Darwin evolution theory, and in last years has increase the interest in the scientific community. A program for calculations was developed in MATLAB language. For program evaluation were realize three experiments: i – Simultaneous sugars determination in solution with midi infrared spectroscopy using attenuated total reflection (ATR); ii – Simultaneous determination of Metyl-terc-butyl-ether and Ethanol in gasoline based on near infrared spectroscopy; iii – Simultaneous optimization in more than one variable dimension using genetic algorithm in generalized rank annihilation method, applied in flow system for determination of copper and zinc. In the first two calibration methods was used partial least squares (PLS) and multiple linear regression (MLR) and the results were compared with obtained, using variable selection.

The third experiment consisted in the variable selection in bilinear method, determing the best sets of time (pH) and wavelengths, in order to optimize the best condition for the analytes determination.

INTRODUÇÃO

Introdução

A partir de meados da década de oitenta, o desenvolvimento da eletrônica e a popularização dos computadores, catalisaram o processo de nças e inovações em vários segmentos da sociedade.

Na química, o reflexo dos avanços tecnológicos incidiu pesadamente sobre a das análises instrumentais, possibilitando a aquisição de maneira fácil e a de um grande número de dados de uma mesma amostra, e qüentemente o tratamento de dados obtidos passou a exigir cálculos mais lexos. Tal fato, aliado ao crescente número de análises, a diversidade das tras e a necessidade de resultados rápidos, abriram um leque de unidades para o desenvolvimento de uma nova área denominada ometria.

A quimiometria consiste essencialmente na aplicação de técnicas máticas e estatísticas, assim como de lógica matemática, em problemas icos[1]. O nome quimiometria originou-se da analogia a psicometria, etria, econometria, etc.

A quimiometria é mais que um ramo da química, por estar atuando mente e de forma relevante a inúmeros segmentos da química. Esta se entra na análise de dados químicos – assegurando que os dados rimentais contenham o máximo de informações – para definir ou selecionar melhores condições para realizar experimentos e medidas – realizar rminações simultâneas de mais de uma espécie de interesse, etc [2].

A quimiometria não é apenas uma simples ferramenta, pois, envolve entes métodos como:

I - Método de otimização;
II - Planejamento experimental;
III - Ferramentas estatísticas;
IV - Ajuste de curva;
V - Pesquisa de biblioteca;
VI - Processamento de sinal;

VII - Análises de fatores;

VIII – Resolução;

IX - Detecção;

X - Reconhecimento de padrões;

XI - Rede neural;

XII - Calibração.

A partir do início da década de noventa, a calibração multivariada apresentou uma explosão no número de publicações em química analítica. Isto provavelmente ocorreu, devido às novas perspectivas que surgiram, como a possibilidade de determinações simultâneas de múltiplos componentes, através de múltiplas medidas instrumentais sem a necessidade de medidas muito seletivas e realização de análises químicas, mesmo na presença de interferentes.

É interessante salientar que a eficiência de qualquer modelo matemático construído, a partir de um dos métodos de calibração multivariada, encontra-se relacionado com o planejamento experimental desenvolvido durante a elaboração do experimento. O planejamento experimental deve ser arquitetado de forma a abranger toda a região de interesse, e ser o mais representativo possível[3].

As análises envolvendo espectroscopia foram as que mais tiraram proveito da calibração multivariada. Isto certamente ocorreu, em decorrência da sua versatilidade no número de aplicações e diversidade de amostras, bem como do fato de permitir algumas vezes análises qualitativas e/ou quantitativas sem necessariamente destruir a(s) amostra(s) [4,5].

Uma das características dos dados gerados em análises espectroscópicas é o elevado número de variáveis independentes (absorbância em cada comprimento de onda)[4,5]. Em algumas situações, o número elevado de variáveis independentes pode agir de forma depreciativa à análise dos dados, pois há situações onde muitas destas variáveis contribuem com pouca ou nenhuma informação para o modelo matemático. Na utilização da calibração multivariada surgiu a necessidade de usar um algoritmo que procure selecionar o melhor conjunto de variáveis, que apresente melhor sensibilidade e linearidade para o(s)

analito(s) de interesse, minimizando ou até mesmo eliminando características potenciais dos interferentes, bem como não-linearidades.

Recentemente tem-se observado uma tendência na aplicação do algoritmo genético na seleção de variáveis[6], entretanto há outros métodos de seleção de variáveis[7-10]. A preferência pelo algoritmo genético deve-se a sua eficiência, versatilidade e robustez [6]. A seleção de variáveis torna possível um uso mais intensivo de modelos mais simples como a regressão linear múltipla (RLM) em calibração multivariada, que até então se limitava a aplicações onde o número de variáveis independentes era menor ou igual ao número de amostras[11].

Objetivos:

Os objetivos deste trabalho encontram-se dentro das perspectivas do Laboratório de Quimiometria em Química Analítica (LAQQA), do IQ-UNICAMP, relacionados ao aprofundamento e domínio das técnicas quimiométricas e otimização de sistemas complexos. Este trabalho teve como objetivo mais específico, estudar a possibilidade do uso do algoritmo genético na seleção de variáveis em calibração multivariada, visando:

- Estudar a aplicação da regressão linear múltipla ao invés do método dos mínimos quadrados parciais, considerado atualmente como método padrão de calibração multivariada, (uma vez que este é mais robusto (permite que os parâmetros da modelagem sejam expandidos para diferentes amostras)).
- Analisar espécies de interesse mesmo em regiões que possuam pouca informação do analito em questão.
- Selecionar variáveis em sistemas químicos que possuam mais de uma dimensão, como dados bilineares gerados por espectrofotômetro com detector de arranjo linear de fotodiodos, utilizando o Método Generalizado de Aniquilação de Posto (GRAM), como metodologia de calibração multivariada.

No capitulo 1, será mostrada uma breve introdução da notação algébrica e teoria básica dos métodos de calibração multivariada, bem como a interpretação

geométrica da regressão linear múltipla (RLM), mínimos quadrados parciais (PLS) e o método generalizado de aniquilação de posto (GRAM).

No capitulo 2, será apresentada a base sobre a teoria geral do algoritmo genético e no caso especial para a seleção de variáveis.

O capítulo 3, irá abordar as aplicações do algoritmo genético na seleção de variáveis em métodos de calibração multivariada, para um caso teórico de otimização e para três casos reais. Na primeira aplicação pratica, foram quantificadas misturas de açúcares (glicose, maltose e frutose) em solução na região do infravermelho médio. A segunda aplicação foi a determinação de aditivos (etanol e metil-terc-butil-éter) em combustível (gasolina) através de espectros na região do infravermelho próximo, utilizando um espectrofotômetro do infravermelho próximo baseado em um filtro óptico acústico sintonizável (AOTF). A terceira aplicação foi a determinação de cobre e zinco por complexação com zincon, explorando o gradiente de pH. Neste caso, a aquisição dos dados foi realizada por um espectrofotômetro visível com detector de arranjo linear de fotodiodos, gerando matrizes de dados referentes a cada amostra. No final será apresentada uma conclusão do trabalho.

CAPÍTULO I

Métodos de Calibração

Multivariada

1. Introdução.

1.1 *Arranjo dos dados e notação de Álgebra linear.*

A Figura 1 ilustra a representação de uma matriz de dados, com *n* linhas e *p* colunas, onde o elemento da matriz x_{np}, indica o valor do n-ésimo objeto e da p-ésima variável.

Os objetos geralmente são espécies químicas, e as variáveis são muitas vezes derivadas das propriedades das espécies químicas dos objetos.

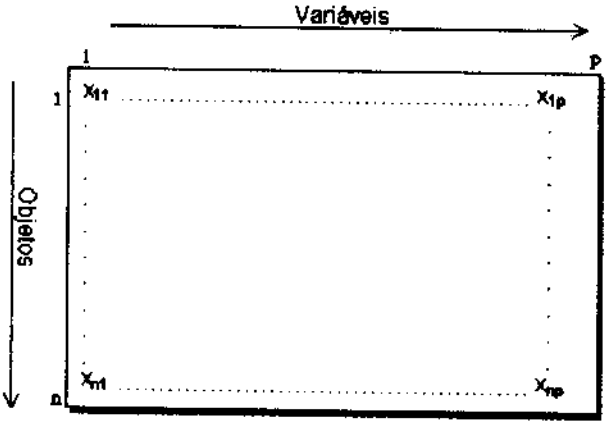


Figura 1. Representação matricial dos dados.

Em problemas como os de calibração multivariada, onde o número de objetos e de variáveis é muito grande, torna-se absolutamente indispensável à disposição ordenada dos dados em forma de matriz, para tornar mais fácil a manipulação dos dados.

Escolheu-se uma estrutura matricial onde cada linha representa um objeto e as colunas suas respectivas variáveis.

Quando se trabalha com equipamentos que geram dados bilineares como, por exemplo: HPLC (cromatografia gasosa de alta eficiência) com arranjo de fotodiodos, cromatografia gasosa (CG) com espectrômetro de massa (MS),

fluorescência emissão/excitação, etc., cada amostra gera uma matriz de dados, onde as linhas não serão as amostras, mas sim alguma outra propriedade que está sendo monitorada.

Neste trabalho será usada a notação padrão de álgebra linear na descrição dos métodos matemático empregados. As matrizes serão denotadas por letras maiúsculas em negrito (**X**), os vetores por letras minúsculas em negrito (**y**). A transposta de uma matriz ou vetor será representada por um sobrescrito "T" ($\mathbf{X}^T$), a inversa pelo sobrescrito "-1" ($\mathbf{y}^{-1}$). A priori, todos os vetores serão vetores coluna, os correspondentes vetores linha serão representados por vetores transpostos. As matrizes diagonais serão denodadas por letras maiúsculas (S). As letras comuns indicarão os escalares que podem ser elementos de matrizes (a_{ij}), de vetores (a_j) ou outras constantes como coeficientes de regressão (b)[12].

1.2 *Posto de uma Matriz*

Posto de uma matriz é o número de vetores linearmente independentes que compõe a matriz, ou seja, são os vetores que não podem ser escritos como uma combinação linear de outros vetores que pertençam ao mesmo espaço vetorial. A interpretação química para o posto é o número de espécies distintas contidas nas amostras químicas, desprezando os ruídos aleatórios inerentes às medidas.

1.3 *Autovetores e Autovalores*

Quando um operador (representado por uma matriz) é aplicado a um espaço vetorial e o produto dessa operação retorna o próprio espaço vetorial multiplicado por uma constante, tem-se uma equação de autovetores e autovalores [13]:

$$\mathbf{M}\Psi = \Lambda\Psi \qquad (01)$$

Onde a matriz **M** é o operador que é aplicado no espaço vetorial formado pelos vetores da matriz Ψ resultando em uma constante (Λ) multiplicada pelo próprio espaço vetorial Ψ.

1.4 *Decomposição de Valores Singulares*

Este algoritmo permite decompor uma matriz no produto de três matrizes, gerando um novo conjunto de bases vetoriais [13]:

$$\mathbf{X} = \mathbf{USV}^T \qquad (02)$$

As colunas de **U** são autovetores de $\mathbf{XX}^T$, abrangendo assim o espaço vetorial das linhas de **X**. As colunas de **V** são autovalores de $\mathbf{X}^T\mathbf{X}$, abrangendo o espaço vetorial das colunas de **X**. **S** é uma matriz diagonal que indica o quanto de variância é descrita por cada um dos vetores bases correspondentes, definidos por **V**. Estes são arranjados em ordem decrescente de variância explicada. O produto das matrizes **US** define as coordenadas na nova base que é denominada de scores.

1.5 *Pré-tratamento dos dados*

O pré-tratamento é análogo as operações de pesagem, extração, secagem, centrifugação, precipitação de uma análise química, e freqüentemente é a fase mais difícil, consome mais tempo e determinam a qualidade e eficiência do método de calibração multivariada empregada.

Antes de desenvolver um modelo utilizando qualquer dos métodos de calibração multivariada, é aconselhável um pré-tratamento dos dados, para eliminar amostras anômalas, minimizar ruídos e informações superpostas de espécies de interesse, bem como de interferentes.

Durante a etapa de manipulação preparação, acondicionamento ou aquisição dos dados, pode-se ter variações nos procedimentos utilizados nas

amostras, dando origem ao que chamamos de amostras anômalas. Os problemas mais típicos que podem originar o aparecimento de amostras anômalas são [14]:

- interferentes físicos (ex. espalhamento de luz na amostra)
- interferentes químicos (ex. instabilidade do analito)
- composição extrema (ex. níveis alto da resposta não linear)
- erro nos dados de referência (ex. erro na digitalização ou digitação dos dados)
- erros nos dados instrumentais (ex. relação sinal/ruído muito baixa)

Umas das maneiras de se determinar estas amostras é através da análise das componentes principais[14].

Entre os pré-processamentos o mais utilizado é o escalonamento dos dados, que consiste em atribuir a mesma influência relativa das diferentes variáveis independente das suas unidades, uma vez que todas as variáveis passam a ser expressas em unidades de desvio padrão.

Para minimizar o ruído pode-se utilizar o filtro com média móvel, transformada de Fourier, transformada de Wavelet, Savitsky-Golay, etc.[14]

A derivada é freqüentemente usada para uma melhor definição de picos que se encontram sobrepostos em uma mesma região e para a correção de linha base.

Outro procedimento para o pré-tratamento dos dados, que será melhor abordado ao longo deste trabalho, é a seleção de variáveis que permite eliminar os termos que não são relevantes na modelagem, gerando uma submatriz com apenas as variáveis que possuem informação.

1.6 *Relação entre Matrizes de dados*

A calibração multivariada utiliza modelos matemáticos para estabelecer uma relação entre uma propriedade que possa ser monitorada, com alguma outra propriedade de interesse.

A calibração multivariada pode ser dividida em duas fases: calibração e previsão. Na etapa de calibração é encontrada uma relação entre as medidas físicas (matriz **X**, variáveis independentes) e as propriedades de interesse (matriz

Y, variáveis dependentes). Na etapa de previsão, avalia-se a performance do modelo matemático construído a partir de um segundo conjunto de amostras. Os dados usados nesta fase recebem o nome de conjunto de validação.

A representação esquemática da etapa de calibração está ilustrada na Figura 2.

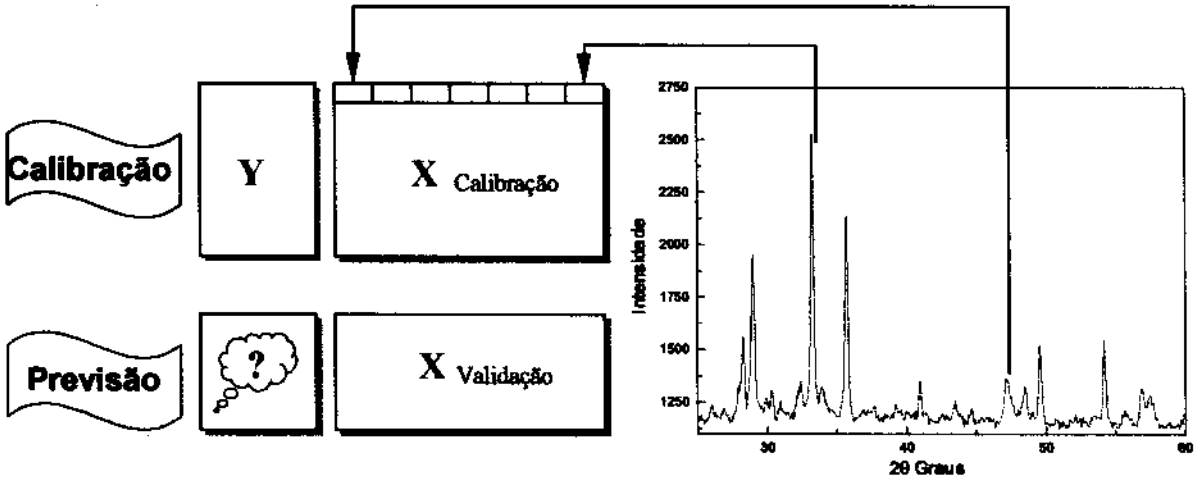


Figura 2. Organização dos dados para a calibração multivariada.

1.7 Regressão linear Múltipla (RLM).

O objetivo da regressão linear múltipla é encontrar uma combinação linear entre as variáveis independentes (medidas instrumentais) e as variáveis dependentes (propriedade de interesse) que minimizem o erro na estimativa dessas propriedades de interesse[12]. Tal relação pode ser escrita como:

$$\hat{y} = x_1 b_1 + x_2 b_2 + x_3 b_3 + ... + x_p b_p + e \qquad (03)$$

$$\hat{Y} = \mathbf{x}^T\mathbf{b} + \mathbf{e} \qquad (04)$$

Onde "$\hat{Y}$" é propriedade estimada (variável dependente)

"**x**" é vetor com as medidas instrumentais (variáveis independentes)

"**b**" é o vetor dos coeficientes de regressão

"**e**" é o vetor de resíduo ou erro

A regressão linear múltipla nada mais é que a resolução de sistemas lineares, portanto as seguintes condições são validas (p é o número de variáveis e n é o número de amostras):

Se p > n, (onde se tem mais variáveis que amostras e) não existe solução para **b**.

Se p = n, (onde se tem número de amostras e variáveis iguais e) há uma única solução para **b**.

Se p < n, (onde se tem mais amostras que variáveis e) não existe uma solução exata para **b**.

Uma maneira mais fácil de interpretar geometricamente o que significa a resolução de um sistema de equações, é ilustrado nas Figuras 3 e 4, onde cada plano é definido por uma equação do tipo: Z= aX + bY.

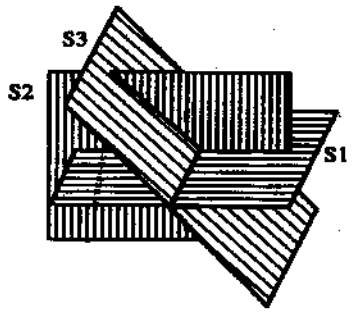


Figura 3. Interseção de três planos em um único ponto.

A Figura 3 representa um caso, onde o sistema de equações possui apenas uma solução e três equações, onde a interseção dos três planos representa a resposta do sistema linear de equações.

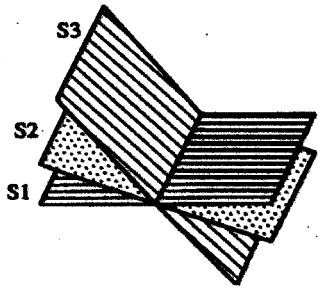


Figura 4. Interseção de três planos definindo uma linha.

A Figura 4 representa o caso particular de um sistema linear de equações que possui infinitas soluções, já que a interseção das três equações determina uma reta com infinitos pontos. Cada ponto desta reta é uma possível solução para este sistema de equações.

A regressão linear múltipla também pode ser entendida como a projeção da(s) coluna(s) das variáveis dependentes sobre o espaço definido pelas colunas da matriz de variáveis independentes. Uma maneira de representar geometricamente esta interpretação é construir um gráfico das colunas da matriz **X** (variáveis independentes), conforme ilustrado na Figura 5. Em seguida projetam-se as colunas da matriz de variáveis dependentes sobre o espaço definido pelos vetores das colunas das variáveis independentes, conforme representado na Figura 6.

$$X = \begin{vmatrix} 6 & 15 & 29 \\ 11 & 18 & 39 \end{vmatrix}$$

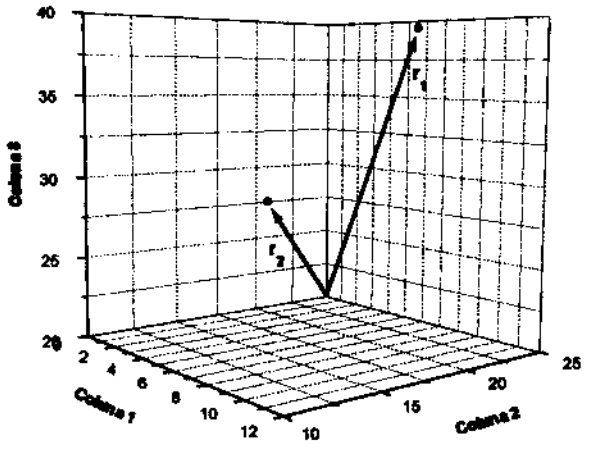


Figura 5. Representação da matriz **X** no espaço das colunas.

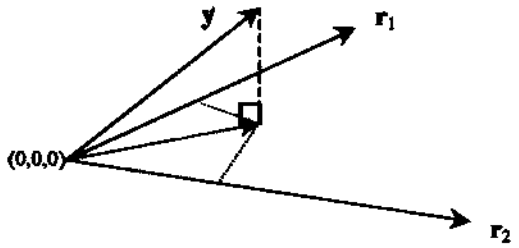


Figura 6. Projeção das colunas das variáveis dependentes sobre o espaço gerado pelas colunas das variáveis independentes.

Os coeficientes da regressão linear múltipla (equação 04) são calculados utilizando o critério dos mínimos quadrados, conforme apresentados na equação 05 :

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (05)$$

As limitações da regressão linear múltipla concentram-se no número de amostras que não pode ser menor do que o número de variáveis, e em determinar a inversa de $(\mathbf{X}^T\mathbf{X})$, pois se as variáveis independentes possuírem uma elevada

correlação, esta inversa pode não existir. Mesmo assim, se $(X^TX)^{-1}$ existir com correlações altas, os valores previstos para "b" terão erros elevados.

1.8 Mínimos Quadrados Parciais (PLS).

O método dos mínimos quadrados parciais é um modelo baseado em variáveis latentes (fatores), onde cada fator é definido como uma combinação linear das variáveis originais da matriz X (variáveis independentes) ou Y (variáveis dependentes)[15,16].

A primeira componente principal correspondente ao maior autovalor é, pela definição, a direção no espaço de X que descreve a máxima quantidade de variância nas amostras. Quando toda a variância de um conjunto de amostras não puder ser explicada por apenas uma componente principal, uma segunda componente principal perpendicular ou ortogonal a primeira será usada. Após a modelagem, teoricamente a matriz dos quadrados dos resíduos deverá conter apenas a variância não explicada associada ao ruído.

A importância da ortogonalidade das componentes principais é devido ao fato de que somente desta forma pode-se garantir que a nova base formada resulta de uma combinação de vetores linearmente independentes, portanto constituindo um novo espaço vetorial.

O método dos mínimos quadrados parciais (PLS – Parcial Least Squares) foi proposto inicialmente por H. Wold e baseia-se nas propriedades do algoritmo NIPALS. Este consiste em escrever a matriz X de posto *p* como uma soma de matrizes de posto 1, conforme a equação 06.

$$X = M_1 + M_2 + M_3 + \ldots + M_p \qquad (06)$$

ou graficamente representando pela Figura 7:

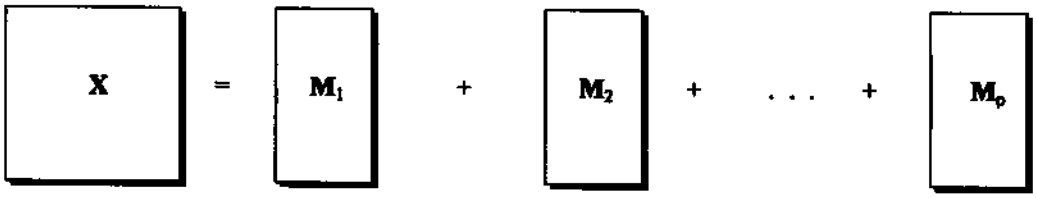


Figura 7. Esquema da decomposição de uma matriz **X** na soma de matrizes de posto 1.

Uma matriz de posto 1 pode ser escrita como produto interno de dois vetores, os scores (t_h ou u_h) e os loadings (p^T_h ou q^T_h), semelhante a Figura 8.

$$X = t_1 p^T_1 + t_2 p^T_2 + \ldots + t_p p^T_p \qquad (07)$$

$$X = T P^T + E \qquad (08)$$

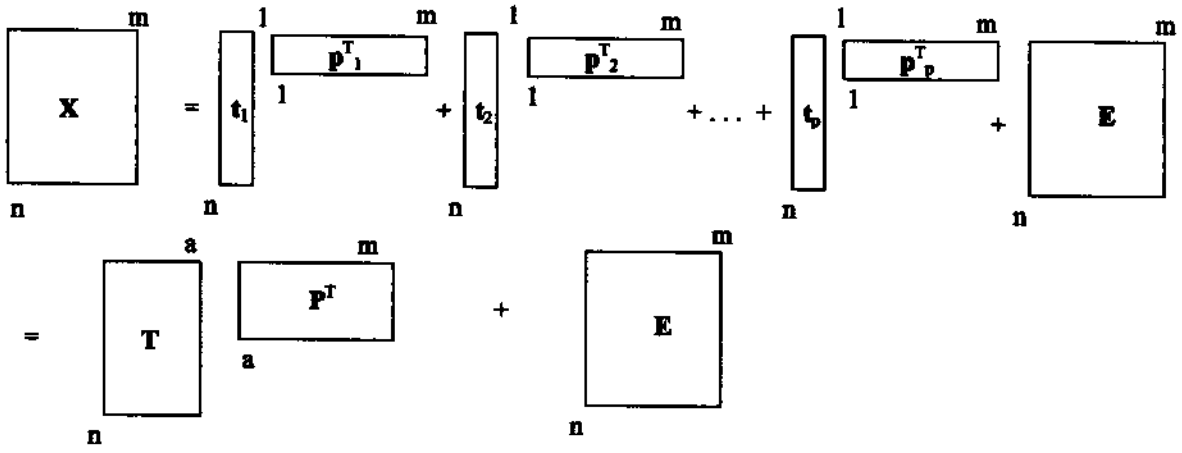


Figura 8. Esquema da decomposição de uma matriz **X** em loadings e scores mais uma matriz de erros.

Da mesma forma a matriz Y (matriz de variáveis dependentes) também é decomposta como:

$$Y = UQ^T + F \qquad (09)$$

Após obter-se as relações externas das componentes principais de X (X = T P^T + E) e Y (Y = UQ^T + F), realiza-se uma relação interna entre os scores de X e Y, afim de se determinar uma relação linear da forma $u_h = b_h t_h$ onde:

$$b_h = u_h t_h / t_h t_h \qquad (10)$$

O processo de desenvolvimento do modelo com o PLS está ilustrado na Figura 9.

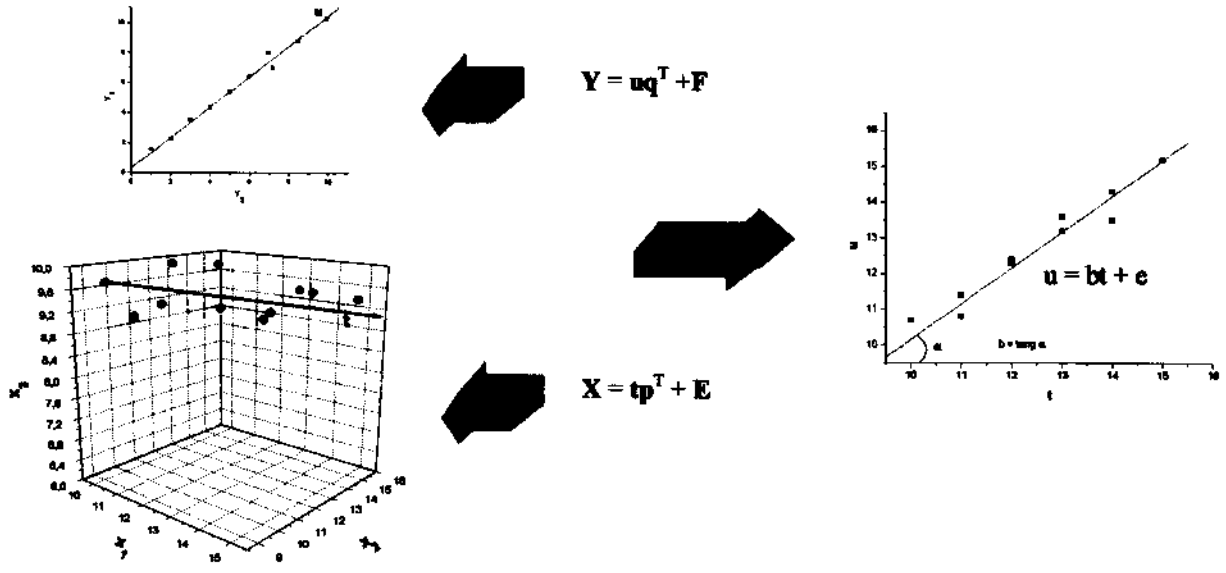


Figura 9. Representação da relação ente os scores da matriz X e Y.

Na Figura 9, o b representa o coeficiente angular do vetor que melhor explicar a relação existente entre os scores de X e Y. Como as componentes principais de X e Y são calculadas independentemente, a relação entre elas não é

necessariamente a melhor. Para minimizar este problema, faz-se uma leve rotação nos eixos das componentes principais de forma a não comprometer a ortogonalidade. Esta leve rotação das componentes principais provoca um desvio da direção de maior variância das matrizes originais.

1.9 Método generalizado de aniquilação de posto (GRAM).

Esta metodologia de calibração foi desenvolvida [17] para ser aplicada em dados bilineares.

Nos instrumentos que geram dados bilineares, uma espécie pura apresentará uma matriz de resposta (**N**) do tipo:

$$\mathbf{N} = \mathbf{x} c_n \mathbf{y}^T \qquad (11)$$

Onde **x** e **y** são vetores coluna que representam as respostas em cada ordem (espectro puro e perfil de concentração dependente do tempo), e c_n é a concentração da respectiva espécie.

Para uma amostra com mais de uma espécie e composição desconhecida, a matriz de resposta pode ser descrita como:

$$\mathbf{M} = \sum_{k=1}^{i} \mathbf{x}_k c_{k,n} \mathbf{y}_k^T = \mathbf{X}\, C_m\, \mathbf{Y}^T \qquad (12)$$

Onde i é o número total de componentes, **X** e **Y** são matrizes cujas colunas são $\mathbf{x}_k$ e $\mathbf{y}_k$, C_m é uma matriz diagonal cujos elementos são $c_{k,n}$.

De forma análoga, pode-se decompor a matriz de resposta para a amostra de calibração (padrão) como:

$$\mathbf{N} = \mathbf{X}\, C_n\, \mathbf{Y}^T. \qquad (13)$$

As matrizes **M** e **N** tem em comum os blocos **X** e $\mathbf{Y}^T$ (os dados de calibração e da amostra de composição desconhecida foram gerados pela mesma técnica

nas mesmas condições), diferindo apenas nas concentrações. Resolvendo para as equações 12 e 13, temos:

$$\mathbf{X}\, C_m = \mathbf{M}(\mathbf{Y}^T)^+ \qquad (14)$$

$$\mathbf{X}\, C_n = \mathbf{N}(\mathbf{Y}^T)^+ \qquad (15)$$

Onde o sinal "+" sobrescrito indica a pseudo-inversa [18]. Multiplicando a equação 14 por C_n e a equação 15 por C_m e combinando os termos temos:

$$\mathbf{N}(\mathbf{Y}^T)^+ C_m = \mathbf{M}(\mathbf{Y}^T)^+ C_n \qquad (16)$$

substituindo $\mathbf{Z} = (\mathbf{Y}^T)^+$ temos:

$$\mathbf{N}\,\mathbf{Z}\, C_m = \mathbf{M}\,\mathbf{Z}\, C_n \qquad (17)$$

Quando a amostra de composição desconhecida (**M**) contém espécies que não estão presente na amostra padrão (**N**) e vice-versa, tem-se o caso mais geral na determinação simultânea de componentes em misturas. Tanto a superfície relacionada à amostra padrão quanto a da amostra desconhecida contém dados relativos a espécies distintas e espécies em comum. Neste caso o posto da soma das matrizes é maior que o posto de qualquer uma das matrizes independentes e menor que o a soma dos postos das duas matrizes.
A equação 17 pode ser reescrita como:

$$\mathbf{M}\,\mathbf{Z} = \mathbf{N}\,\mathbf{Z}\,\Lambda \;;\; \Lambda = C_m\, C_n^{-1} \qquad (18)$$

Para a resolução da equação 18 de autovalores e autovetores, é necessário inicialmente que as matrizes sejam quadradas e inversíveis.

Uma transformação das matrizes **N** e **M** deve ser realizada de modo que se formem matrizes quadradas, mantendo inalterado o posto de ambas. Para essa transformação é necessário determinar dois conjuntos de vetores ortonormais $\mathbf{p}_i$ e

q_i que formarão uma base para a representação das duas matrizes de dados em questão. Os vetores p_i formarão as colunas de uma nova matriz **P** e os vetores q_i as colunas de outra nova matriz **Q**. As matrizes **P** e **Q** são calculados a partir da decomposição em valores singulares de duas novas matrizes formadas colocando-se **M** e **N** lado a lado e uma sobre a outra.

$$\mathbf{M} \mid \mathbf{N} = \mathbf{P}\, S_1\, \mathbf{V}^T \qquad (19)$$

$$\underline{\mathbf{M}} = \mathbf{U}\, S_2\, \mathbf{Q}^T \qquad (20)$$
$$\mathbf{N}$$

Onde **P**, **V**, **U** e **Q** são autovetores e S_1 e S_2 são matrizes de valores singulares (matrizes diagonais). Como os dados experimentais apresentam ruído que não devem ser modelados, faz-se essa decomposição considerando apenas os valores significativos de **M** | **N** e **M** / **N**. Essa truncagem é realizada tomando apenas as "i" colunas significativas de **U** e **V** que contém o ruído minimizado e a submatriz formada no canto superior direito da matriz **S** com dimensões "i" por "i". A constante "i" é o número de componentes principais, ou seja, o número de autovetores necessários para se explicar a variância nos dados de **M** e **N**:

Determinadas as matrizes **P** e **Q**, a equação 18 pode ser resolvida substituindo as matrizes **M** e **N** pelas projeções $\mathbf{M}_{PQ}$ e $\mathbf{N}_{PQ}$ das mesmas sobre a base formada por **P** e **Q**.

$$\mathbf{M}_{PQ} = \mathbf{P}^T \mathbf{M}\, \mathbf{Q} = (\mathbf{P}^T \mathbf{X})\, C_m\, (\mathbf{Q}^T \mathbf{Y})^T \qquad (21)$$

$$\mathbf{N}_{PQ} = \mathbf{P}^T \mathbf{N}\, \mathbf{Q} = (\mathbf{P}^T \mathbf{X})\, C_n\, (\mathbf{Q}^T \mathbf{Y})^T \qquad (22)$$

$$\mathbf{M}_{PQ}\, \mathbf{Z} = \mathbf{N}_{PQ}\, \mathbf{Z}\, \Lambda; \quad \Lambda = C_m\, C_n^{-1} \qquad (23)$$

Resolvendo então a equação generalizada de autovalores e autovetores, encontra-se a razão entre as concentrações para das espécies presentes nas duas amostras (padrão e desconhecida).

Para resolver esta equação, utiliza-se o algoritmo QZ. Somente é possível garantir que os autovetores e autovalores da equação 23 são reais e maiores

que zero se $\mathbf{M}_{pq}$ e $\mathbf{N}_{pq}$ forem positivos e não singulares. Como tanto $\mathbf{M}_{pq}$ quanto $\mathbf{N}_{pq}$ são matrizes quadradas reais e assimétricas, não se pode afirmar que o algoritmo QZ só produzirá autovetores e autovalores reais.

Assim, após o uso do algoritmo QZ para resolver a equação 23, as matrizes de autovalores e seus autovetores correspondentes ficam distribuídas em ordem de magnitude (do maior para o menor) dos autovalores em Λ, agrupando os complexos conjugados em pares. Em seguida encontra-se uma matriz de transformação de similaridade $\mathbf{T}^{\otimes}$, a qual transforma $\mathbf{Z}$ e Λ em $\mathbf{Z}^{\otimes}$ e $\Lambda^{\otimes}$. Autovalores reais ficam sempre associados a autovetores reais, e autovetores complexos conjugados ficam sempre associados a autovetores complexos conjugados. Finalmente encontra-se a segunda matriz de transformação de similaridade $\mathbf{T}^{\otimes}$, a qual converte $\mathbf{Z}^{\otimes}$ e $\Lambda^{\otimes}$ em $\mathbf{Z}^{\otimes}$ e $\Lambda^{\otimes}$, que não contém mais números complexos.

Uma vez que $\mathbf{Z}$ e Λ são transformados em $\mathbf{Z}^{\otimes}$ e $\Lambda^{\otimes}$, os espectros puros e os perfis de concentração com o pH podem ser calculados de acordo com as equações 24 e 25. Como não há valores complexos em $\mathbf{Z}^{\otimes}$, os espectros e perfis de concentração $\mathbf{X}$ e $\mathbf{Y}$ são reais.

$$\mathbf{X} = \mathbf{P}\mathbf{N}_{PQ}\mathbf{Z}^{\otimes} \qquad (24)$$

$$\mathbf{Y} = \mathbf{Q}((\mathbf{Z}^{\otimes})^{+})^{T} \qquad (25)$$

Após estimar-se os componentes puros, a concentração dos constituintes de interesse na amostra desconhecida pode ser calculada pelas equações 26, 27 e 28. Uma matriz $\mathbf{R}$ fornece as razões entre as concentrações dos constituintes de interesse nessas duas matrizes:

$$\mathbf{C}_m = (\mathbf{P}^T\mathbf{X})^{-1}\,\mathbf{M}_{PQ}\,((\mathbf{Q}^T\mathbf{Y})^T)^{-1} \qquad (26)$$

$$\mathbf{C}_n = (\mathbf{P}^T\mathbf{X})^{-1}\,\mathbf{N}_{PQ}\,((\mathbf{Q}^T\mathbf{Y})^T)^{-1} \qquad (27)$$

$$\mathbf{R} = \mathbf{C}_m\mathbf{C}_n^{-1} \qquad (28)$$

A partir do método generalizado de aniquilação de posto é possível através de uma amostra padrão obter o perfil de concentração e os espectros das espécies puras presentes na amostra teste e na amostra padrão, conforme esquematizado na Figura 10.

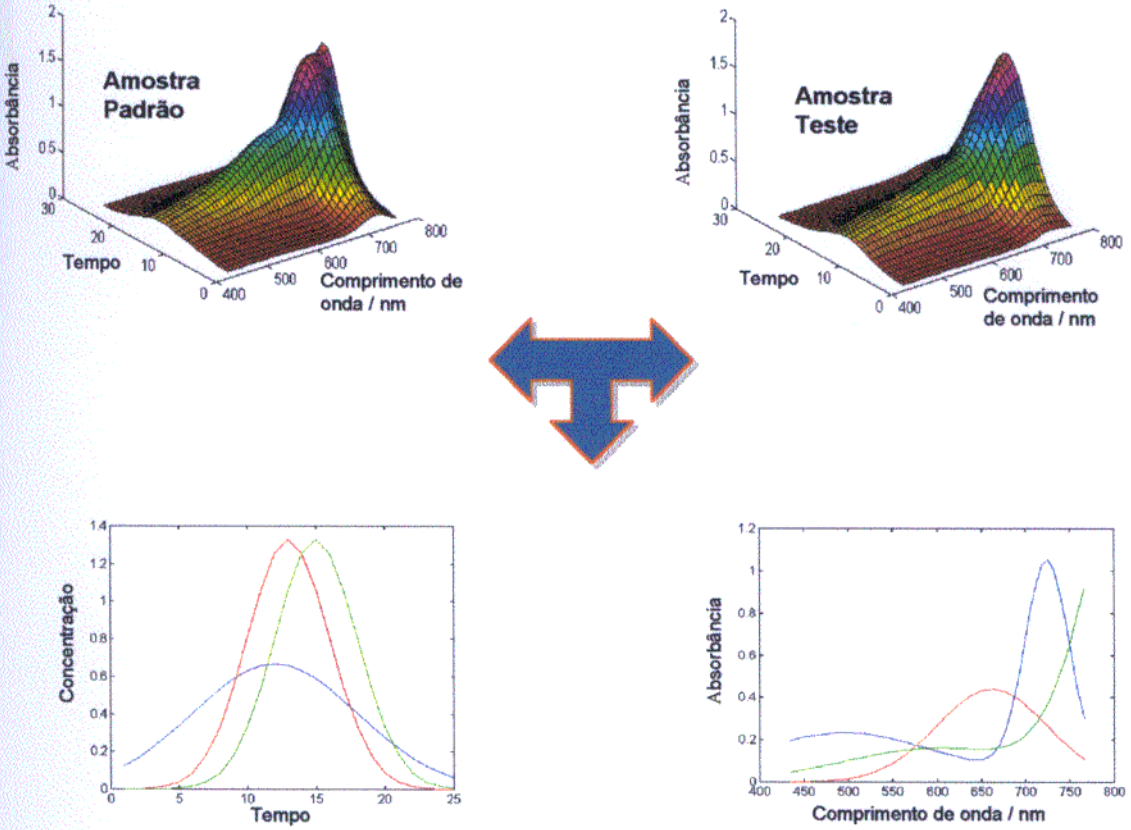


Figura 10. Deconvolução do espectro das espécies puras de interesse e perfil de concentração, através do método generalizado de aniquilação de posto (GRAM).

CAPÍTULO II

Algoritmo Genético

2. Introdução ao Algoritmo Genético.

2.1. *Teoria de Darwin.*

A teoria da evolução das espécies originou-se das observações realizadas por Charles R. Darwin, durante uma expedição cientifica ao redor do mundo. Nesta ocasião Darwin assumia a tarefa de naturalista do navio Eagle.[19] Durante esta jornada de cinco anos, Darwin presenciou inúmeras situações que o levaram a conjecturar sobre a possível seleção natural das espécies. No arquipélago dos Galápagos, ele verificou que cada ilha possuía uma fauna característica, principalmente com relação à pássaros, tartarugas e lagartos. Em 1858, Darwin apresentou a teoria da seleção natural numa sessão da Academia Britânica de Ciência. No ano seguinte, publicou um resumo do livro "Natural Selection", onde introduzia a teoria da Evolução das espécies. Segundo Darwin, no mecanismo de seleção natural, os indivíduos mais adaptados têm maior probabilidade de sobrevivência do que os menos adaptados. Como conseqüência disso, os mais adaptados teriam maior número de descendentes , o que provocaria, na geração seguinte, um aumento no seu número, em relação aos não adaptado. Tais mudanças remetem a evolução biológica, a qual ocasiona mudanças nas propriedades das populações dos organismos que transcendem o período de vida de um único indivíduo. As mudanças nas populações que são consideradas evolutivas são aquelas herdáveis via material genético, de uma geração para a outra. A evolução biológica pode ser pequena ou substancial, sendo estes efeitos responsáveis pelas adaptações dos organismos a diferentes ambientes. [19]

2.2 Algoritmo Genético.

A partir da teoria de Darwin evolutiva das espécies, John H. Holland um pesquisador da Universidade de Michigan, propôs nos anos 60 a construção de um algoritmo matemático para otimização em sistemas complexos, batizado de Algoritmo Genético .[6]

Foi nos anos oitenta, que o algoritmo genético recebeu um grande impulso em diversas áreas de aplicação cientifica, devido principalmente à popularização dos computadores e consequentemente aparecimento de computadores cada vez mais rápidos e potentes. Suas áreas de maior aplicação são : [20]

- Processamento de imagem;
- Modelagem e identificação de sistemas;
- Filtros adaptivos, para o cancelamento de ruídos;
- Robótica;
- Parâmetros de redes neurais.

Em química, a primeira aplicação de algoritmo genético foi no início dos anos noventa por Lucasius e Kateman [21]. As aplicações deste algoritmo cresceram, principalmente devido as vantagens que este apresenta quando comparado a outros métodos de otimização. Como principais vantagens pode-se destacar : [6]

i - Não requer informações sobre o gradiente da superfície de resposta;

ii - Eventuais descontinuidades da superfície de resposta não afetam a performance da otimização;

iii - A presença de mínimos locais não reduzem a eficiência do algoritmo;

iv - A performance do algoritmo tem apresentado excelentes resultados para problemas de otimização de grande escala.

Basicamente há quatro tipos principais de algoritmos genéticos: Algoritmo genético de Holland, programa genético de Koza, programa evolucionário de Fogel e estratégica evolucionária de Rechenberg. Eles se diferem pelo mecanismo de código genético, procedimento de seleção para a reprodução. O mais popular é o de Holland, pois este codifica as informações na forma de código binário (0,1).[22-23]

2.2.1 *Algoritmo Genético Básico.*

O algoritmo genético foi estruturado de forma que as informações referentes a um determinado sistema pudessem ser codificadas de maneira análoga a dos cromossomos biológicos. Desta forma o algoritmo proposto assimilara-se muito ao processo evolutivo biológico, facilitando a confecção do algoritmo matemático.

O algoritmo genético básico envolve cinco passos: codificação das variáveis, criação da população inicial, avaliação da resposta, cruzamento e mutação.

i. **Codificação das variáveis.**

O cromossomo biológico é composto de genes, os quais são responsáveis por uma determinada característica do indivíduo, como por exemplo cor dos olhos, altura, cor da pele, etc. Através de uma analogia, é possível construir um cromossomo artificial, num processo químico onde cada gene representará um determinado parâmetro que será otimizado como: temperatura, agitação, tempo, quantidade do catalisador (conforme Figura 12).

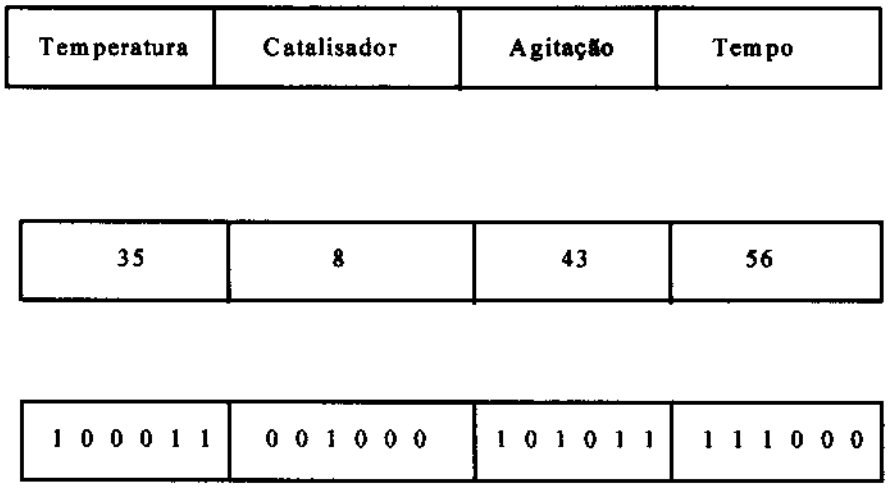


Figura 11. Codificação dos parâmetros do cromossomo em números binários de seis bits.

Neste tipo de sistema, a codificação em número binários é a forma mais simples de representar as informações relativas a cada parâmetro que será otimizado (como temperatura, catalisador, agitação e tipo de catalisador). Assim cada cromossomo artificial será representado por uma seqüência de zeros e uns (forma binária para cada parâmetro). As vantagens deste tipo de código são: [6]

i - Torna as operações genéticas mais fáceis de serem executadas;

ii - Torna possível o trabalho com variáveis, que assumem valores contínuos;

iii - Facilita a variação da faixa de trabalho dos parâmetros;

iv - Apresenta relativa simplicidade na variação da precisão dos valores de cada parâmetro.

A Figura 11, mostra a codificação dos valores relativos a quatro parâmetros que serão otimizados. Neste caso deve-se codificar para binário o valor de temperatura igual a 35 graus, o valor do catalisador igual a 8 gramas , o valor de agitação igual a 43 rpm e o valor do tempo igual a 56 minutos. Tomando-se como seis bits cada um desses parâmetros, tem-se assim o cromossomo formado por 24 unidades binárias. Nesse caso os seis primeiros bits formam o gene correspondente à temperatura, e na seqüência de seis bits os genes para catalisador, agitação e tempo.

ii - **Criação da população inicial.**

Através de um gerador randômico, cria-se uma população inicial de cromossomos, como mostrado na Figura 12, com uma população de sete indivíduos (onde cada indivíduo é representado por um cromossomo).

[1] 1000000000010010001010
[2] 0001001001001000000010
[3] 1000010000000000000010000
[4] 0000000001000000000001000
[5] 0100010000000100000000000
[6] 0000000010000000010000000
[7] 0000010000000001000000100

Figura 12 . Representação de uma população aleatória de sete cromossomos.

Sendo a população inicial aleatória, garantimos que não haverá nenhum tipo de influência tendenciosa do meio externo. Desta forma tenta-se garantir a idoneidade da pesquisa. Entretanto há casos onde existem informações adicionais que podem ser introduzidas na geração da população inicial, minimizando o tempo de computação.

Na população inicial, pode-se estipular qual será o número máximo e mínimo de unidades binárias que um determinado cromossomo poderá possuir. Esse número é algumas vezes um fator limitante, determinado pelo método utilizado para calcular a resposta, associada a performance de determinado cromossomo.

iii - **Avaliação da resposta ou ' Aptidão'.**

Deve-se encontrar o valor associado à performance de cada cromossomo, relacionada ao sistema de interesse. A resposta de cada cromossomo é o resultado mais importante no procedimento do algoritmo genético.

A resposta (aptidão) é uma característica intrínseca ao indivíduo. Esta indicará qual o nível de habilidade que um indivíduo possui para sobreviver a predadores, pestes e a outros obstáculos para a subsequente reprodução. Transportando para o algoritmo matemático, seria sua habilidade de produzir a melhor resposta. Portanto será a resposta a responsável direta pela vida ou morte dos indivíduos.

Em alguns algoritmos é a resposta que irá determinar qual a probabilidade com que um determinado cromossomo irá cruzar. Portanto desta forma determinados indivíduos poderão realizar mais de um cruzamento durante uma mesma geração.

iv - **Cruzamento**.

Representa a permuta entre o material genético de diferentes indivíduos. A partir de uma população inicial, uma nova população é formada pelo cruzamento aleatório entre os cromossomos. Portanto os filhos recebem informações de seus progenitores, através do material genético proveniente deste cruzamento.

Durante a permuta de material genético entre os indivíduos, haverá uma tendência de que os indivíduos que possuem características dominantes, transmita-as para as gerações futuras. Após algumas gerações pode-se observar que uma elevada taxa dos indivíduos possuem a presença de variáveis '*dominantes*'. O cruzamento é a etapa responsável pela convergência levando para a situação de otimização desejada.

O processo de cruzamento dá-se pela escolha aleatória dos indivíduos e da posição onde ocorrerá a ruptura seguida pela troca de material genético. O número de rupturas geralmente varia entre 1 e 2. Isto deve-se ao fato de que quanto maior for o número de rupturas menor será a semelhança entre os pais e os filhos, dificultando a convergência.

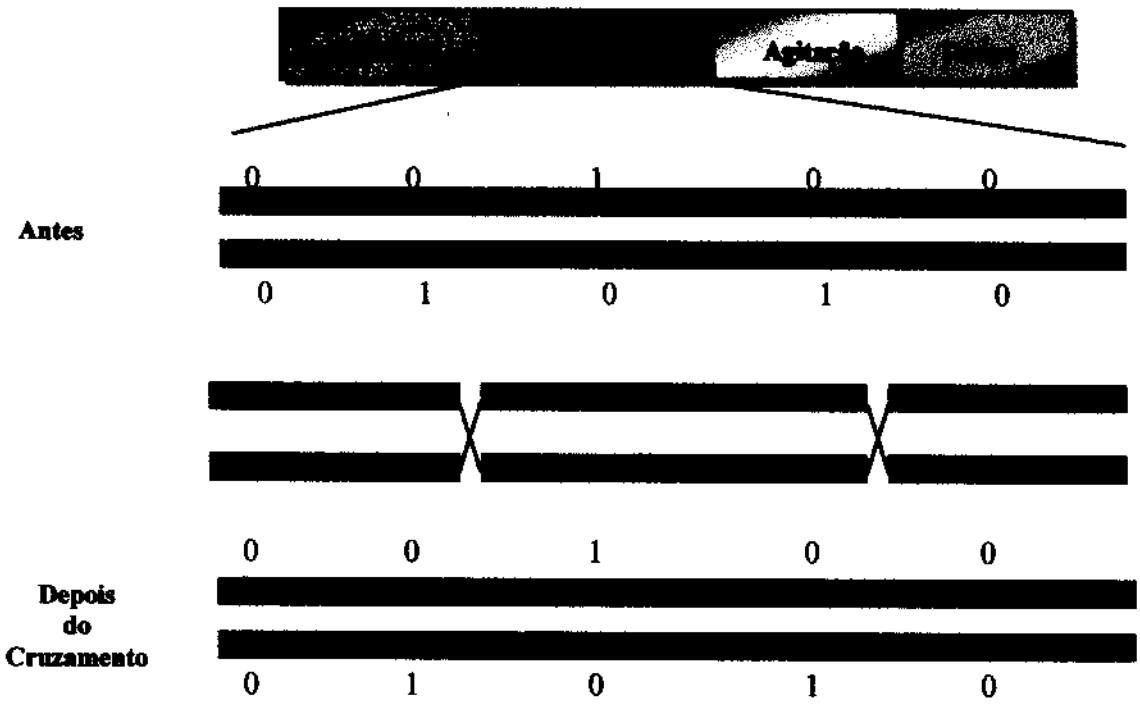


Figura 13. Operação de cruzamento em um cromossomo.

A Figura 13 mostra a etapa do cruzamento entre dois indivíduos, para o gene (pedaço do cromossomo) relacionado à variável catalisador, a qual está sendo otimizada.

v - **Mutação.**

Durante o processo de reprodução pode ocorrer uma perturbação das informações contidas dentro dos genes ocorrendo uma mutação.

A mutação promove alterações no código genético de uma pequena parcela da população, onde muitas vezes pode atuar de forma benigna. Se a

alteração gerar características melhores, estas serão transmitidas para os demais indivíduos ao longo das próximas gerações.

A mutação ajuda a solucionar o problema do confinamento a mínimos locais, pois promove alterações que direcionam a pesquisa para fora desta região. A Figura 15 representa dois cromossomos, onde ocorreu mutação no material genético de um deles. A mutação na representação binária é realizada pela troca de um por zero e vice-versa, conforme pode ser notado na Figura 14.

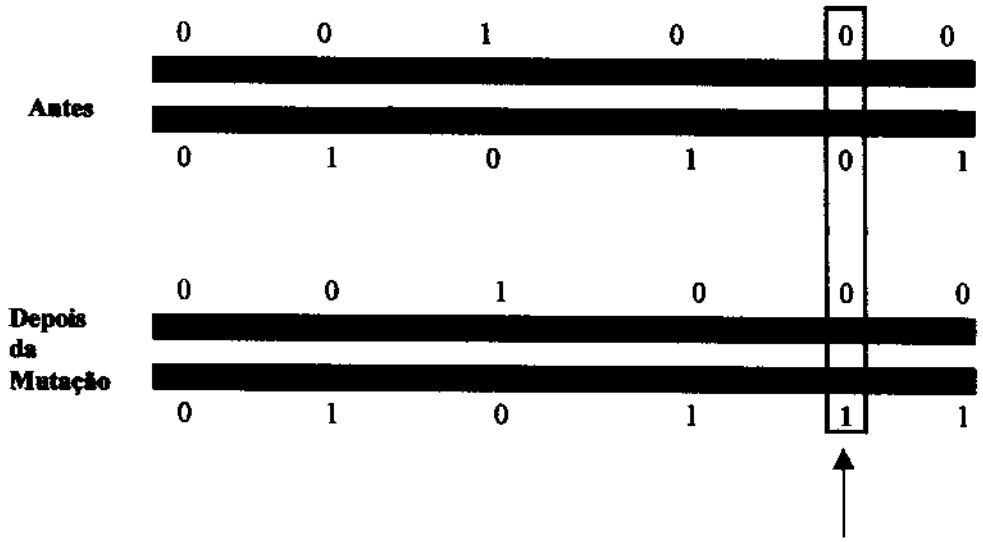


Figura 14. Operação de mutação nos cromossomos após o cruzamento.

vi - **Principais Etapas do Algoritmo Genético básico**.

A Figura 15 mostra as principais etapas do algoritmo genético básico. O processo é continuamente repetido até que uma condição seja satisfeita. Como critério de parada para o algoritmo genético normalmente adota-se como sendo um número máximo de gerações ou até que um erro máximo desejado seja alcançado.

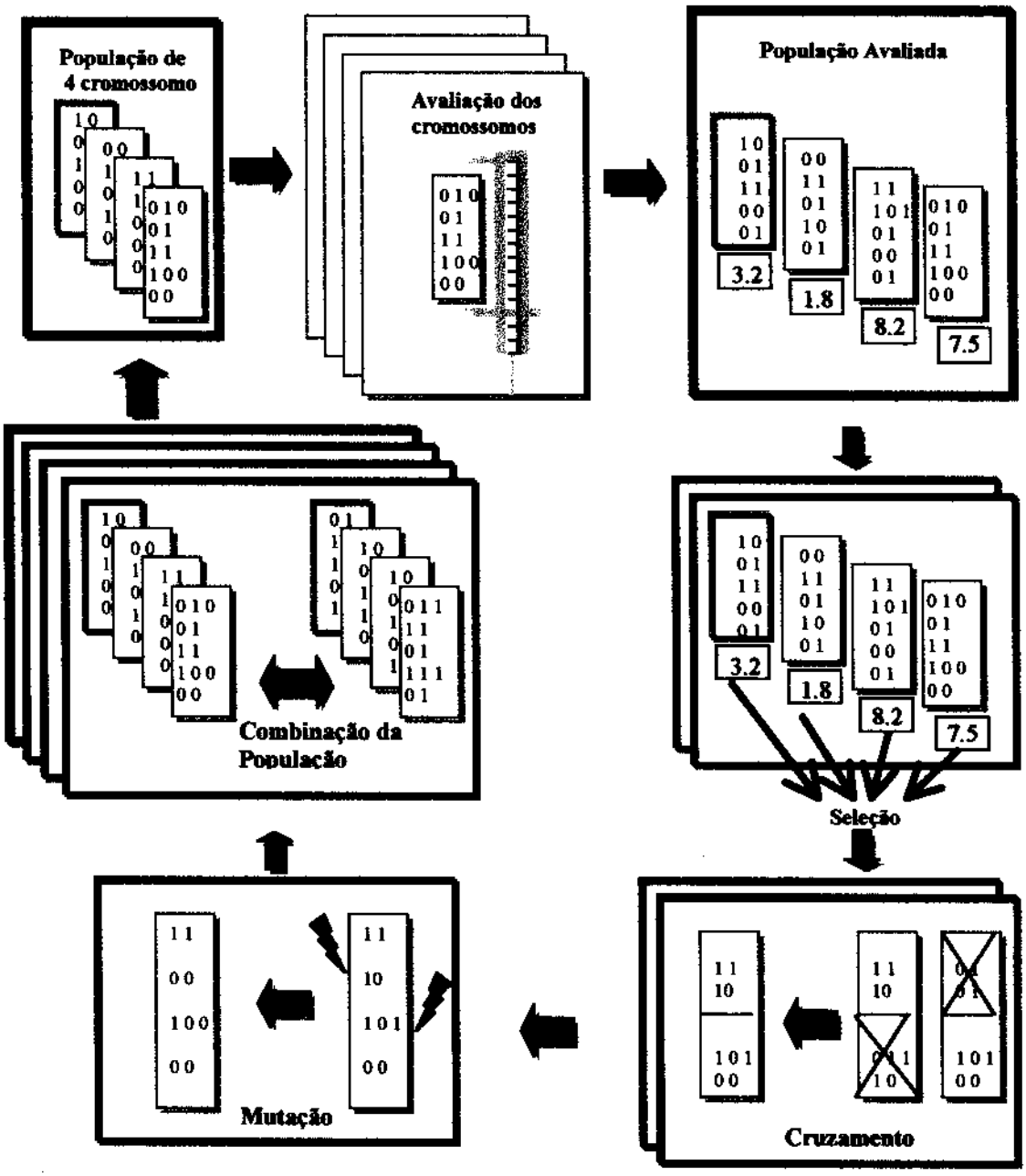


Figura 15. Principais etapas do algoritmo genético básico.

Para uma boa performance do algoritmo faz-se necessário otimizar parâmetros como, probabilidade de mutação, probabilidade de cruzamento, tamanho da população inicial e número de gerações.[6] O tamanho da população inicial é escolhido entre 50-500 membros. Normalmente utiliza-se probabilidade de mutação de aproximadamente 1% , probabilidade de cruzamento entorno de 90%. O número de gerações irá depender do tamanho da população inicial. [6]

2.3 Algoritmo Genético em Otimização.

A otimização envolve a maximização ou otimização de uma função ou uma série de variáveis num processo. Dentre os métodos de otimização comumente utilizados podem ser destacados o método de otimização Simplex [24], a Tempera simulada (Simulated Anneling)[10], Algoritmo Genético, e os Métodos Clássicos de Gradiente[25].

O estudo desses métodos de otimização é importante pois, sistemas químicos geralmente podem ser complexos quando trabalham com um grande número de variáveis. Isto torna muitas vezes impossível ou impraticável sua otimização em processos de tentativa e erro.

Em ordem, para ter-se oportunidade de encontrar um ótimo global em sistemas complexos, o algoritmo deve ter alguns mecanismos inteligentes, para que possam caminhar em direção a região de melhor performance.

De acordo com o grau de informações que se possui de um determinado sistema, os métodos de otimização podem ser classificado em - métodos fracos ou fortes. Os métodos fracos são aqueles onde se fazem poucas aproximações. Este é caracterizado por exigir um alto tempo de processamento, contudo é um método mais robusto[20].Os métodos fortes, dedicados ou tailorianos assumem um grande número de aproximações. Este método apresenta como grande vantagem o curto tempo de processamento, contudo fornece um modelo frágil, tendo uma estreitada faixa de aplicação.[20]

Nenhum dos métodos de otimização garante que a resposta encontrada seja a melhor, portanto esta resposta é chamada de solução estimada. Apesar

de não ser a melhor solução verdadeira, esta apresenta resultados suficientemente bons para muitos propósitos experimentais.[6]

O algoritmo genético encontra-se em uma faixa intermediária entre os métodos de otimização fracos e fortes - sendo considerado moderado. Pois este concilia a vantagem de ser um método robusto e gastar um médio tempo de processamento[6].

O diferencial do algoritmo genético encontra-se no fato deste realizar uma pesquisa paralela na superfície de resposta e não uma pesquisa ponto a ponto como os demais. Isto confere ao algoritmo genético uma maior robustez.[6] As regras de seleção usadas pelo algoritmo genético podem ser classificadas em probabilisticas e não determinadas.[6] Há essencialmente quatro fatores que diferenciam o algoritmo genético das técnicas convencionais de otimização: [6,20]
i - Pesquisa a partir de uma população e não de um simples ponto; ii - Pesquisa via amostragem; iii - Pesquisa usando operadores estocásticos (regras não deterministas).

2.4 Operações e Técnicas Avançadas.

Após o desenvolvimento do primeiro algoritmo genético inúmeras modificações foram propostas com o intuito de aperfeiçoar a performance e robustez do programa.

Os primeiros algoritmos eram compostos de basicamente quatro etapas: geração da população inicial, cruzamento, mutação e avaliação. Atualmente em programas mais sofisticados pode-se encontrar inúmeros outros operadores genéticos, como :

- Dominância
- Distinção de sexo :
  - . Homossexuais
  - . Heterossexuais
  - . Assexuados
- Incestos

Dominância: Quando um determinado gene é responsável por uma característica que faz com que um indivíduo se destaque entre um determinado grupo. Desta forma, torna-se interessante a preservação desta informação, e a propagação para gerações futuras, podendo acelerar o processo de convergência.

O operador de dominância possui a função de localizar e preservar os genes responsáveis pelo destaque em sua performance, que um determinado cromossomo apresenta em relação aos demais. Isso é implementado no algoritmo matemático, fazendo com que após um determinado número de gerações, inicia-se uma sub-rotina que emparelhe dois cromossomos selecionados aleatoriamente e crie um novo indivíduo que possua as características dominantes dos progenitores. Esta tipo de operação pode ser observada conforme esquema apresentado, segundo a Figura 16, onde os ***zeros*** representam as informações dominantes e os ***uns*** as recessivas. Portanto além da população proveniente do cruzamento há também o acréscimo de mais estes indivíduos à população final desta geração.

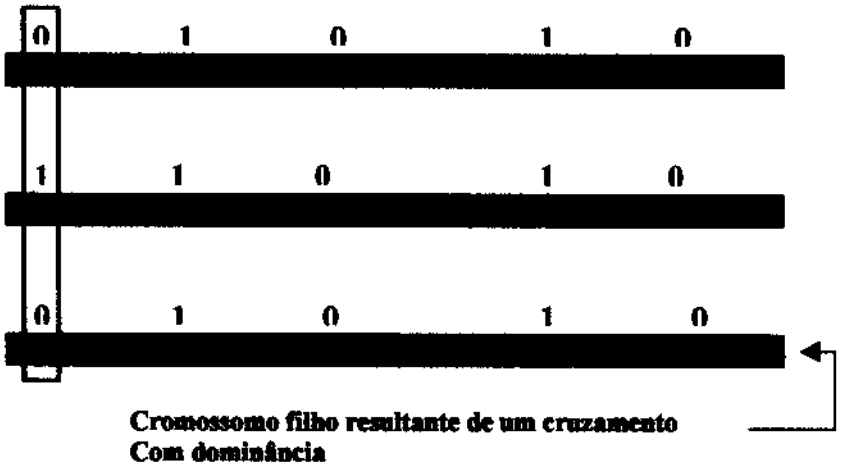


Figura 16. Figura ilustrativa de um cruzamento com dominância.

Distinção de Sexos: O cruzamento é a operação onde são gerados novos indivíduos a partir de cromossomos pais. O cruzamento pode apresentar-se de três formas distintas : [6]

. Homossexuais;

. Heterossexuais;

. Assexuado.

O cruzamento por distinção de sexo é realizada através de um bit reservado para distinguir o sexo dos cromossomos, como por exemplo representando-se um indivíduo do sexo masculino como tendo o bit "*1*", em uma determinada posição do cromossomo. A partir deste bit é possível controlar o cruzamento entre os indivíduos.[6]

Nos cruzamentos homossexuais ocorre o cruzamento entre cromossomos de mesmo sexo e nos heterossexuais ocorre o cruzamento entre cromossomos de sexo distintos. Já no cruzamento assexuado ocorre troca de material genético com o próprio indivíduo.

O controle pela distinção de sexo faz com que a diversificação seja maior, já que determinadas características provavelmente, somente um determinado sexo apresentará. Isso pode ser observado na natureza, já que apenas seres inferiores são hermafroditas.

Incestos: O cruzamento entre indivíduos muito semelhantes leva a geração de um cromossomo muito parecido com os dos pais. Desta forma não há nenhuma melhora significativa para a nova geração[6]. Pode-se cria uma sub-rotina que evite o incesto, durante a etapa de cruzamento. O incesto é mais provável ocorrer após várias gerações.

Computação paralela: Programas onde existem um grande número de operadores genéticos tornam-se demorados, uma vez que muitas operações devem ser realizadas. Para de tornar a execução do programa mais rápida, tem-

se proposto o uso da computação paralela, minimizando o tempo de computação. A computação paralela é caracterizada por: [26-28]

i - Um ambiente onde há múltiplos processadores;

ii - Uma rede interligando os processadores para a distribuição de partes do programa a serem executadas em diferentes maquinas simultaneamente;

iii - Um ambiente que possibilite criar e manipular um processamento paralelo:

iv - Um algoritmo paralelo e um programa paralelo.

A construção de um programa em paralelo requer que o programador siga os seguintes passos, conforme ilustrado na Figura 17:

i- Decomposição do algoritmo, ou dos dados, em várias partes;

ii - Distribuição dos pedaços para os diferentes processadores que estão trabalhando simultaneamente;

iii - Coordenação do trabalho e da comunicação entre esses processadores;

iv - Considerações :

- Tipo da arquitetura paralela;
- Tipo da comunicação entre processadores.

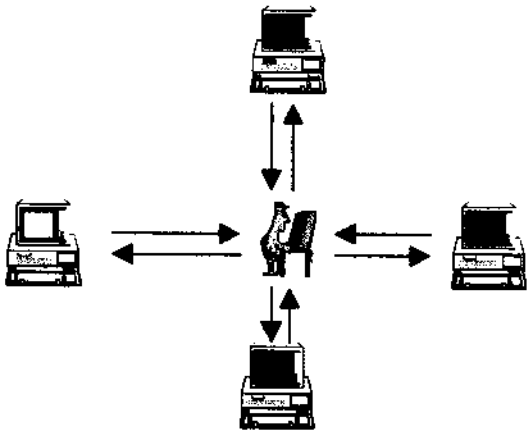


Figura 17. Ilustração de uma rede de computadores interligada na troca de informações.

No algoritmo genético a computação paralela pode ser implementada por exemplo na avaliação da resposta de cada indivíduo, conforme Figura 18.

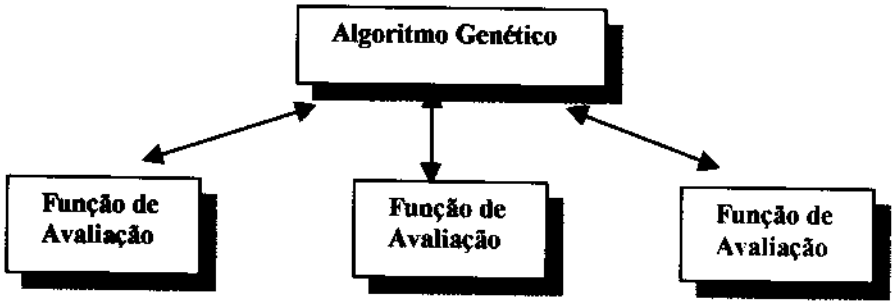


Figura 18. Esquema da distribuição de um programa.

2.5 Aplicação do Algoritmo Genético em Química.

Na química tem-se observado uma acentuada tendência na aplicação de algoritmo genético em inúmeros problemas de otimização de sistemas complexos essencialmente devido à robustez e versatilidade do algoritmo.

Na química analítica tem-se utilizado o algoritmo genético para otimização de parâmetros internos (pesos) de rede neural [29]. Uma de suas aplicações é o treinamento da rede para análise quantitativa de espectros de fluorescência de raios-X.[30] Outra aplicação bastante difundida é a seleção de comprimentos de onda em determinações espectroscópicas, principalmente no infravermelho próximo, utilizando calibração multivariada[31]. Geralmente nestes casos utiliza-se o método dos mínimos quadrados parciais[31] ou a regressão linear múltipla[32] para a construção de um modelo que possa correlacionar os espectros com as espécies de interesse. Este tem como objetivo determinar o número mínimo de comprimentos de onda que possuam informação relevante para a análise de interesse.

O algoritmo genético também tem sido utilizado no mapeamento de átomos, afim de se determinar a mínima distância química[33]. Outro campo que está sendo explorado é a pesquisa de planejamentos experimentais, que consiste no planejamento de *n* experimentos, onde há variação de alguns fatores para otimizar o sistema.[34]

Muitos sistemas químicos apresentam um alto nível de complexidade, sendo muitas vezes necessário o uso de uma função para explicá-los. Para solucionar este tipo de problema utiliza-se uma poliotimização. Um exemplo de poliotimização em química analítica é a otimização de distâncias entre picos cromatográficos, como função das condições cromatográficas (composição da fase móvel, temperatura do injetor, valor de pH, etc.)[35]. Pode-se encontrar outro exemplo de poliotimização na determinação de glicose em sangue humano por fotometria, baseado em algoritmo genético em poliotimização. A otimização envolve 12 parâmetros químicos e tecnológico, com o intuito de monitorar seis propriedades ópticas. O resultado obtido foi comparado com os dos métodos clássicos regressão dos mínimos quadrados parciais não linear [35]. Encontra-se ainda exemplos da aplicação em determinação da constante de estabilidade de sistemas de cloreto de cádmio em meio perclorado com chumbo e éter coroa diclorohexil-18-coroa-6, a partir de dados calorimetricos e polarográficos[36]. Na seleção e detecção de amostras anômalas dentre um conjunto de amostras, para construção do modelo de calibração.[37-38]. Na otimização da separação de base fracas ou soluções acidas por cromatografia líquida de fase reversa com variação de pH e modificação orgânica na fase movel[39]. Na otimização da estimativa dos parâmetros cinéticos de uma reação[40]. Para a otimização da intensidade linhas de emissões atômicas de traços de elementos em pó de alumina como uma função da condição de excitação.[41] Análise conformacional de biopolimero.[42]

2.6 Algoritmo Genético na seleção de variáveis.

Tem-se observado nos últimos anos uma tendência na implementação de técnicas espectróscopicas em análises químicas rotineiras. Isto deve-se essencialmente à simplicidade, rapidez, baixo custo, versatilidade, bem como outros atributos de diferenciação proporcionados pela técnica.

A calibração multivariada contribuiu de forma significativa para a consolidação desta técnica, viabilizando a determinação simultânea de mais de uma espécie mesmo em situações adversas, onde os métodos convencionais de análises seriam difíceis ou impossíveis de serem realizados (por exemplo determinações "on-line" em processos). Tornou também possível análises de cunho qualitativos e quantitativos, bem como a detecção de amostras anômalas.

Em análises espectrofotométricas, muitas das aplicações de calibração multivariada deste método envolvem um elevado número de variáveis. Muitas vezes o uso de somente algumas das variáveis que carregam mais informações, podem proporcionar maior segurança e facilidade na interpretação do modelo. Assim as variáveis (absorbância em diferentes comprimentos de onda) que possuem uma menor relação sinal/ruído ou/e não linearidade podem ser descartadas.

Atualmente o número de algoritmos que propõem-se a selecionar variáveis é pouco expressivo. Dentre estes destacam-se a regressão linear múltipla por passos (stepwise multiple linear regression)[14] e os que empregam o uso do algoritmo genético.

A implementação do algoritmo genético na seleção de variáveis difere-se das demais aplicações no que tange a codificação do problema e a função de resposta, já que as outras etapas permanecem inalteradas.

Na codificação do problema em seleção de variáveis de espectros, considera-se que o cromossomo possui p genes, onde cada gene representa uma das variáveis do espectro (absorbância em cada comprimentos de onda).

Assim sendo, o cromossomo terá o mesmo número de variáveis contidas no espectro, conforme representada na Figura 19.

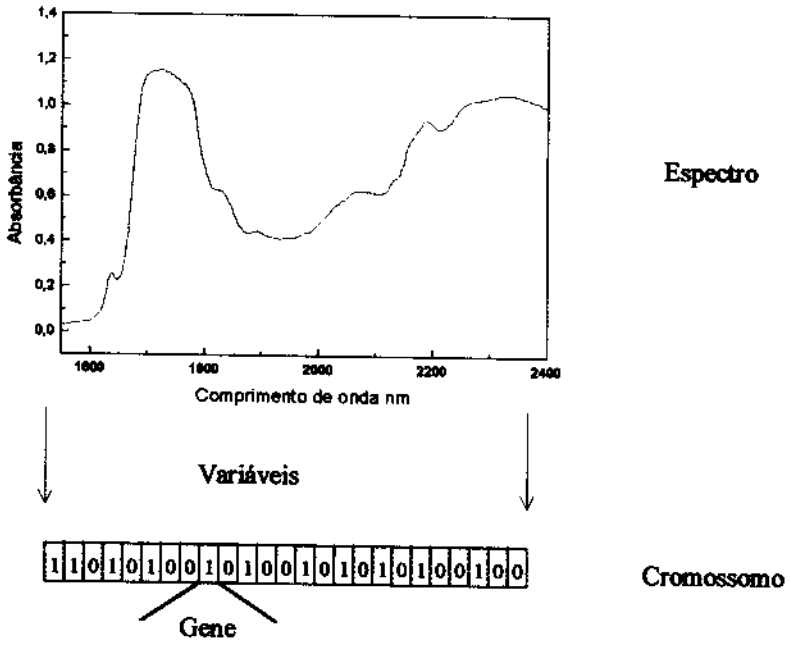


Figura 19. Codificação de um espectro na forma de um cromossomo.

Quando utiliza-se o algoritmo genético para a seleção de variáveis, é necessário, dependendo do modelo de calibração, utilizar algumas "condições de contorno ", como por exemplo pré-estabelecer o número máximo de variáveis selecionadas, caso da regressão linear múltipla.

Quando a seleção de variáveis é feita em mais de uma dimensão, o problema pode ser codificado na forma de um cromossomo, apenas emparelhando-se as varias dimensões a serem otimizadas, conforme apresentada na Figura 20, onde temos o caso especial para dados bilineares. Este procedimento viabiliza a otimização simultânea de todas as dimensões de interesse.

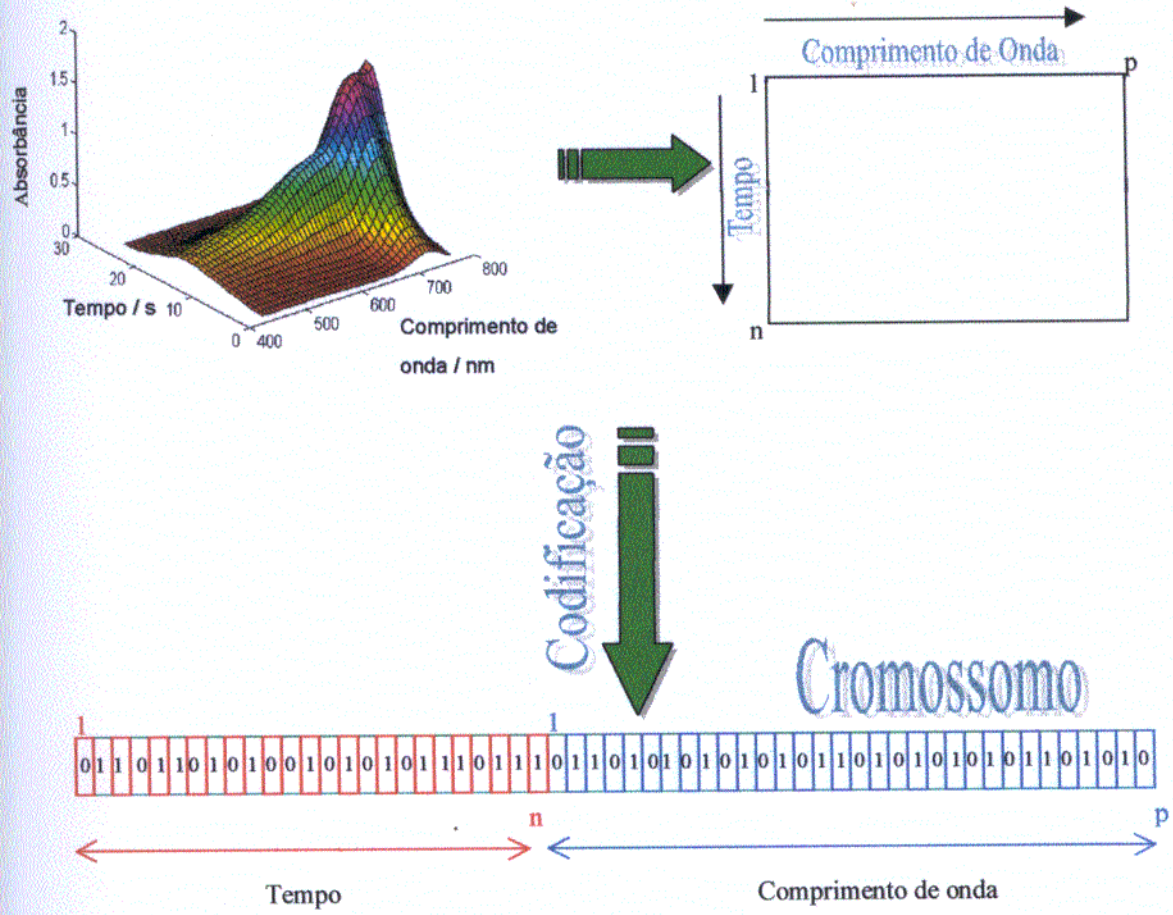


Figura 20. Codificação de uma matriz de dados bilineares na forma de um cromossomo.

Na seleção de variáveis utiliza-se também o auxílio do código binário (0,1) para codificar o problema. Cada gene pode assumir o valor ***um*** ou ***zero***, sendo que quando a posição referente a uma determinada variável for igual a ***um***, isto implica na seleção da variável. Por outro lado, se a posição conter o valor ***zero***, isto indica que a variável não foi selecionada.

CAPÍTULO III

Aplicação do Algoritmo Genético na Seleção de Variáveis.

3.1 Desenvolvimento dos Programas.

Um dos objetivos do trabalho concentrou-se no desenvolvimento de um programa para o Algoritmo Genético em Matlab 4.2c. O programa desenvolvido consistiu na construção de um programa principal (apêndice A) e sub-rotinas para a geração de uma população inicial, cruzamento sexuado em duas ou três partes e cruzamento assexuado. Os indivíduos para o cruzamento são selecionados de forma aleatória, entretanto cada uma das modalidades foi atribuída uma probabilidade para ser selecionada. Para o cruzamento assexuado foi atribuída uma probabilidade 6%, para o cruzamento em duas partes 74% e em três partes entre 20%. Também existem sub-rotinas para a mutação, dominância, incesto, avaliação dos cromossomos e eliminação de indivíduos semelhantes.

3.2 Aplicação Teórica.

A seguir, será apresentada uma aplicação teórica do processo de otimização do algoritmo genético em um sistema de três dimensões.

O intuito desta aplicação é possibilitar a visualização do processo de otimização, avaliar a performance do algoritmo e observar o comportamento adotado para superar obstáculos (máximos e mínimos locais), afim de atingir o objetivo proposto.

3.2.1 Exemplo de funcionamento do Algoritmo Genético.

Uma maneira simples para exemplificar o funcionamento e a performance do algoritmo é sua aplicação na determinação do ponto máximo da função[25]:

$$Z = (3\times(1-X)^2\times\exp(-X^2-(Y+1)^2)) - (10\times(\frac{X}{5}-X^3-Y^5)\times\exp(-X^2-Y^2))$$
$$-(\frac{1}{3}\times\exp(-(X+1)^2-Y^2)+\frac{3}{5}\times(\frac{17}{10}\times\text{sen}(pi\times X)+\frac{17}{10}\times\cos 2\times Y))$$

com o domínio de X e Y entre -3 e +3, conforme pode ser observado na superfície de resposta da Figura 21. Neste caso pretende-se determinar quais os valores de X e Y que, dentro do domínio especificado, irão produzir o maior valor de Z.

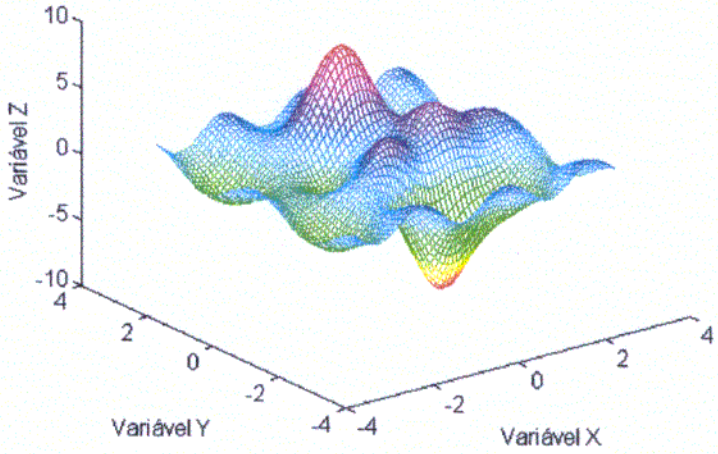


Figura 21. Superfície correspondente a função:

$$Z = (3 \times (1 - X)^2 \times \exp(-X^2 - (Y+1)^2)) - (10 \times (\frac{X}{5} - X^3 - Y^5) \times \exp(-X^2 - Y^2))$$
$$- (\frac{1}{3} \times \exp(-(X+1)^2 - Y^2) + \frac{3}{5} \times (\frac{17}{10} \times \text{sen}(pi \times X) + \frac{17}{10} \times \cos 2 \times Y))$$

Como se pode observar, trata-se de uma função complexa, com inúmeros máximos locais e um máximo global.

Para a otimização da função, foi criado um cromossomo com dois genes, sendo cada um destes relacionado a uma variável da função (valores de X e Y). O próximo passo foi representar cada gene por código binário de 16 bits, obtendo uma precisão de 0,00003 unidade. Foi utilizado mais um bit para representar o sinal do número.

Desta maneira, cada cromossomo é representado por 34 unidades binárias, onde os 17 primeiros estão relacionados ao valor de X e os 17 últimos (outro gene) ao valor de Y. Assim cada cromossomo (que representa os valores de X e Y), pode ser associado a uma resposta que é representada pelo valor correspondente de Z da função. O objetivo do algoritmo genético é então encontrar qual cromossomo que irá produzir o maior valor de Z possível, dentro do domínio especificado.

Pelo fato dos cromossomos possuírem uma dimensão relativamente pequena, foi criados 10 cromossomos e 25 gerações. A Figura 22 mostra a melhor resposta na população inicial e os respectivos valores de melhor resposta para as próximas gerações deste exemplo.

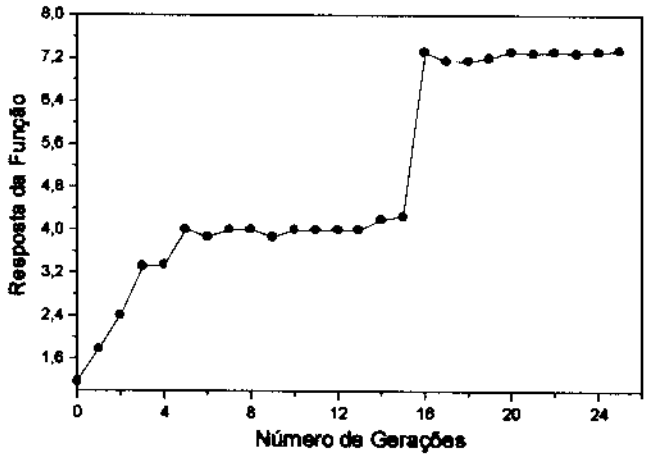


Figura 22. Resposta dos melhores indivíduos de cada geração.

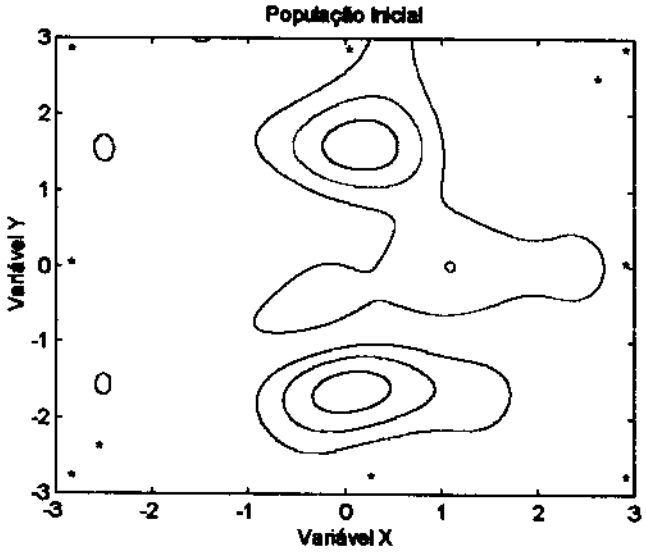
No gráfico da Figura 22 são observados dois patamares: - O primeiro patamar encontra-se relacionado à determinação de um máximo local, o qual permaneceu como sendo uma situação de pseudo-ótimo global por 10 gerações. Após a 15ª geração observa-se a convergência para as proximidades do máximo global que se encontra caracterizado pelo segundo patamar da Figura 22.

Para uma melhor exemplificação da performance do algoritmo, no decorrer do processo, optou-se por apresentar a visualização do processo de pesquisa em algumas das gerações, conforme ilustrado nas Figuras 23.a e 23.b. Pode-se observar que na Figura 23.a a população inicial apresentou uma resposta de pouco interesse de acordo com as expectativas desejadas. Outro fator que deve ser salientado é o fato dos indivíduos (cromossomos) apresentarem-se bem distribuídos de forma quase homogênea ao longo do espaço amostral e ao mesmo tempo distante das regiões de máximo. Na nona geração a pesquisa é realizada basicamente na região próxima de um máximo local.

Na Figura 23.b pode-se observar que na décima quinta geração esta em uma região de máximo local, próxima a região do máximo global. Na vigésima

quinta geração, o máximo global foi encontrado. Entretanto observa-se que a pesquisa ao longo do espaço amostral ainda persiste, contudo em menor proporção quando relacionada ao número de indivíduos que se encontram ocupados em determinar o valor exato do suposto máximo global.

I)



II)

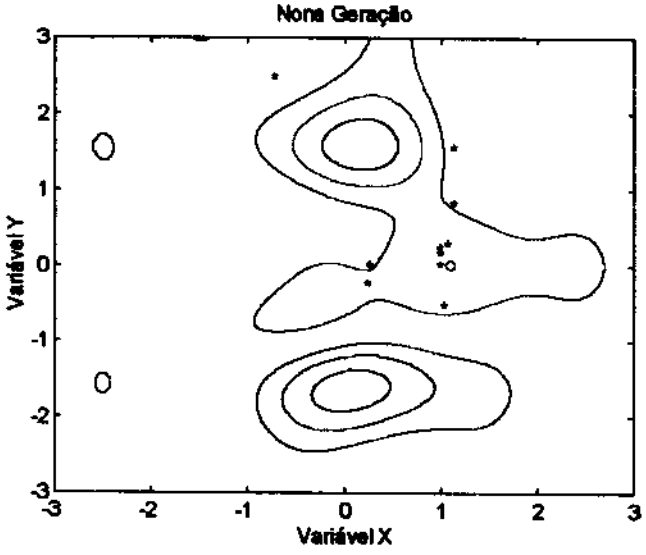
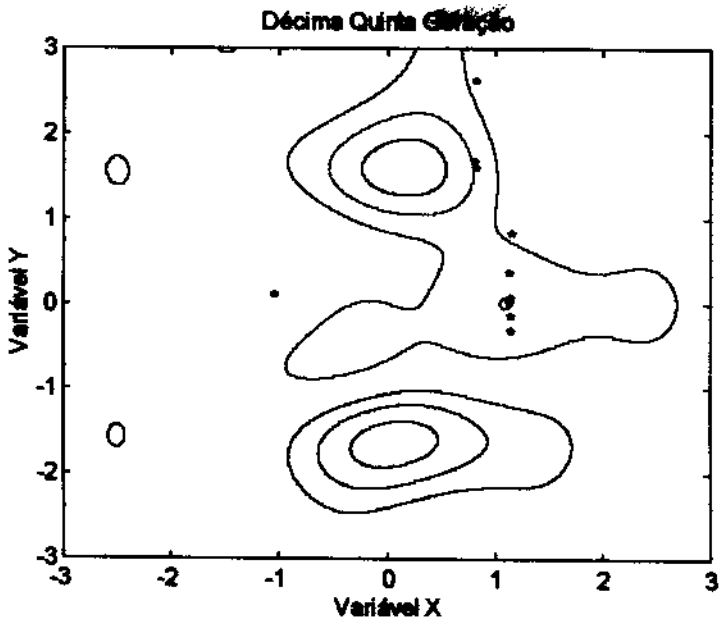


Figura 23.a. Superfície de contorno da Figura 22, onde os asteriscos indicam a posição dos indivíduos na geração.

I) População Inicial

II) Nona Geração

I)



II)

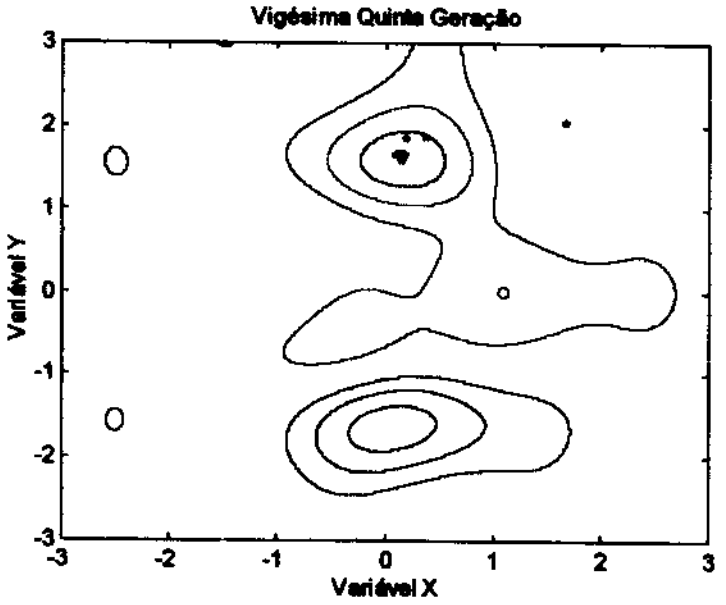


Figura 23.b. Superfície de contorno da Figura 22, onde os asteriscos indicam a posição dos indivíduos na geração.

I) Décima Quinta Geração

II) Vigésima Quinta Geração

A Figura 24 mostra o percurso realizado sempre pelos melhores indivíduos em cada geração na otimização da função Z. Este se inicia no lado superior direito da superfície de contorno e encontra-se traçado pela linha em azul com os asteriscos em preto sobre a superfície de contorno.

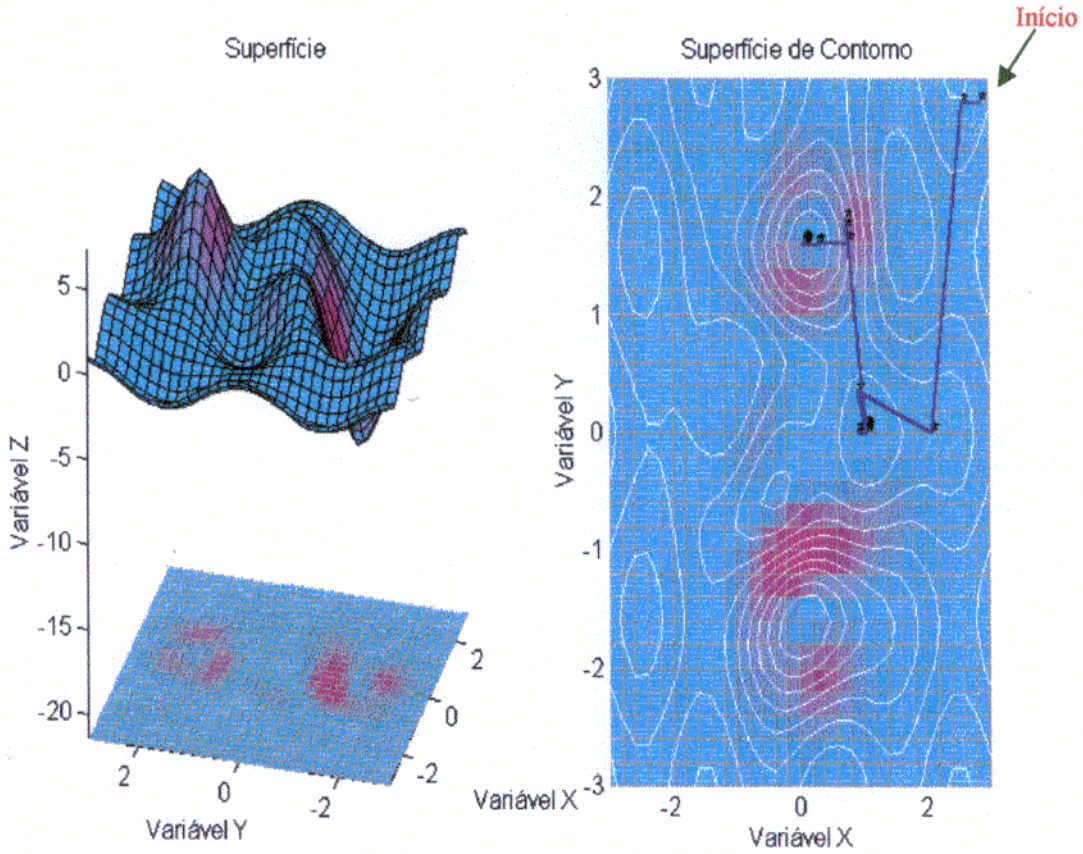


Figura 24. Percurso realizado sobre a superfície de contorno durante a otimização da função Z.

Através deste exemplo teórico é possível observar que o algoritmo genético é capaz de encontrar uma região bem próxima do máximo global, mesmo em sistemas onde existem vários máximos locais, evidenciando desta forma a eficiência do algoritmo.

3.3 Determinação de Açúcares.

Neste tópico será tratada a seleção de variáveis na determinação simultânea de açúcares em solução. Inicialmente será apresentada uma breve introdução teórica da técnica de espectroscopia no infravermelho utilizando a reflexão total atenuada (ATR), posteriormente a descrição do procedimento experimental e finalmente a discussão dos resultados obtidos.

3.3.1 Introdução.

A espectroscopia estuda a interação da radiação eletromagnética com a matéria, sendo um dos seus principais objetivos o estudo dos níveis de energia de átomos ou moléculas. No caso de moléculas, a região espectral onde estas transições são observadas dependem do tipo de nível envolvido: eletrônico, vibracional ou rotacional[43]. Normalmente as transições eletrônicas são situadas na região do ultravioleta ou visível, as vibracionais na região no infravermelho e as rotacionais na região de microondas (em casos particulares também no infravermelho longínquo).

Historicamente uma das primeiras aplicações da espectroscopia no infravermelho como ferramenta analítica, foi durante o período da segunda guerra mundial[43]. Nesta ocasião, ela foi usada no controle de qualidade em várias indústrias químicas. Após o desenvolvimento da instrumentação da cromatografia gasosa, a espectroscopia no infravermelho restringiu-se basicamente a aplicações qualitativas, entretanto este quadro tem sido revertido nos últimos anos, após a consolidação das técnicas quimiométricas. A calibração multivariada, em especial viabilizou o desenvolvimento de novas metodologias de análises, envolvendo técnicas espectroscópicas com resultados tão bons quando dos obtidos por cromatografia. Além disso, a união entre a calibração multivariada e a espectroscopia vem acelerando o tempo de análises e diminuindo os custos associados à preparação, reagentes, resíduos, etc.

Entre as técnicas espectroscópicas temos a absorbância e reflectância. As técnicas de reflectância geralmente são usadas com amostras que não podem ser analisadas pelo método normal de transmitância como: polímeros, plásticos, madeiras, metais, papeis, couros, colas, etc[44].

A reflectância pode ser subdividida em:

- Reflectância especular ;
- Reflectância total atenuada (ATR);
- Reflectância interna múltipla (MIR).

Essencialmente a técnica de reflectância mede a radiação refletida a partir de uma superfície a qual pode ser difusa ou especular. Entretanto, o que se mede geralmente é a mistura de ambos fenômenos[44,45].

A reflexão total atenuada (ATR) é uma técnica que se baseia no fenômeno da reflexão total da radiação na interface de matérias com índices de refração diferentes (Figura 25). Quando o índice de refração de um cristal de ATR (n_2) é maior que o índice de refração da amostra (n_1) e a luz incidente no infravermelho excede o ângulo crítico α_c. Isto faz com que a luz seja totalmente refletida na interface amostra / elemento de reflexão. Entretanto no ponto de reflexão, o feixe no infravermelho penetra fracamente dentro da amostra, em conseqüência da interferência entre o feixe de luz incidente e refletido, formando uma onda padrão na superfície de reflexão[44]. A profundidade de penetração é da ordem de poucos micrômetros e respeita a equação (29)[46].

$$d_p = \frac{\lambda}{2\pi n_1 \sqrt{\text{sen}^2\,\theta - \frac{n_2^2}{n_1^2}}} \qquad (29)$$

Onde: d_p é definido como distância necessária para causar um decréscimo na amplitude do campo elétrico por um fator de e^{-1}.

n_2 é o meio de maior índice de refração (cristal)

n_1 é o meio de menor índice de refração (amostra)

θ é o ângulo de incidência

λ é o comprimento de onda da radiação

O equivalente caminho óptico b_e para celas de multi-reflexões pode ser obtido a partir do número de reflexões N por $b_e = d_e * N^{(47)}$.

Onde:

d_e é definido como a espessura de uma cela de transmissão que possua a mesma absorbância de uma única reflexão sobre o ATR.

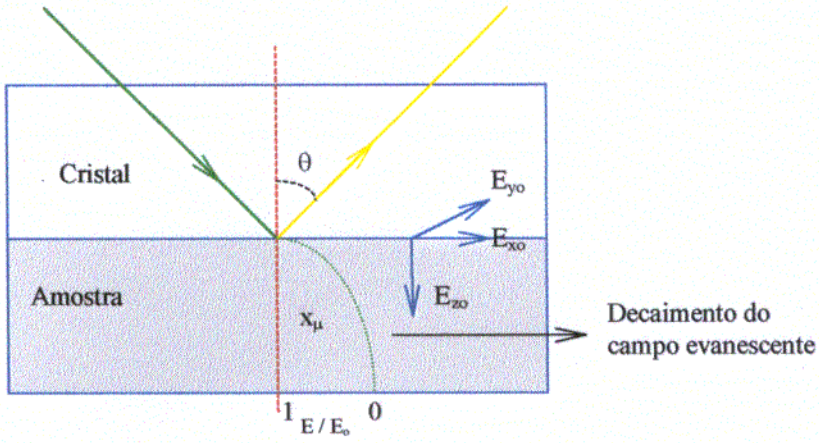


Figura 25. Desenho esquemático de um feixe de radiação infravermelha incidente sobre um cristal de ATR.

O campo elétrico que existe na amostra é chamado de onda evanescente, e sua amplitude decresce exponencialmente, respeitando a seguinte equação $E = E_0 e^{(-x/dp)}$ onde[47]:

E_0 é a amplitude do campo elétrico inicial

x é a distância a partir da superfície do ATR

d_p a profundidade da luz penetrada na amostra, esta é determinada a partir da equação 29.

Se a amostra absorve em um certo comprimento de onda evanescente, há uma redução na intensidade do feixe refletido emergente do elemento de reflexão, proporcionando uma atenuação da reflexão detectada, originando o espectro de ATR.

A técnica é freqüentemente empregada para evitar interferentes na impressão digital do espectro de filmes finos. Tem grande utilidade para examinar materiais densos ou com alta absorção onde a transmissão não é possível[44-48].

Uma limitação do ATR é que o método requer que a amostra tenha um contato íntimo com o elemento de reflectância interna. O trabalho quantitativo utilizando o ATR é dificultado pelo fato que a refletividade em um dado comprimento de onda é função do índice de refração da amostra, ângulo de polarização de incidência de radiação, bem como da concentração de absorção da espécie. O índice de refração pode variar durante a reação de uma solução concentrada, e desta forma o cálculo da concentração a partir da medida de reflectância pode ser dificultada ou impossibilitada. Há também uma dificuldade na precisão da determinação do ângulo de incidência de reflexão interna[44-48].

Apesar da técnica apresentar um elevado número de desvantagens, esta é aconselhada em situações especificas, onde uma ou mais vantagens superam em muito algumas das possíveis desvantagens existentes. Este tipo de problema é comumente verificado em situações onde o solvente (água) comporta-se como interferente, por apresentar estiramentos característicos na mesma região de absorção da espécie de interesse, inviabilizando a análise pelos métodos tradicionais[48].

3.3.2 Determinação de açúcares em solução por ATR.

Hoje em dia nota-se uma nítida preocupação no desenvolvimento de metodologias de análises "on-line" que sejam rápidas, gerem pouco resíduos, sejam não destrutivas e simples, para o monitoramento de sistemas químicos.

Em análises de celulose e açúcares em solução, já se tem realizado o estudo de espécies químicas através da reflexão total atenuada. O interesse da análise da celulose deve-se ao fato desta ser o composto orgânico mais abundante na biosfera, e por ser o principal polímero estrutural de plantas. A celulose é constituída de cadeias lineares de unidades de glicose. Em células animais a glicose é armazenada na forma de glicogênio, o qual serve como

reservatório nutricional para as células. Além de parte integrante e essencial do reservatório nutricional dos animais (em animais superiores é essencial a presença de glicose no sangue) e estrutural das plantas, possui um grande interesse industrial, visto o grande número de aplicações destas espécies químicas.

Os açúcares sacarose e glicose são obtidos em escala industrial a partir de vários tipos de plantas. Mais do que 100.000.000 toneladas de sacarose são produzidas por ano em todo mundo, quantidade comparável a produção de ácido sulfúrico em 1989 nos EUA que foram de 43.400.000 toneladas.

Neste trabalho foi realizada a determinações quantitativa de três açúcares em solução glicose, maltose e frutose. As estruturas moleculares dos respectivos açúcares encontram-se apresentadas a seguir.

3.3.2.1 Estruturas Moleculares.

As Figuras 26, 27 e 28 representam a estruturas dos açúcares utilizados. A partir destas ilustrações é possível observar a semelhança entre as estruturas das espécies estudadas.

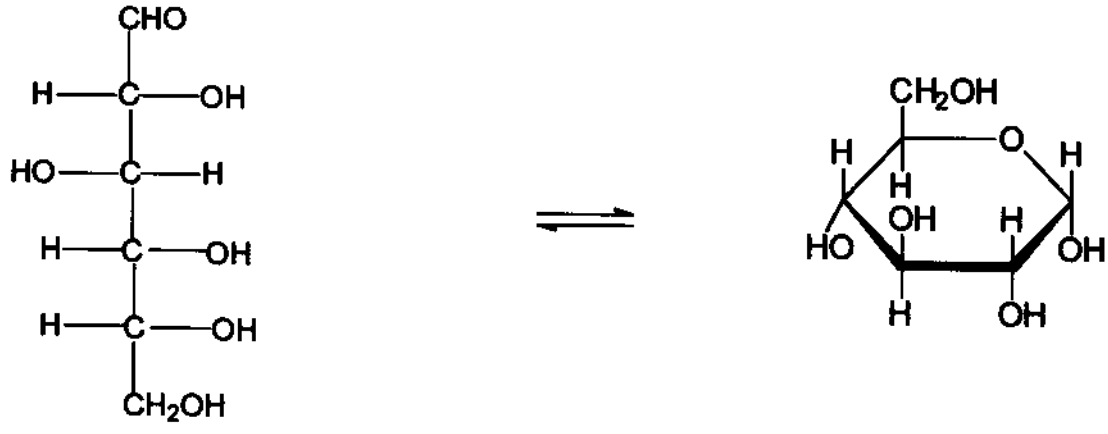


Figura 26. Estrutura da Molécula de Glicose.

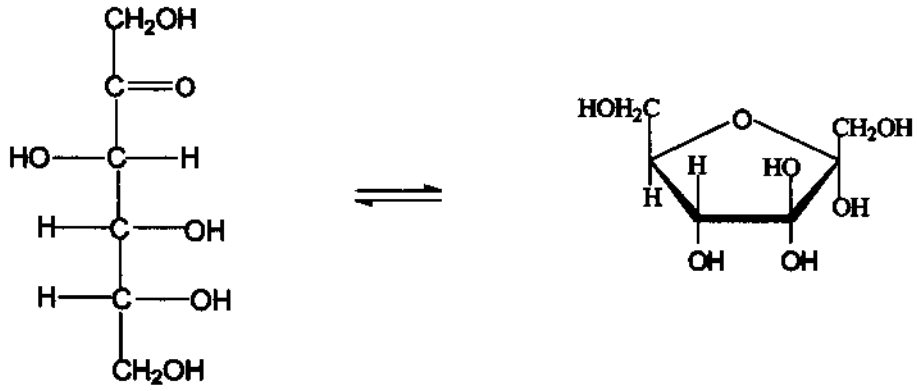


Figura 27. Estrutura da Molécula de Frutose.

As formas predominantes de glicose e frutose em solução não são cadeias abertas, mas sim na forma de anéis[48].

Na maltose, duas unidades de glicose unem-se por uma ligação glicosídica α-1,4. A maltose é originada da hidrólise do amido que por sua vez, é hidrolisada a glicose pela maltase[49].

Tanto a glicose, maltose e frutose tem o espectro no infravermelho similar, porque o módulo vibracional do anel da glicose é o responsável principal pelo perfil dos espectros[50].

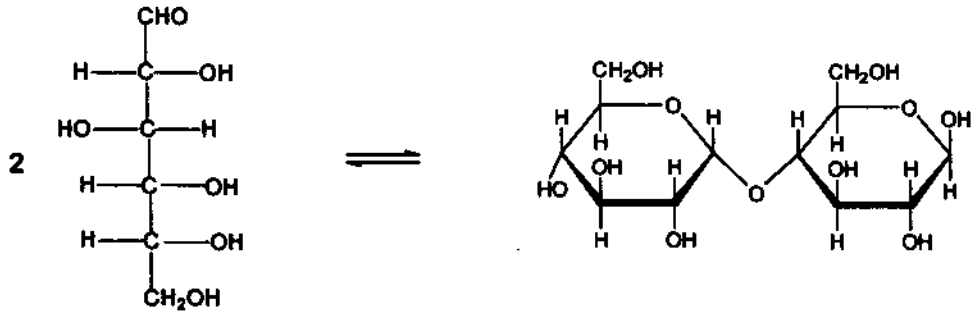


Figura 28. Estrutura da Molécula de Maltose.

3.3.3 Procedimento Experimental

A fase preliminar de preparação das amostras envolveu um planejamento experimental para a preparação das 64 misturas de soluções contendo glicose, maltose e frutose na faixa de 21,99 a 28,28 gL^{-1} ; 15,99 a 28,30 gL^{-1} e 9,97 a 16,30 gL^{-1} , respectivamente, por pesagem e posterior diluição.

Para a preparação das soluções optou-se por realizar 3 planejamentos experimentais distintos, para obter-se um número ideal de amostras para o experimento. Para isso foi preparado um planejamento experimental com 3 níveis e três variáveis totalizando um número total de 27 experimentos, conforme ilustrado na Figura 29.

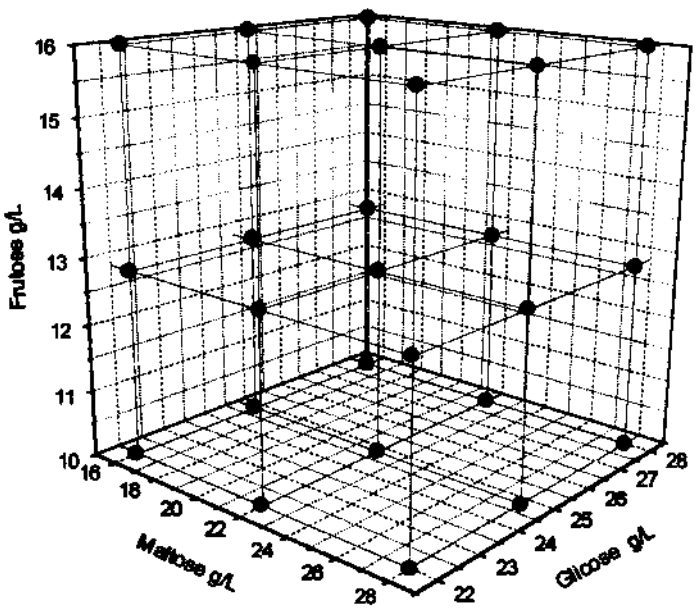


Figura 29. Representação do planejamento experimental da mistura de Açúcares.

Em seguida foi realizado um segundo planejamento na forma de um cubo menor (Figura 30) contendo 14 amostras, que se encontra no interior do primeiro cubo (Figura 29).

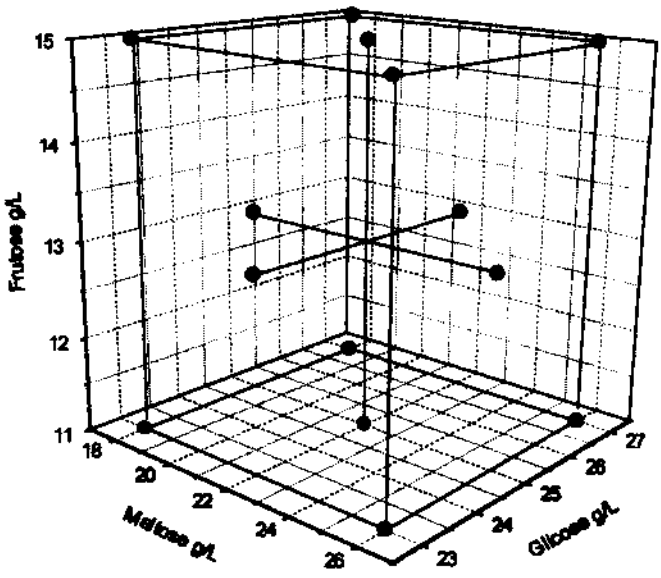


Figura 30. Representação do planejamento experimental da mistura de Açúcares.

O terceiro planejamento envolveu a preparação de mais oito amostras, formando um terceiro cubo menor (Figura 31), o qual encontra-se no interior dos dois primeiro cubos (Figuras 29 e 30).

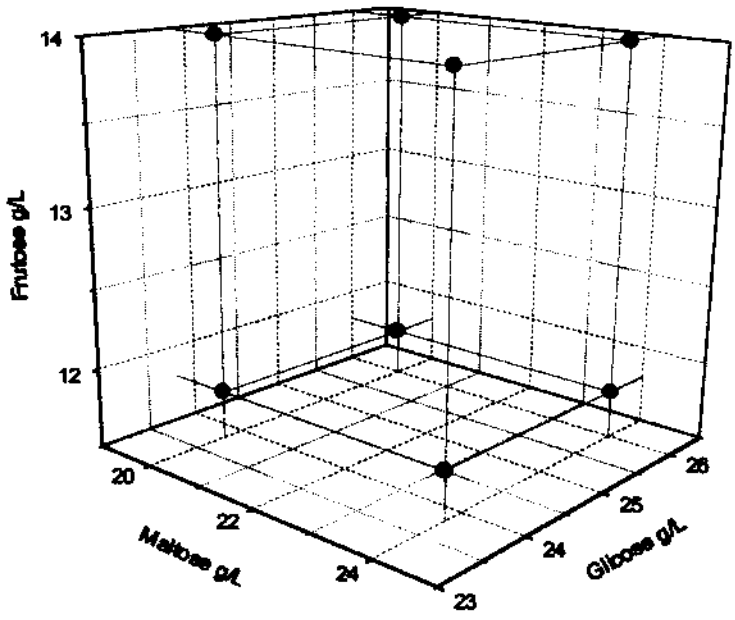


Figura 31. Representação do planejamento experimental da mistura de Açúcares.

Para alcançar o total de 64 amostras, foram preparadas mais 15 misturas das soluções dos açucares de forma aleatória dentro do limite superior (cubo maior) e limite inferior (cubo menor).

A aquisição dos espectros foi realizada por um espectrofotômetro de infravermelho Nicolet 520 FT-IR com a técnica de reflexão total atenuada (ATR), empregado um acessório do tipo bote com um cristal de seleneto de zinco. Os espectros foram obtidos nas seguintes condições: purgado com nitrogênio; 32 leituras por espectros; resolução de 2 cm^{-1} e região de aquisição de 4000 a 400 cm^{-1}.

A Tabela 1 apresenta a concentração em gL^{-1} presente nas amostras empregadas para a construção do modelo de calibração, sendo que a segunda amostra (com asterisco) foi eliminada do conjunto.

Tabela 1. Amostras usadas para a construção do modelo de calibração.

Amostra	Maltose / gL^{-1}	Glicose / gL^{-1}	Frutose / gL^{-1}
1	22,04	28,04	16,04
2*	28,30	28,11	16,00
3	16,00	28,06	12,86
4	22,05	28,00	12,91
5	27,92	28,18	13,01
6	15,99	24,84	12,89
7	22,10	24,84	12,89
8	27,97	24,82	12,92
9	27,94	28,13	10,04
10	28,21	24,80	10,06
11	18,25	27,00	14,98
12	18,04	26,98	10,98
13	17,98	22,98	10,99
14	22,27	24,99	15,01
15	22,03	26,98	13,00
16	24,02	26,08	14,30
17	20,02	26,01	14,08
18	24,10	23,82	11,81
19	25,01	23,46	11,42
20	19,00	23,41	14,20
21	25,01	26,24	14,21
22	18,08	22,53	10,44
23	17,65	27,20	10,47
24	25,40	26,83	10,41
25	25,41	26,80	15,03

A Tabela 2 apresenta a concentração em gL^{-1} presente nas amostras utilizadas para a construção da validação do modelo de calibração. Neste conjunto foi necessário a eliminação das amostras 1 e 2 (com asterisco). As amostras utilizadas na fase de validação são empregadas para a seleção das variáveis do modelo.

Tabela 2. Amostras usadas para a construção da validação do modelo.

Amostra	Maltose / gL^{-1}	Glicose / gL^{-1}	Frutose / gL^{-1}
1*	16,04	24,88	16,16
2*	28,10	24,86	16,19
3	16,08	22,03	16,23
4	22,02	22,16	15,96
5	16,00	22,14	13,05
6	28,09	21,99	12,94
7	16,15	28,28	9,97
8	22,06	22,09	10,22
9	26,24	22,98	14,99
10	18,30	23,00	14,99
11	22,00	25,00	10,99
12	22,02	23,01	13,03
13	18,21	24,96	13,00
14	24,12	25,98	11,80
15	24,97	26,24	11,44
16	24,96	23,44	14,22
17	18,04	22,39	15,08
18	19,65	22,78	15,08
19	17,62	27,22	15,10
20	28,16	22,10	10,21
21	26,00	26,99	15,00

Os métodos de inteligência artificiais como o algoritmo genético, exigem o uso de um terceiro grupo de amostras (conjunto teste) que não tenha sido usado na etapa de seleção das variáveis. Isso porque durante a seleção de variáveis o algoritmo pode selecionar variáveis que apresente bons resultados somente para o caso particular das amostras do conjunto de validação, não podendo ser aplicada a outras amostras. Com o uso de um terceiro conjunto este tipo de problema pode ser detectado.

A Tabela 3 apresenta a concentração em gL^{-1} presente nas amostras utilizadas para o teste do modelo de calibração. Neste conjunto foi necessário a eliminação da amostra 2 (com asterisco).

Tabela 3. Amostras usadas para o teste do modelo

Amostra	Maltose / gL^{-1}	Glicose / gL^{-1}	Frutose / gL^{-1}
1	16,09	28,04	16,19
2*	22,02	24,89	16,20
3	27,96	22,05	16,16
4	22,18	22,01	13,00
5	22,12	28,07	10,00
6	16,06	24,89	10,29
7	22,02	24,95	10,04
8	16,08	22,02	10,12
9	26,18	26,98	11,00
10	26,05	22,99	10,98
11	26,06	25,00	13,04
12	24,00	23,78	14,00
13	20,02	23,82	14,09
14	20,03	26,20	11,83
15	19,04	23,39	11,42
16	19,00	26,18	11,42
17	18,99	26,31	14,21
18	19,72	22,80	10,41

3.3.3.1 Parâmetros de Configuração do Algoritmo Genético.

- Tamanho da população Inicial: 100 cromossomos
- Número de Gerações: 100
- Probabilidade de Cruzamento: 90 %
- Probabilidade de Mutação: 1 %
- Erro máximo para a finalização do processo: 1%
- Número máximo de variáveis selecionadas: 10
- Número de componentes principais: 4
- Método selecionado: PLS e RLM

3.3.4 Resultados e discussão.

3.3.4.1 Determinação do Número ideal de Componentes Principais e de Amostras Anômalas.

Fazendo a análise visual da coleção de espectros obtidos dos açúcares (Figura 32), observa-se a presença de 4 amostras com absorbância superior a 0,7, com comportamento diferenciado da coleção de espectros.

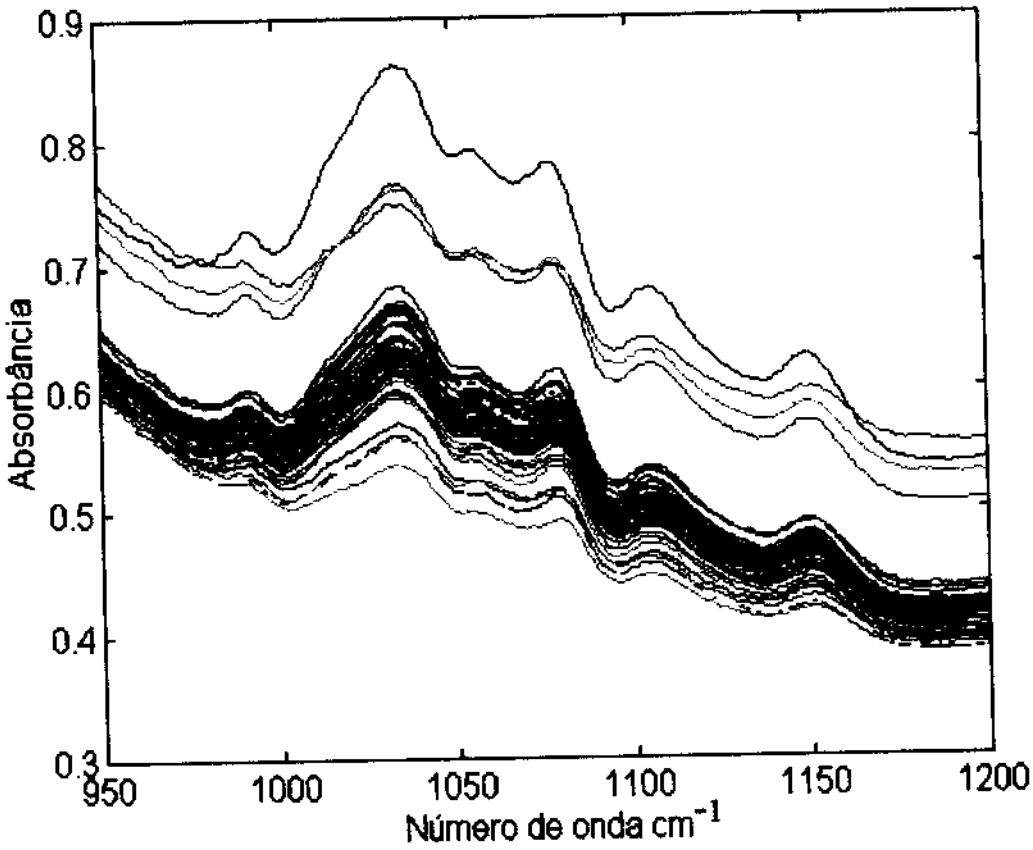


Figura 32. Coleção de espectros no infravermelho das 64 soluções de açúcares.

Realizando uma análise das componentes principais[13,14] (Figura 33), verifica-se que as amostras com absorbância acima de 0,7 podem ser classificadas como amostras anômalas (provavelmente devido algum erro durante a pesagem dos açúcares). Estas amostras foram eliminadas do estudo de

quantificação de açúcares, pois além de terem um perfil diferente das demais, podem apresentar comportamento não linear[51].

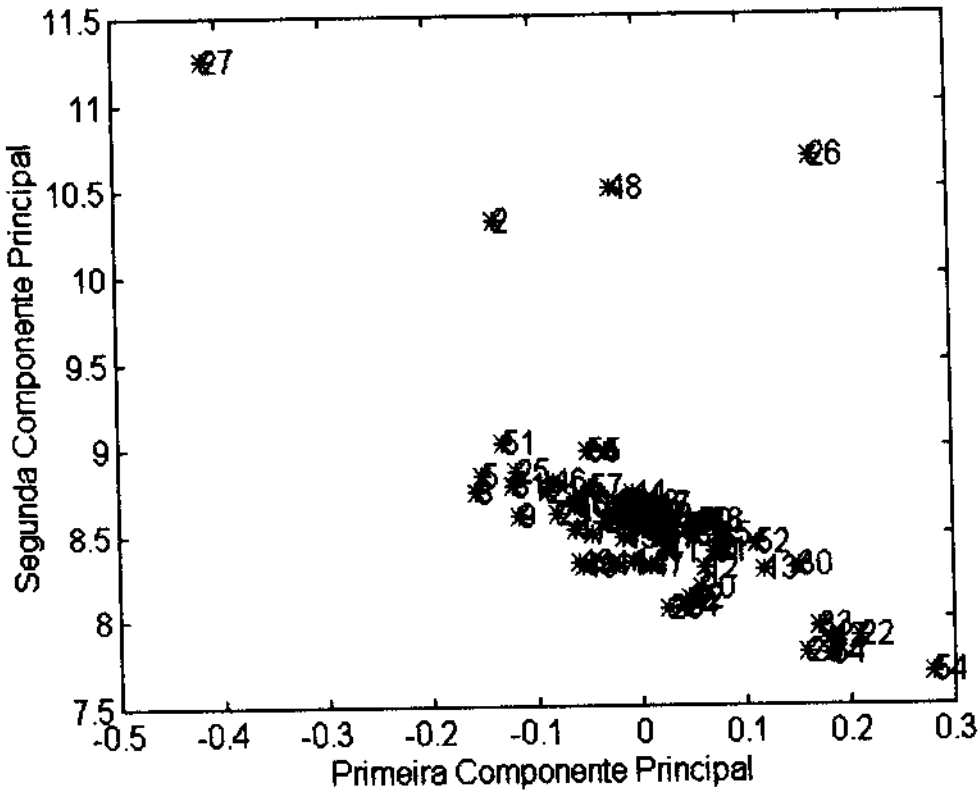


Figura 33. Análise das componentes principais das 64 amostras de açúcares.

As sessenta amostras restantes foram separadas em três grupos: 24 amostras para o desenvolvimento do modelo de calibração, 19 amostras de validação e seleção de variáveis e 17 amostras para testar o modelo, sendo somente utilizado após ter sido definido quais variáveis seriam utilizadas. Para a análise quantitativa dos açúcares, utilizou-se o modelo dos mínimos quadrados parciais (PLS) e da regressão linear múltipla (RLM), com seleção de variáveis pelo algoritmo genético.

Um estudo prévio utilizando-se validação cruzada ("cross-validation" [52]), indicou que o número ideal das componentes principais seria quatro, conforme

Figura 34. Em todo o estudo realizado para os açúcares, empregou-se quatro componentes principais, inclusive durante a fase de seleção de variáveis.

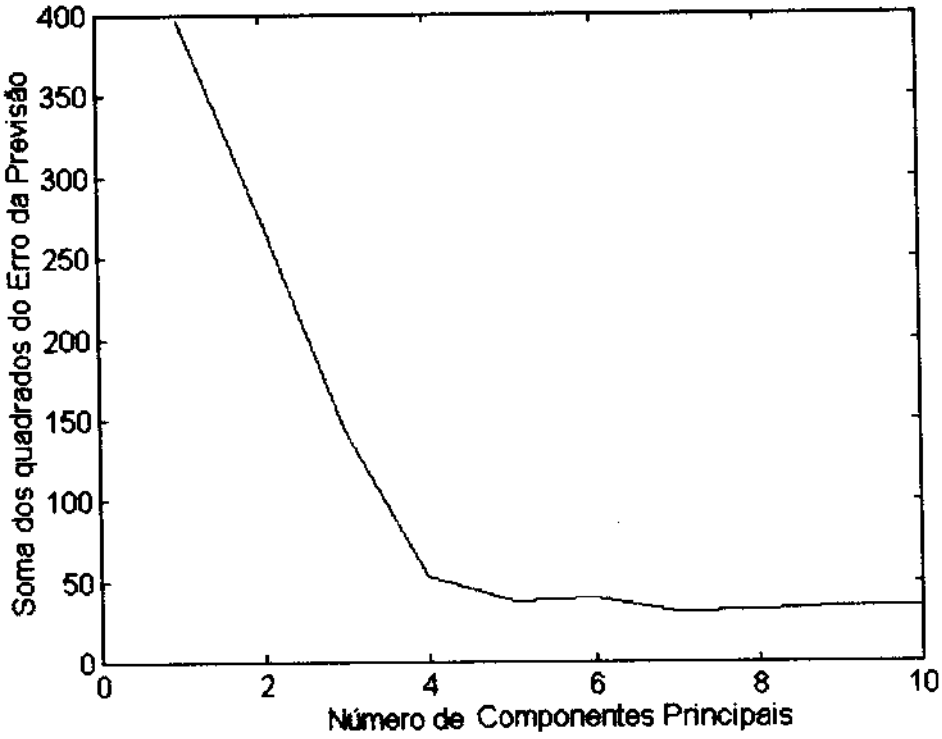


Figura 34. Determinação do Número das Componentes Principais ideais, utilizando validação cruzada.

3.3.4.2. Método dos Mínimos Quadrados Parciais na Construção do Modelo, na Faixa espectral entre 950 a 1200 cm^{-1} com 260 pontos.

Em uma primeira análise, optou-se por não realizar um pré-tratamento dos dados, para avaliar a performance dos dois métodos de calibração multivariada (Mínimos Quadrados Parciais ' PLS ' e Regressão Linear Múltipla ' RLM ') em situações mais extremas.

A Tabela 4 apresenta os erros relativos para o conjunto teste do modelo de calibração, utilizando a região dos espectros compreendida entre 950 a 1200 cm^{-1}. As linhas mais escuras da Tabela 4 procuram destacar as amostras que apresentaram erro relativo superiores a 10 %.

Além do erro relativo, foi utilizado o erro padrão de previsão (SEP) como, um segundo parâmetro para avaliar o erro obtido durante as fases de validação e teste do modelo construído[54]. O erro padrão de previsão é comumente empregado, já que representa um erro médio do modelo com as mesmas unidades da propriedade estimada, e é mais sensível a presença de amostras com erros elevados.

O erro padrão de previsão é calculado como: $SEP = \sqrt{\frac{\sum_{i=1}^{i=n} (\hat{y}_i - y_i)^2}{n}}$

Onde y_i é o valor real, $\hat{y}_i$ representa o valor estimado pelo modelo e n é o número de amostras.

A Tabela 5 apresenta o erro padrão de previsão para o conjunto de validação e teste.

Tabela 4. Erros relativos do conjunto teste.

Amostra	Maltose / %	Glicose / %	Frutose / %
1	4,74	0,51	0,36
2	17,18	-21,53	21,58
3	5,01	-8,75	5,45
4	-24,28	5,20	-33,18
5	-20,71	2,78	-15,70
6	-8,28	1,86	-10,06
7	7,51	-3,59	8,01
8	-3,52	8,79	-7,21
9	3,97	-6,06	1,71
10	3,62	-1,73	4,71
11	7,64	-7,93	5,47
12	1,86	-7,12	0,21
13	-9,26	2,96	-9,05
14	-7,67	-5,78	-4,21
15	-4,67	0,52	-5,30
16	-1,78	-0,79	-2,79
17	8,53	-4,36	8,25

Tabela 5. Erro padrão de previsão dos conjuntos de validação e teste.

		Maltose	Glicose	Frutose
Conjunto de Validação	SEP (gL^{-1})	2,10	2,67	1,11
Conjunto Teste	SEP (gL^{-1})	2,24	1,67	1,38

3.3.4.3. Algoritmo Genético na Seleção de Variáveis no Método dos Mínimos Quadrados Parciais, sem Pré-Tratamento dos Dados.

Em uma fase posterior, utilizou-se o algoritmo genético para selecionar o melhor conjunto de variáveis para o método dos mínimos quadrados parciais (PLS). Neste estudo utilizou-se o espectro na faixa entre 950 a 1200 cm^{-1} e sem nenhum pré-tratamento dos dados.

Devido a dificuldade de determinar o número de componentes principais durante o processo de seleção de variáveis, optou-se por mantê-la fixo em quatro. Para determinar o número das componentes principais, seria necessário realizar uma validação cruzada para cada cromossomo gerado, tornado praticamente inviável o processo, devido ao tempo requerido de processamento.

No final do processo de seleção das variáveis, os seguintes comprimentos de onda foram escolhidos como sendo o melhor conjunto que minimiza o erro no processo de validação do modelo de calibração: 1036; 1038; 1051; 1062; 1109; 1117; 1118; 1120; 1148; 1168 cm^{-1}.

A Tabela 6 apresenta os erros relativos para as amostras do conjunto teste, utilizando o algoritmo genético para selecionar o melhor conjunto de variáveis a serem empregadas no modelo gerado pelo PLS. A Tabela 7 traz os resultados obtidos para o SEP do conjunto de validação e teste.

Tabela 6. Erros relativos do conjunto teste.

Amostra	Maltose / %	Glicose / %	Frutose / %
1	2,61	3,62	1,70
2	8,23	-2,33	14,02
3	-4,17	-1,17	-1,16
4	-16,36	-11,02	-11,35
5	-12,89	-6,32	-11,32
6	-5,32	-4,83	-6,26
7	4,61	2,96	1,98
8	-0,81	-1,39	-1,68
9	3,40	-5,30	-3,61
10	3,39	-3,96	3,72
11	5,30	-4,25	0,98
12	-1,62	-3,58	-4,35
13	-2,16	-4,30	-6,18
14	-2,69	-8,53	-7,14
15	-3,26	4,27	-4,96
16	-1,73	-0,60	-3,29
17	12,16	-3,78	6,35

Tabela 7. Erro padrão de previsão dos conjuntos de validação e teste.

		Maltose	Glicose	Frutose
Conjunto de Validação	SEP (gL^{-1})	0,93	1,06	0,66
Conjunto Teste	SEP (gL^{-1})	1,45	1,26	0,82

Ao comparar os resultados obtidos para o PLS com e sem a seleção de variáveis, através das Tabelas 5 e 7, pode-se notar uma melhora significativa na estimativa da concentração dos açúcares.

3.3.4.4 Algoritmo Genético na Seleção de Variáveis para a Regressão Linear Múltipla, sem Pré-Tratamento dos Dados.

Foi realizada a quantificação dos açúcares utilizando-se regressão linear múltipla nos cálculos. A seleção de variáveis para o modelo da regressão linear múltipla apresentou bons resultados conforme apresentado na Tabela 8. Para este modelo foram selecionadas as seguintes variáveis: 975; 996; 1036; 1039; 1068; 1071; 1100; 1103; 1148; 1164 cm^{-1}.

A Tabela 8 apresenta o erro médio e a Tabela 9 o erro padrão de previsão encontrados para o conjunto de validação e teste, após a seleção de variáveis para a regressão linear múltipla.

Tabela 8. Erros relativos do conjunto teste.

Amostra	Maltose / %	Glicose / %	Frutose / %
1	1,93	0,34	4,68
2	4,81	2,91	9,19
3	-1,16	2,21	-0,71
4	-10,06	-9,74	-7,58
5	-5,94	-9,61	-8,27
6	-6,67	-2,68	0,44
7	0,94	3,37	-5,64
8	-0,15	3,16	-3,85
9	-0,79	-2,16	-5,03
10	7,08	1,03	6,03
11	3,59	-1,62	-3,73
12	-1,61	-2,72	-5,81
13	-8,52	-7,08	-9,84
14	-3,98	-5,60	-4,49
15	-3,97	4,72	-2,77
16	-6,13	-1,01	-3,45
17	0,17	-2,94	-1,58

Tabela 9. Erro padrão de previsão dos conjuntos de validação e teste.

		Maltose	Glicose	Frutose
Conjunto de Validação	SEP (gL^{-1})	0,67	1,06	0,56
Conjunto Teste	SEP (gL^{-1})	1,08	1,18	0,70

3.3.4.5. Método dos Mínimos Quadrados Parciais, com Pré-Tratamento dos Dados.

Nesta etapa realizou-se um pré-tratamento dos dados, para verificar se seria possível reduzir ou eliminar ruídos nos espectros, oriundos da aquisição dos dados, e conseqüentemente melhorar os resultados de previsão do modelo.

Como pré-tratamento dos dados utilizou-se o filtro de Savisky-Golay e a primeira derivada[54]. Basicamente o filtro de Savisky-Golay ajusta um polinômio a uma certa janela com um número fixo de pontos do espectro. Após isso, há um deslocamento de um ponto nesta janela, ajustando-se outro polinômio (da mesma ordem), e assim sucessivamente ocorre o deslocamento até o final do espectro. No caso da misturas de açúcares cada janela continha 31 pontos, e foi utilizado um polinômio de ordem dois.

Observando os resultados das Tabelas 10 e 11 verifica-se uma melhora na previsão da estimativa das concentrações dos açúcares.

Tabela 10. Erros relativos do conjunto teste.

Amostra	Maltose / %	Glicose / %	Frutose / %
1	4,93	1,71	0,85
2	6,45	-2,75	8,12
3	-2,40	-2,94	-1,04
4	-9,83	-9,53	-10,31
5	-13,69	-3,95	-6,72
6	-7,51	-0,96	-4,96
7	2,98	1,07	1,46
8	1,10	0,86	1,61
9	-0,32	-2,99	-3,27
10	1,25	0,88	3,54
11	1,61	-0,89	-1,58
12	-2,98	-3,08	-4,85
13	-5,06	-1,88	-3,70
14	-9,14	-5,32	-3,70
15	-2,72	-0,59	-3,21
16	-0,64	-1,61	-0,84
17	5,12	-1,39	3,20

Tabela 11. Erro padrão de previsão dos conjuntos de validação e teste.

		Maltose	Glicose	Frutose
Conjunto de Validação	SEP (gL^{-1})	1,03	1,05	0,52
Conjunto Teste	SEP (gL^{-1})	1,16	0,86	0,55

3.3.4.6. Algoritmo Genético na Seleção de Variáveis no Método dos Mínimos Quadrados Parciais, com Pré-Tratamento dos Dados.

A seleção de variáveis associada ao prévio tratamento dos dados apresentou uma melhora nos resultados. Analisando as variáveis selecionadas com pré-tratamento dos dados (1002; 1055; 1066; 1121; 1125; 1141; 1165 cm^{-1}) e sem o pré-tratamento, observa-se que as variáveis selecionadas após o pré-tratamento encontram-se concordantes com as variáveis selecionadas sem o pré-tratamento dos dados, já que os comprimentos de onda selecionados encontram-se em regiões bem próximas. Entretanto também se observa que foram necessárias algumas variáveis a mais para os dados sem o pré-tratamento, isso provavelmente encontra-se relacionado à presença de ruído e não-linearidades. Os resultados obtidos podem ser avaliados a partir das Tabelas 12 e 13.

Tabela 12. Erros relativos do conjunto teste.

Amostra	Maltose / %	Glicose / %	Frutose / %
1	7,65	1,60	3,23
2	3,41	-1,33	8,37
3	-1,98	-1,29	2,18
4	-8,20	-8,59	-10,80
5	-4,60	-7,47	-3,66
6	-3,01	-4,49	-2,80
7	-2,65	1,73	-2,34
8	2,89	-1,58	2,26
9	-1,02	-3,49	-4,93
10	3,15	-0,59	5,39
11	0,42	-0,58	-0,75
12	-1,57	-4,72	-4,06
13	-2,49	-3,63	-2,63
14	-2,92	-7,62	0,23
15	-2,96	-1,74	-4,94
16	-0,84	-2,39	-0,07
17	4,76	-2,24	2,66

Tabela 13. Erro padrão de previsão dos conjuntos de validação e teste.

		Maltose	Glicose	Frutose
Conjunto de Validação	SEP (gL^{-1})	0,77	1,03	0,44
Conjunto Teste	SEP (gL^{-1})	0,78	1,04	0,56

3.3.4.7 Algoritmo Genético na Seleção de Variáveis para a Regressão Linear Múltipla, com Pré-Tratamento dos Dados.

As Tabelas 14 e 15 apresentam os erros encontrados, após utilizar o modelo construído a partir das seguintes variáveis selecionadas: 973; 977; 985; 992; 1061; 1075; 1095; 1136; 1151; 1199 cm^{-1}.

Pode-se observar que neste caso obteve-se uma estimativa da concentração das amostras do conjunto teste sem que nenhuma delas tivesse erros relativos superiores a 10%.

Tabela 14. Erros relativos do conjunto teste.

Amostra	Maltose / %	Glicose / %	Frutose / %
1	2,46	6,38	3,74
2	7,01	-0,77	7,96
3	0,89	-3,85	1,29
4	-7,85	-9,62	-7,79
5	-3,23	-2,94	-6,81
6	-3,83	-2,86	-3,74
7	7,47	2,76	-0,84
8	0,43	1,13	1,27
9	-1,31	-7,02	-7,85
10	2,86	3,28	8,11
11	2,22	1,59	1,76
12	-4,66	-3,13	-4,01
13	0,58	-0,94	0,82
14	-7,26	-6,69	-2,45
15	-0,21	1,08	-3,58
16	-2,58	-1,95	-1,76
17	2,87	-3,05	1,19

Tabela 15. Erro padrão de previsão dos conjuntos de validação e teste.

		Maltose	Glicose	Frutose
Conjunto de Validação	SEP (gL^{-1})	0,69	0,99	0,48
Conjunto Teste	SEP (gL^{-1})	0,90	1,08	0,58

3.3.4.8. Análise dos Resultados Obtidos.

Para facilitar a interpretação dos resultados obtidos entre os diferentes modelos matemáticos, novas tabelas foram construídas contendo os valores de erro padrão de previsão.

A Tabela 16 apresenta os resultados para o erro padrão da previsão (SEP), para os dados sem pré-tratamento de dados.

Tabela 16. Valores de erro padrão de previsão (gL^{-1}) das amostras de açúcares.

	Maltose		Glicose		Frutose	
	Validação	Teste	Validação	Teste	Validação	Teste
PLS	2,10	2,24	2,67	1,67	1,11	1,38
PLS_AG	0,92	1,45	1,06	1,26	0,66	0,55
RLM_AG	0,67	1,08	1,06	1,18	0,56	0,70

Onde:

PLS: resultados utilizando o PLS sem a seleção de variáveis.

PLS_AG: resultados obtidos usando o PLS com as variáveis selecionadas pelo algoritmo genético.

RLM_AG: resultados obtidos usando a regressão linear múltipla (RLM) com as variáveis selecionadas pelo algoritmo genético.

Como podemos observar pela Tabela 16, que a seleção de comprimentos de onda proporciona uma significativa melhora nos resultados para o modelo dos mínimos quadrados parciais. Os resultados da regressão linear múltipla apresentaram-se ligeiramente melhores que os fornecidos pelos mínimos quadrados parciais, evidenciando a viabilidade do uso de um modelo matemático mais simples como a regressão linear múltipla, na quantificação dos açúcares.

Posteriormente realizou-se um pré-tratamento dos dados, envolvendo o filtro de Savisky-Golay e da primeira derivada, obtendo os seguintes resultados apresentados na Tabela 17.

Tabela 17. Valores do erro padrão de previsão (gL^{-1}) das amostras de açúcares, após o pré-tratamento dos dados.

	Maltose		Glicose		Frutose	
	Validação	Teste	Validação	Teste	Validação	Teste
PLS	1,03	1,16	1,05	0,86	0,52	0,55
PLS_AG	0,77	0,77	1,03	1,03	0,44	0,56
RLM_AG	0,69	0,90	0,99	1,08	0,48	0,58

Através das Tabelas 16 e 17, verificamos uma considerável melhora nos resultados dos mínimos quadrados parciais (sem pré-tratamento e com pré-tratamento dos dados), de aproximadamente 90%. Isso provavelmente deve-se a uma melhora da relação sinal/ruído, bem como problemas de linha base e sobreposição de picos dos espectros, que após o pré-tratamento foram minimizados.

Estudos realizados mostram que a utilização de derivadas na espectroscopia proporciona melhores resultados, principalmente quando se pretende usar calibração multivariada[54]. Contudo, deve-se tomar cuidado com a ordem da derivada usada, pois quanto maior a ordem da derivada, maior será a sensibilidade do modelo a ruídos.

Os resultados obtidos pela regressão linear múltipla com a seleção de variáveis são melhores ou iguais aos com seleção de variáveis envolvendo o método dos mínimos quadrados parciais.

Na Figura 35 estão indicados os comprimentos de onda selecionados pelo algoritmo genético para a regressão linear múltipla, quando se utiliza a primeira derivada dos espectros nos cálculos.

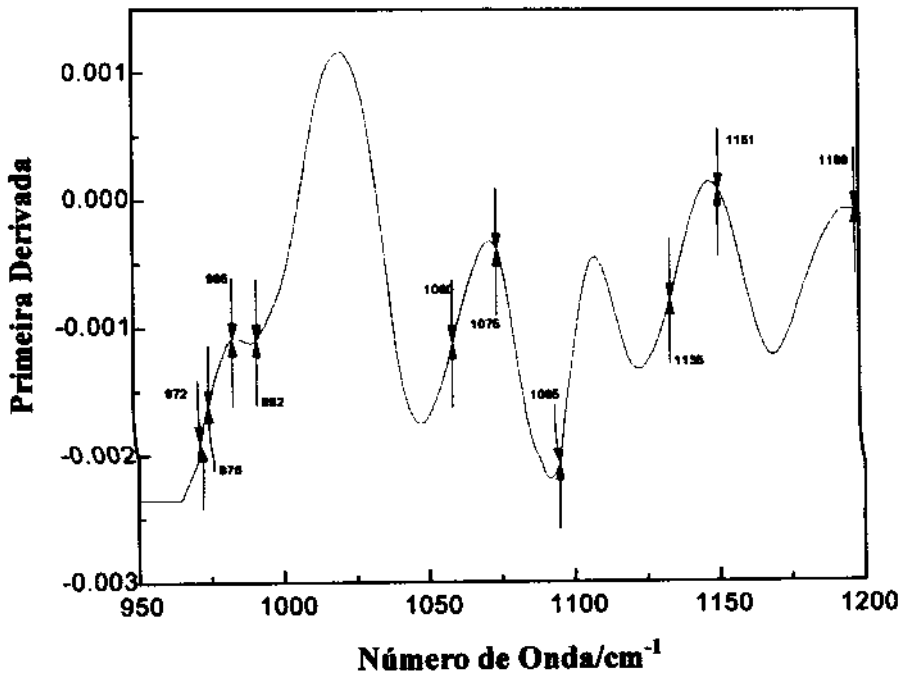


Figura 35. Primeira derivada do espectro da mistura de açúcares.

Após o pré-tratamento encontrou-se um erro médio nas concentrações dos açúcares na faixa de 3,2 a 3,5 %. De acordo com a literatura, o erro médio aceitável industrialmente na determinação destes açúcares é por volta de 3%, faixa esta obtida pelo método padrão que utiliza cromatografia líquida de alta eficiência[51,53]. Assim, a utilização de espectroscopia no infravermelho pode torna-se interessante para determinações "on-line" em processo industrial.

A Tabela 18 apresenta uma tentativa para atribuições dos números de onda selecionados para a regressão linear múltipla, relacionando-os com os respectivos açúcares.

Tabela 18. Atribuição dos números de onda selecionados[56].

Números de onda Selecionado (cm^{-1})	Estiramento das Espécies puras (cm^{-1})	Atribuição	
972	972	C-C	(Frutose)
976	972	C-C	(Frutose)
985	984	C-C	(Glicose)
992	984	C-C	(Glicose)
1060	1056,1065	C-O-C	(Frutose, Maltose)
1075	1080	C-O-C, C-O	(Maltose)
1095	1092	C-O	(Maltose)
1135	1136	C-O	(Glicose)
1151	1148	C-O	(Glicose)
1199	1192	O-C-H, C-C-H, C-O-H	(Frutose)

Segundo a literatura, os estiramentos encontrados na região entre 1153 a 904 cm^{-1} são atribuídos aos módulos vibracionais C-O e C-C. Na região entre 1199 a 1474 cm^{-1} ocorrem as deformações angulares das ligações O-C-H, C-C-H e C-O-H [51,53].

Através da Figura 36, é possível observar que a maior freqüência dos erros relativos na previsão do conjunto teste da mistura dos açúcares ocorre entre 0 a 5%, indicando portanto que a metodologia empregada apresenta-se satisfatória para determinações analíticas.

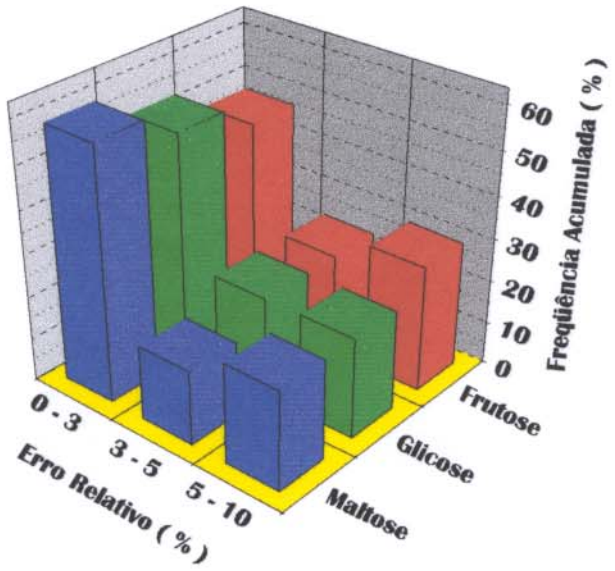


Figura 36. Freqüência acumulada do erro relativo absoluto nas determinações das misturas dos açúcares.

3.3.4.9 Estudo da Permuta de Variáveis Selecionadas entre os Modelos do PLS e RLM.

Com o intuito de verificar a possibilidade de intercâmbio entre as variáveis selecionadas para diferentes métodos de calibração multivariada, construiu-se novos modelos para os mínimos quadrados parciais a partir das variáveis selecionadas para a regressão linear múltipla e vice versa, tanto para o caso sem e com pré-tratamento dos dados. Os resultados obtidos podem ser observados nas Tabelas 19 e 20.

Tabela 19. Permuta de variáveis entre o modelo do PLS e RLM, sem pré-tratamento dos dados.

	SEP/ gL^{-1}		
	Teste		
Método	Maltose	Glicose	Frutose
PLS → RLM	1,41	1,22	0,82
RLM → PLS 6 componentes	1,38	1,12	0,79

Tabela 20. Permuta de variáveis entre o modelo do PLS e RLM, com pré-tratamento dos dados.

	SEP/ gL^{-1}		
	Teste		
Método	Maltose	Glicose	Frutose
PLS → RLM	1,04	1,28	0,76
RLM → PLS 7 componentes	0,87	1,12	0,57

Analisando os resultados da Tabela 19, verificamos um acréscimo significativo no SEP. Desta forma a permuta entre variáveis de diferentes métodos de calibração sem pré-tratamento de dados deve ser realizada de forma controlada. Outro ponto a ser salientado é o número das componentes principais utilizada, já que a troca das variáveis acarretou um aumento no número das componentes principais para a construção do modelo do PLS. Inicialmente todos os modelos envolvendo o PLS foram construídos com 4 componentes principais, com a permuta foram necessárias 6 componentes principais, ou seja, foi necessário o uso de mais duas componentes principais para explicar algumas variações introduzidas pelo novo conjunto de variáveis selecionados.

Pela Tabela 20, observa-se que com o pré-tratamento dos dados a permuta entre variáveis torna-se mais viável, pois os erros apresentaram uma variação menor. Nota-se também que a troca entre as variáveis do modelo da RLM para o PLS é menos significativa, isso provavelmente é devido ao aumento no número das componentes principais.

3.4 Determinação de Aditivos em Gasolina.

Esta aplicação tem como objetivo utilizar o infravermelho próximo para determinação simultânea de etanol e MTBE em gasolina. A calibração multivariada (com o PLS e RLM) foi empregada nos cálculos com a seleção de variáveis pelo algoritmo genético.

Este tópico foi dividido basicamente em três itens: Introdução, procedimento e resultados e discussão.

3.4.1 Introdução.

A gasolina é um combustível empregado comumente em veículos automotores, sendo constituída basicamente de uma mistura complexa de hidrocarbonetos alifáticos de cadeia normal ou ramificada de 5 a 12 carbonos (em alcanos) e alguns hidrocarbonetos aromáticos como benzeno, tolueno e xileno [57].

O uso da gasolina pura como combustível de automotores apresenta problemas relacionados com a octagem (medida relacionada com a compressão máxima da mistura combustível-ar, possível de ser obtida sem que ocorra a explosão espontânea). A gasolina isenta de aditivos geralmente apresenta uma baixa octagem, proporcionando a explosão da mistura antes que ocorra a ignição pela vela, desta maneira diminuíndo a eficiência do processo[58].

Para de minimizar ou eliminar problemas de octagem, utiliza-se aditivos químicos para evitar a explosão espontânea da mistura combustível/ar. Entre os aditivos normalmente empregados estão o metanol, etanol e metil-terc-butil-éter (MTBE)[59]. Outro ponto importante no controle na percentagem máxima de aditivos incorporados no combustível é a formação da separação fases, ocasionando problemas de funcionamento aos motores de veículos aéreos, terrestres ou marítimos.

A adição de éteres como aditivos, além de contribuírem para o aumento da octagem, também atuam como fornecedores naturais de comburente (oxigênio) para a combustão completa da gasolina, melhorando a eficiência do processo, além de minimizar a liberação de CO proveniente de uma combustão incompleta.

Na literatura existem alguns métodos propostos para a determinação de etanol em gasolina, como por exemplo: fluorescência no ultravioleta[60], infravermelho[61,62], ressonância magnética nuclear[57], infravermelho próximo[63]. Para a determinação de MTBE existem métodos baseados em: CG/MS[64], ressonância magnética nuclear[65], infravermelho, cromatografia gasosa. O método padrão internacional para a determinação de etanol ou MTBE em gasolina é o descrito pela ASTM (American Society for Testing and Materials) D 4815, o qual pode ser utilizado com satisfatório resultado para amostras com até 10 % (m/m) de etanol. Como no Brasil é permitido a adição de até 22% (m/m) de etanol, o método da ASTM não é empregado, mas sim o método da proveta, que consiste na remoção do aditivos através de uma extração via solução aquosa. O método da proveta não pode ser empregado na determinação do MTBE, já que éteres não são solúveis em meio aquoso.

Um possível método alternativo para a determinação de aditivos em gasolina baseia-se no uso da técnica no infravermelho próxima aliada a quimiometria. Até pouco tempo atrás, o infravermelho próximo era pouco usado em análises químicas, visto que os espectros obtidos nesta região são oriundos de bandas de combinação e apresentam elevada sobreposição de bandas, tornando a interpretação dos resultados bastante complexa[66]. A introdução da calibração multivariada abriu novos horizontes, e entre estes viabilizou o uso da região no infravermelho próximo, tornando possível análises químicas não destrutivas de plantas, animais e minerais.

Caminhando paralelo ao avanço das novas técnicas de análises de dados, encontra-se a instrumentação. A instrumentação espectroscópica apresentou progressos gradativos com o tempo, inicialmente com o

desenvolvimento da informática que viabilizou o uso da transformada de Fourier na aquisição dos espectros. Recentemente tem ocorrido a substituição de componentes ópticos inteiros como os monocromadores por um cristal. Esta nova tecnologia óptica é denominada de filtro óptico-acústico sintonizável (AOTF), pode reduzir o preço dos equipamentos, viabiliza a construção de equipamentos de campo, e permite a seleção aleatória de determinados comprimentos de onda sem a necessidade de fazer a aquisição completa do espectro [67,68].

3.4.2 Procedimento Experimental.

Foram preparadas 41 amostras de gasolina com etanol e MTBE na faixa 5 a 15 % (m/m). Utilizou-se gasolina isenta de etanol e MTBE, proveniente de diferentes fontes (golfo pérsico e Brasil, fornecida pela PETROBRÁS / REPLAN).

O planejamento experimental utilizado está ilustrado na Figura 37, onde as amostras utilizadas para construir o modelo de calibração encontram-se representadas pelo símbolo de um círculo em vermelho, o conjunto de validação por um asterisco em verde e o conjunto teste por triângulo azul.

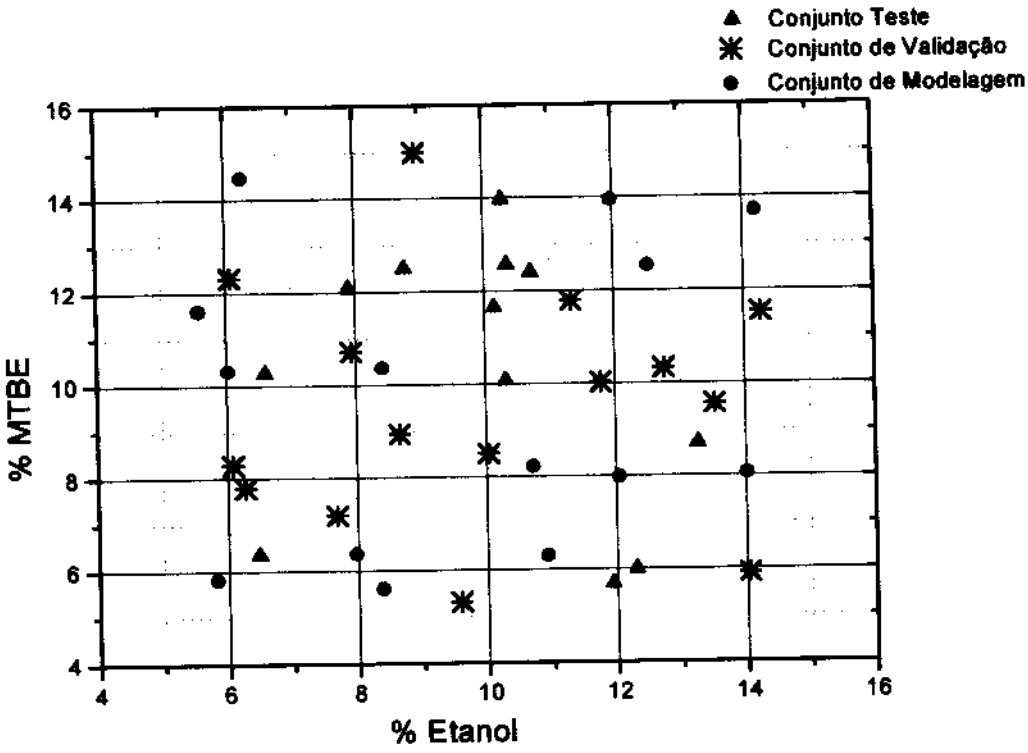


Figura 37. Planejamento experimental da mistura de combustíveis.

Os 41 espectros obtidos das misturas de combustível foram divididos em três subconjuntos: conjunto de modelagem com 15 amostras (apresentados na Tabela 21), conjunto de validação com 14 amostras (conforme Tabela 22) e conjunto teste com 12 amostras (segundo Tabela 23).

Tabela 21. Amostras utilizadas para a construção do modelo de calibração.

Amostra	Etanol (%)	MTBE (%)
1	5,57	11,61
2	6,26	14,48
3	8,37	5,61
4	10,71	8,23
5	11,97	13,96
6	14,01	8,08
7	14,19	13,71
8	5,82	5,82
9	6,02	10,33
10	7,97	6,38
11	8,41	10,37
12	10,92	6,30
13	12,04	8,00
14	12,52	12,54

Tabela 22. Amostras utilizadas para a construção do modelo de previsão da calibração.

Amostra	Etanol (%)	MTBE (%)
1	6,27	7,80
2	7,68	7,20
3	7,93	10,71
4	8,95	14,99
5	9,58	5,31
6	11,78	10,02
7	11,34	11,78
8	14,04	5,90
9	13,53	9,56
10	14,26	11,52
11	6,09	8,30
12	6,07	12,32
13	8,66	8,94
14	10,03	8,52
15	12,75	10,33

Tabela 23. Amostras utilizadas para a construção do modelo de validação da calibração.

Amostra	Etanol (%)	MTBE (%)
1	6,47	6,36
2	6,59	10,27
3	7,89	12,08
4	10,14	11,65
5	10,71	12,40
6	10,27	13,97
7	11,92	5,68
8	13,26	8,71
9	8,76	12,51
10	10,31	10,08
11	10,35	12,57
12	12,29	6,00

O espectrofotômetro no infravermelho próximo utilizado, foi construído no laboratório [69], sendo baseado em um filtro óptico acústico sintonizável (AOTF) [70] que opera na região de 1500-2400 nm, com resolução de 5 a 14 nm, e abertura óptica de 5x5 mm. Para a aquisição dos espectros, a mistura de combustível foi acondicionada a uma cela de quartzo com caminho óptico de 1 cm. Este equipamento é controlado por um gerador rf (rádio freqüência) e amplificador (Brimrose, 56-92 MHz, 4W). O equipamento dispõe de uma fonte de luz de tungstênio de 100W e detetor de PbS (Ealing Electro-Optics). O sinal da rf foi modulado em 190 Hz e o sinal do detetor foi enviado a um amplificador lock-in (Stanford, model 845). Um programa escrito em Visual Basic 3.0 foi empregado na aquisição dos dados e o cálculo dos espectros de absorbância.

3.4.2.1 Parâmetros de Configuração do Algoritmo Genético.

- Tamanho da população Inicial: 100 cromossomos
- Número de Gerações: 100
- Probabilidade de Cruzamento: 90 %
- Probabilidade de Mutação: 1 %
- Erro máximo para a finalização do processo: 1%
- Número máximo de variáveis selecionadas: 10
- Número de componentes principais: 3
- Método selecionado: PLS e RLM

3.4.3 Resultados e Discussão.

A seleção das variáveis torna-se um procedimento interessante para análises "on-line" onde se utiliza espectrofotômetros com filtro óptico acústico sintonizável (AOTF), já que este dispositivo eletrônico permite a seleção de comprimentos de onda randomicamente, tornando as análises mais rápidas [71].

Como desvantagem, o uso da AOTF com seleção de variáveis impedem que se faça o uso de filtros digitais e de derivadas, já que não se tem o espectro completo, e sim pontos discretos.

Para o sistema estudado foram selecionados 10 comprimentos de onda (1582, 1671, 1680, 1773, 1799, 1871, 1903, 2178, 2204 e 2395 nm) para o método da regressão linear múltipla, e 9 comprimentos de onda (1547, 1607, 1619, 1686, 1855, 2035, 2065, 2182 e 2324 nm) para os mínimos quadrados parciais.

Na Figura 38, é apresentada a coleção de espectros da mistura de combustível com os comprimentos de onda selecionados pelo algoritmo genético, utilizando a regressão linear múltipla na modelagem de calibração multivariada.

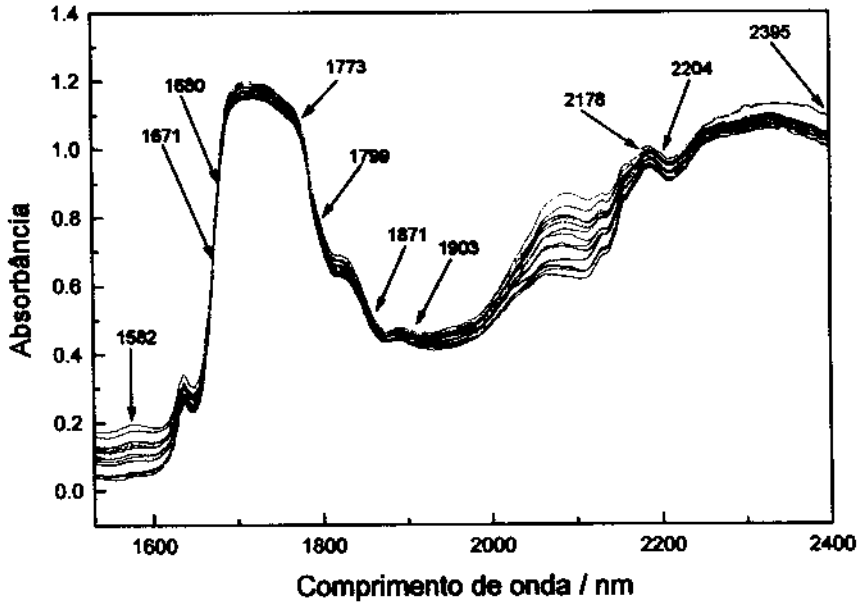


Figura 38. Coleção de Espectros no infravermelho próximo (NIR) da mistura de combustível e as variáveis selecionadas.

O grande problema no estudo quantitativo da mistura é a determinação de MTBE, pois esta espécie contribui com poucas informações para o espectro. Isso porque o grupo -C-O-C (grupo característico do MTBE) é inativo na região de estudo, e as informações sobre esta espécie são oriundas apenas dos grupos - CH_1 -CH_2 e -CH_3, que também se encontram presentes na matriz e no etanol [55]. Em contrapartida a determinação do etanol é relativamente simples de ser realizada, pois este possui o grupo -OH (ativo na região entre 1400 a 1700 nm, devido ao primeiro sobretom do grupo –OH, e a outra entre 2000 e 2150 nm devido à banda de combinação do mesmo grupo) que o diferencia da matriz e do MTBE, contribuindo de forma significativa na sua determinação[60]. Portanto somente a análise do etanol seria trivial, podendo ser realizada pelos métodos tradicionais de calibração.

Observando a Tabela 24, constata-se que a determinação de MTBE pelo método dos mínimos quadrados parciais com todas as variáveis, apresenta um erro padrão de previsão mais alto do que com a seleção de variáveis. Utilizando-se a regressão linear múltipla com seleção de variáveis, os erros do conjunto teste encontram-se menores que os demais métodos de calibração.

Um estudo prévio utilizando-se validação cruzada indicou que o número ideal das componentes principais, para serem empregadas no PLS seria três.

Tabela 24. Valores de erro padrão de previsão (%) das amostras das misturas de combustíveis.

	Etanol		MTBE	
	Validação	Teste	Validação	Teste
PLS	0,37	0,20	0,75	0,99
PLS_AG	0,38	0,32	0,31	0,63
RLM_AG	0,45	0,39	0,53	0,21

Onde:

PLS: resultados utilizando o PLS sem a seleção de variáveis.

PLS_AG: resultados obtidos usando o PLS com as variáveis selecionadas pelo algoritmo genético.

RLM_AG: resultados obtidos usando a regressão linear múltipla (RLM) com as variáveis selecionadas pelo algoritmo genético.

Se a escolha das variáveis fosse realizada utilizando-se apenas os comprimentos de onda onde há maiores variações entre os espectros (Figura 39), observam-se erros bem superiores para o MTBE, conforme mostrado na Tabela 25. Contudo para determinação do etanol os erros não são muito diferentes daqueles obtidos com a seleção de variáveis. Isso se deve ao fato de que o etanol possui uma absorbância na região entre 1400 a 1700 que é característico da espécie. A absorbância na região de máximo da banda responde linearmente à concentração de etanol. Desta forma fica evidente que para a modelagem tornar-se mais satisfatória para a determinação de MTBE, é necessário a inclusão de outras variáveis, que contribuem de forma significativa para o desenvolvimento modelo de calibração.

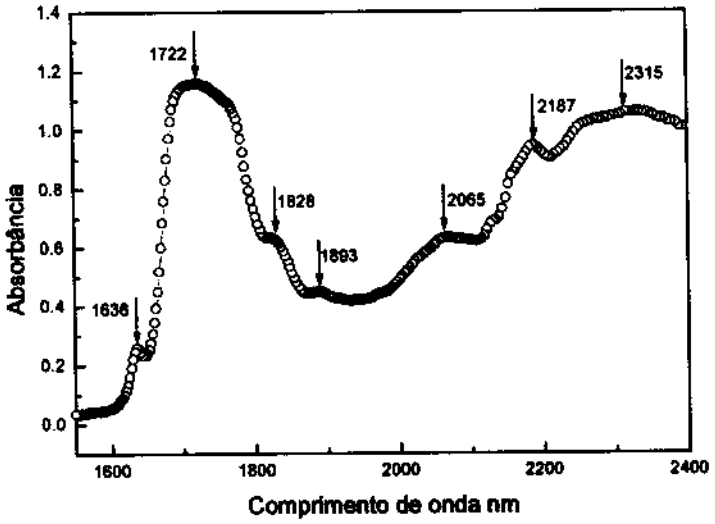


Figura 39. Espectro no infravermelho próximo (NIR) da mistura de combustível, com a seleção das variáveis onde há maiores variações entre os espectros.

Tabela 25. Valores de erro padrão de previsão (%) das amostras das misturas de combustíveis.

	Etanol		MTBE	
	validação	Teste	Validação	Teste
PLS	0,40	0,28	1,28	1,54
RLM	0,51	0,50	1,73	1,69

Com a utilização da seleção de variáveis em regressão linear múltipla, pode obter-se resultados semelhantes (como na determinação de etanol) e até melhores (determinação de MTBE), utilizando-se um modelo muito mais simples de ser explicado e entendido do que os mínimos quadrados parciais, atualmente considerado método padrão de calibração multivariada.

Aparentemente o uso de gasolinas de procedências diferentes não influenciou nos resultados, já que diferenças na constituição química da matriz poderiam acarretar diferenças de resultados na análise.

Com a finalidade de avaliar quais eram as variáveis mais importantes na determinação quantitativa dos aditivos, construiram-se curvas de calibração univariadas para etanol e MTBE, segundo as variáveis selecionadas pelo algoritmo genético, conforme Figuras 40.a e 40.b para o etanol e Figuras 41.a e 41.b para o MTBE. Observa-se claramente que na determinação do etanol o comprimento de onda a 1572 nm apresenta uma alta correlação entre a absorbância e a concentração de etanol. O comprimento de onda a 1572 nm encontra-se na região do primeiro sobretom do grupo -OH, grupo este característico dos álcoois.

Os demais comprimentos de onda possuem uma performance inferior, de acordo com a Tabela 26. Portanto os erros obtidos utilizando apenas uma variável é maior que do modelo multivariado (Tabela 24).

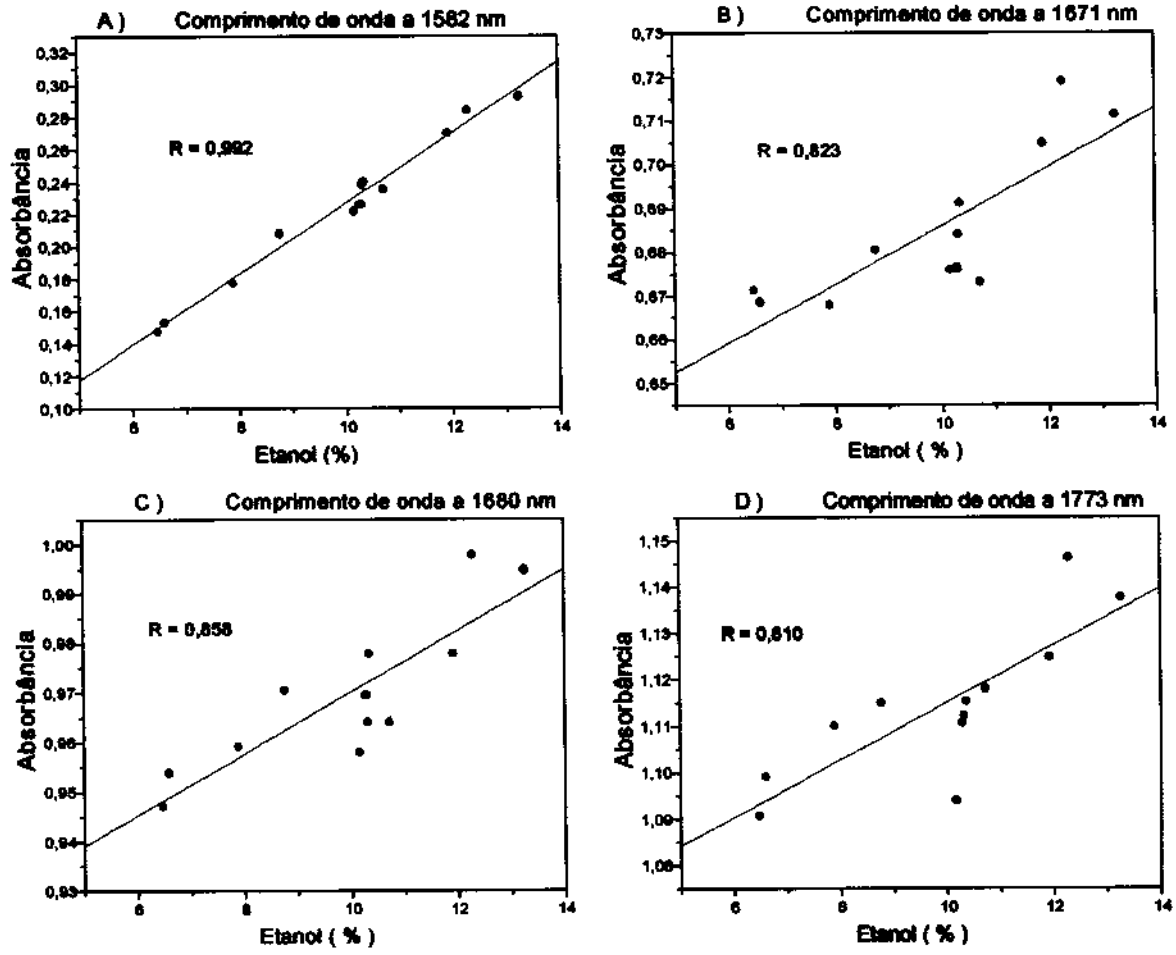


Figura 40.a. Curva de Calibração univariada para o etanol, com os comprimentos de onda selecionados pelo Algoritmo Genético.

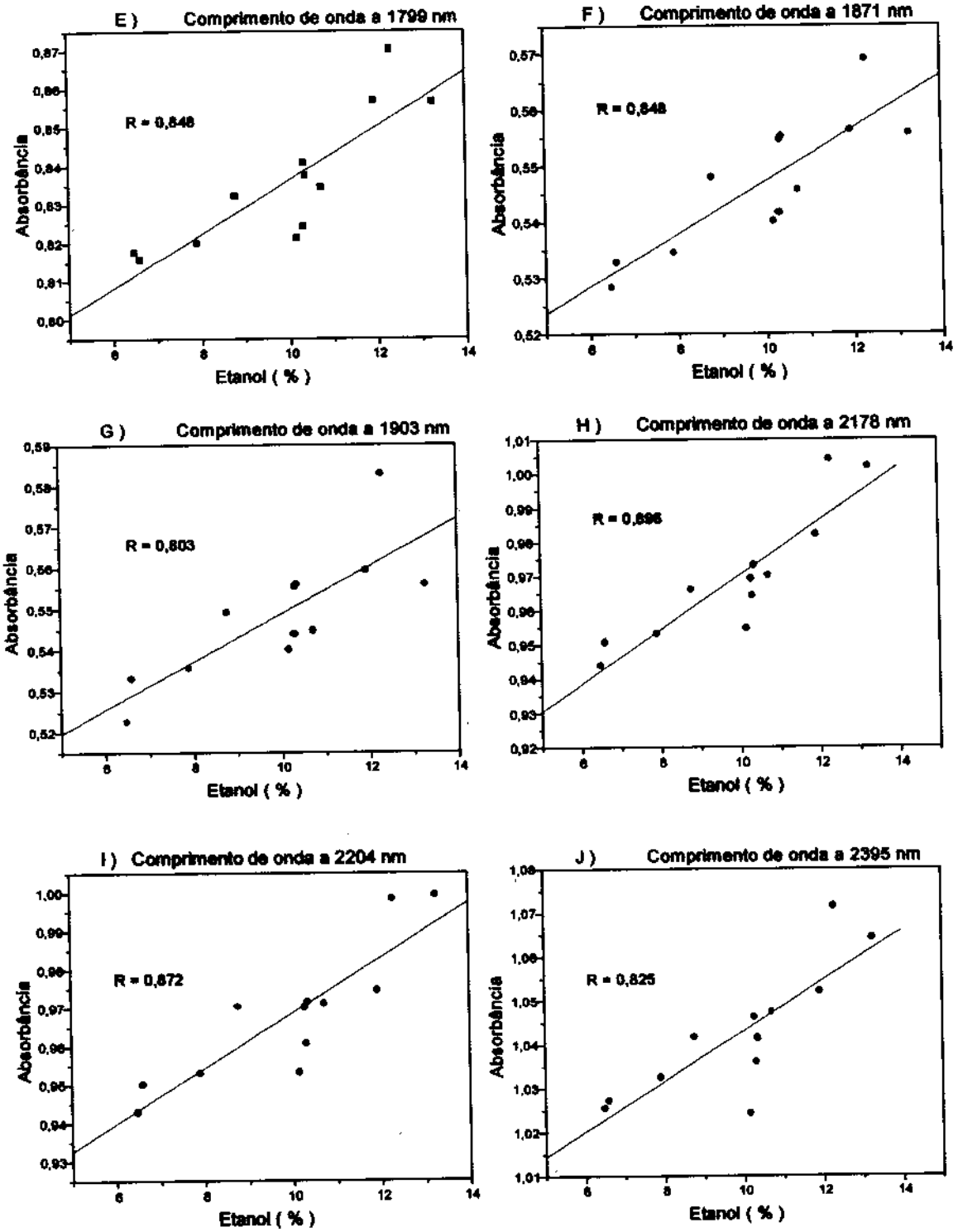


Figura 40.b. Curva de Calibração univariada para o etanol, com os comprimentos de onda selecionados pelo Algoritmo Genético.

Tabela 26. Erro relativo (%) para a previsão e validação do modelo de calibração univariado.

Amostras	Validação (%)	Teste (%)
1	11,67	12,37
2	6,07	9,76
3	7,99	9,42
4	10,52	7,53
5	3,60	5,89
6	2,09	6,50
7	3,64	1,09
8	1,54	2,50
9	-2,81	0,65
10	1,37	0,88
11	-3,02	0,41
12	1,25	-1,75
13	-0,52	
14	-6,75	
15	-8,24	

Na determinação do MTBE observa-se que a calibração univariada com os comprimentos selecionados pelo algoritmo genético possuem baixa correlação entre a absorbância e a concentração (Figura 41.a e 41.b), indicando portanto que as variáveis selecionadas possuem sinergia no modelo multivariado (regressão linear múltipla). Tal comportamento já era esperado, já que os éteres não possuem estiramento ativo nesta região no infravermelho próximo.

Devido a baixa correlação entre a absorbância e concentração para os comprimentos de onda selecionados para o MTBE, optou-se por não fazer a previsão das concentrações das amostras de previsão e validação.

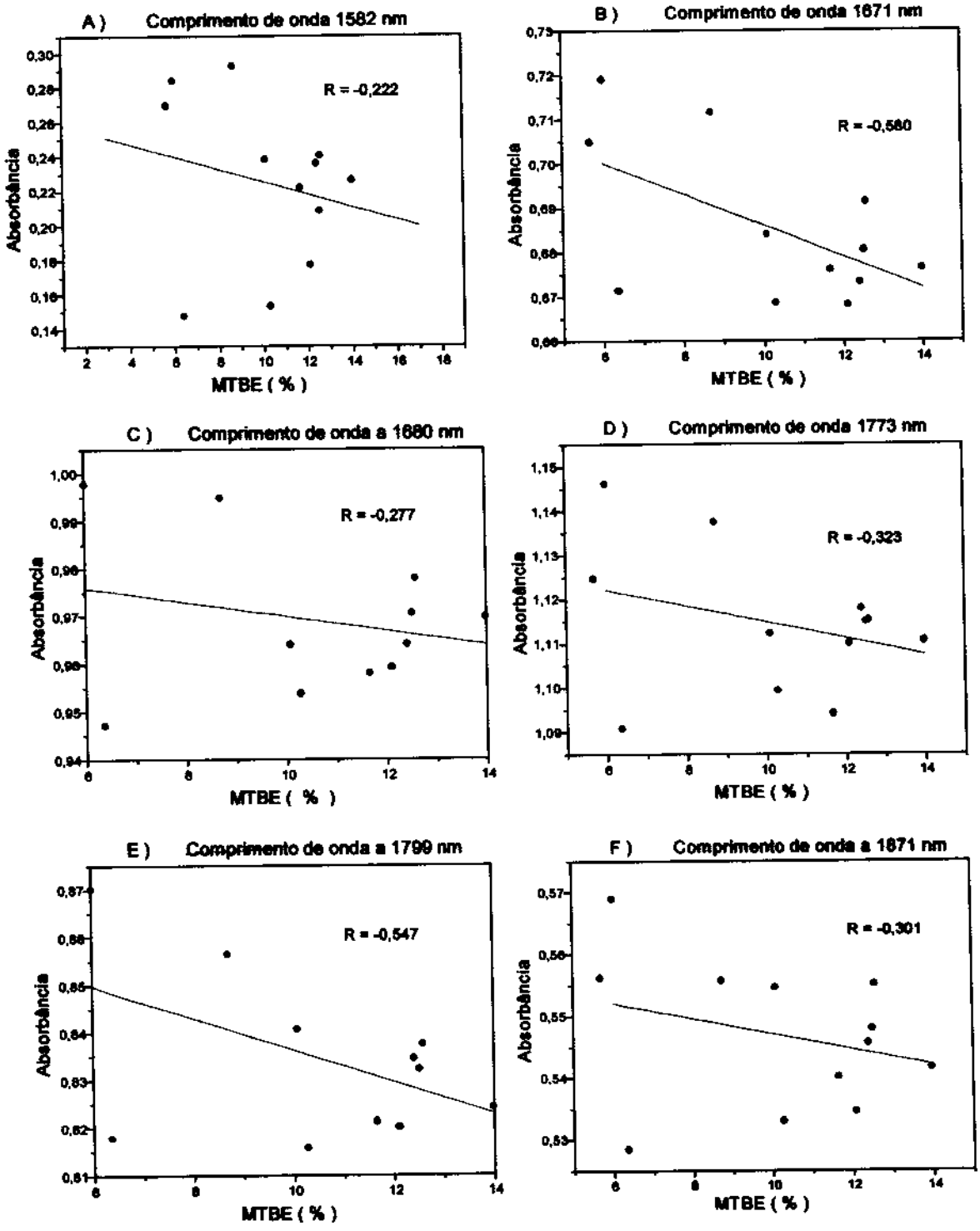


Figura 41.a. Curva de Calibração univariada para o MTBE, com os comprimentos de onda selecionados pelo Algoritmo Genético.

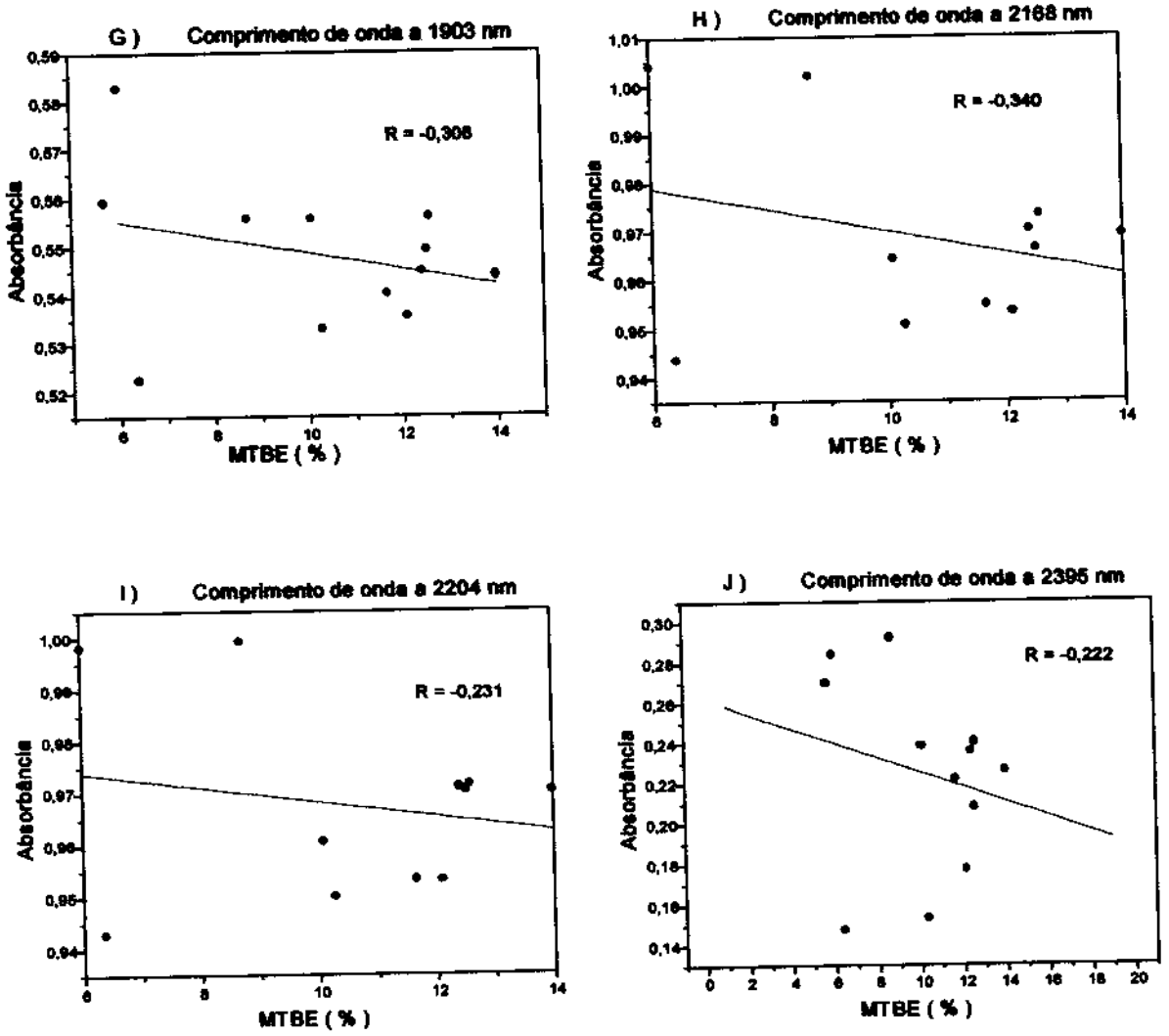


Figura 41.b. Curva de Calibração univariada para o MTBE, com os comprimentos de onda selecionados pelo Algoritmo Genético.

3.5 Determinação de Zinco e Cobre em Sistema de Fluxo.

O objetivo desta aplicação foi estudar a possibilidade de utilizar o algoritmo genético na otimização de sistemas em mais de uma dimensão. Para isso foi realizada a determinação de cobre e zinco por complexação com zincon em um sistema de fluxo com gradiente de pH, utilizando algoritmo genético e o método generalizado de aniquilação de posto (GRAM) na quantificação de cobre e zinco. Neste exemplo busca-se encontrar os melhores tempos e comprimentos de onda na determinação dos metais.

Será realizada uma breve introdução sobre a importância industrial e biológica dos metais estudados, em seguida será mostrado o procedimento experimental, finalizando com a interpretação dos resultados.

3.5.1 Introdução.

Nos dias atuais o cobre é muito conhecido pelo seu emprego em ligas, canos, fios elétricos e radiadores de automóveis. Entretanto, o seu comportamento bioquímico vem despertando o interesse nas últimas décadas quando foram estabelecidas e esclarecidas funções de alguns compostos de coordenação contendo cobre.

O cobre é um dos principais metais de transição presente no corpo humano. Os teores mínimos em vários fluidos são, em μg/100 mL: soro (72), plasma (75), eritrócitos (71), fluido espinhal (6), saliva (10) e fluidos digestivos (63). O metal também é encontrado em muitos tecidos e em teores mínimos diferentes. Alguns exemplos, dados em μg/100 mL, em tecidos são: cérebro (40), pulmão (500), coração (340), músculo (640), dentes (520), ossos (370), baço (120), rim (200) e fígado (1500). Para um metabolismo balanceado é necessária a presença de cobre na dieta alimentar.

O cobre é encontrado em carnes, frutos do mar, em muitos vegetais, cereais e nozes. Estima-se que na dieta são necessários de 2 a 5 mg de cobre

por dia. Isso implica que a mesma quantidade de cobre deve ser excretada para que haja um equilíbrio do seu teor no organismo.

O elemento químico zinco, é extraído da natureza principalmente a partir do mineral esfarelita ((ZnFe)S), sendo convertido a zinco metálico após uma série de etapas de beneficiamento industrial. Há um incontável número de aplicações industriais envolvendo o uso direto ou indireto do zinco, como por exemplo: galvanização de materiais ferrosos (aumento da resistência a corrosão), em baterias de Zn-Ni, como catalisador em processos de vulcanização (Óxido de zinco), em tintas (Sulfeto de zinco), em produtos farmacêuticos e cosméticos, agricultura (análise de solos).

O zinco possui elevada importância biológica, já que se encontra presente em diversas enzimas e cofatores essenciais para manutenção da vida. O zinco também se encontra envolvido no metabolismo de proteínas e ácidos nucléicos, sendo, portanto, fundamental no processo de multiplicação celular.

A preocupação do estudo e determinação de cobre e zinco, bem como de outros metais pesados, concentra-se no impacto ambiental que o acúmulo destas espécies podem ocasionar ao ecossistema. É sabido que hoje em dia há um grande volume de efluentes industriais sem nenhum tratamento são diariamente despejados no leito de diversos rios, levando a um desequilíbrio ecológico em diversas cadeias ambientais (marinhas e terrestres). Portanto para controlar as concentrações desses metais, o desenvolvimento de técnicas de análises mais rápidas e precisas são cada vez mais solicitadas e necessárias.

Há vários métodos para determinação de cobre zinco, em diferentes matrizes. Os mais comuns são os métodos espectrofotométricos [71-73], os polarográficos [74,75] e os métodos colorimétricos [76,77]. A determinação de cobre em plantas e solos pode ser feita com dietilditiocarbamato de sódio [77,78], com ditizona [79,80] ou com cuproína[81], porém envolve reação de extração e é susceptível à interferência de outros metais. A determinação de zinco em rochas

e material biológico é feita com ditizona [74,76,82] em pH de 6,0 a 9,5 e também com o reagente α-naftiltiocarbazona[83-84] em pH 8,0 a 10,0.

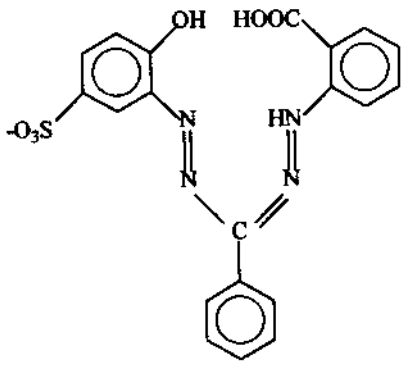
Além dessas metodologias, o zinco e o cobre podem ser determinados espectrometricamente sob a forma de complexos com zincon simultaneamente.

3.5.2 Complexante.

O ácido o-[1-(2-hidroxi-5-sulfofenil)–3-fenil–5-formazano] benzóico, mais conhecido como zincon é um complexante que pode ser usado na determinação de vários metais entre eles o cobre e zinco, entretanto o zincon não é especifico para estes metais, sendo suscetível a presença de interferentes. A Figura 42 apresenta a estrutura desse complexante[85-86].

Utilizando o complexante zincon é possível determinar cobre na presença de zinco, já que estas duas espécies complexam em diferentes pH's. O complexo cobre-zincon é estável no intervalo de pH 5,0 a 9,5, enquanto o complexo zinco-zincon é estável no intervalo de pH de 8,5 a 9,5. O zincon apresenta absobância máxima em 470 nm, quando complexado com zinco apresenta máxima absorbância em 620 nm , e com o cobre em 595 nm[87].

A)



B)

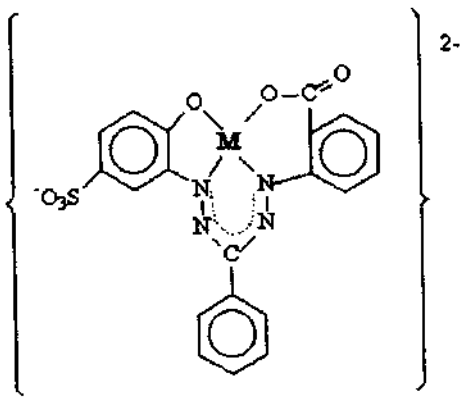


Figura 42. A) Estrutura da molécula de zincon., B) Estrutura da molécula de zincon Complexada com um metal (**M**).

3.5.3 Procedimento Experimental

Para a determinação de cobre e zinco em misturas, foi proposto um experimento com 14 amostras, conforme apresentado na Tabela 27.

Tabela 27. Soluções preparadas da mistura dos metais.

Solução	Zinco / $molL^{-1}$	Cobre / $molL^{-1}$
1	$2,84x10^{-5}$	$9,72x10^{-6}$
2	$2,84x10^{-5}$	$1,94x10^{-5}$
3	$2,84x10^{-5}$	$2,92x10^{-5}$
4	$2,84x10^{-5}$	$3,89x10^{-5}$
5	$5,68x10^{-5}$	$1,94x10^{-5}$
6	$5,68x10^{-5}$	$2,92x10^{-5}$
7	$5,68x10^{-5}$	$3,89x10^{-5}$
8	$5,68x10^{-5}$	$4,86x10^{-5}$
9	$8,52x10^{-5}$	$9,72x10^{-6}$
10	$8,52x10^{-5}$	$1,94x10^{-5}$
11	$8,52x10^{-5}$	$2,92x10^{-5}$
12	$8,52x10^{-5}$	$3,89x10^{-5}$
13	$8,52x10^{-5}$	$4,86x10^{-5}$
14	$1,14x10^{-4}$	$9,72x10^{-6}$

3.5.4 Preparação das Soluções para Padronização do Cu^{2+} e Zn^{2+}.

Solução Estoque de Cu^{2+} e Zn^{2+}.

A partir dos sais de $(CuSO_4)5H_2O$ e $(ZnSO_4)7H_2O$ preparam-se 500,00 mL de soluções estoque de Cu^{2+} e Zn^{2+}, com $8,10x10^{-4}$ $molL^{-1}$ e $7,10x10^{-4}$ $molL^{-1}$ respectivamente.

Solução do Complexante.

Uma solução de EDTA 1,53 x 10^{-3} $molL^{-1}$ foi preparada, dissolvendo-se 0,5692 g de EDTA em água destilada deionizada e completando-se o volume a 1000,00 mL.

Indicadores.

Preparou-se uma mistura do indicador murexida em cloreto de potássio, triturando e misturando 0,10 g do indicador em 10,00 g de KCl em um Almofariz, tomando-se para cada titulação uma massa de aproximadamente 0,20 g da mistura. Utilizou-se solução de indicador Negro de Eriocromo T

Soluções Tampões.

A partir de um balão volumétrico de 100,00 mL foi preparada uma solução tampão contendo cloreto de amônio (NH_4Cl) 1 $molL^{-1}$ e hidróxido de amônio (NH_4OH) 1 $molL^{-1}$.

Preparou-se 100,00 mL de uma solução 1 $molL^{-1}$ de bicarbonato de potássio ($KHCO_3$) e 1 $molL^{-1}$ carbonato de potássio (K_2CO_3).

3.5.4.1 Padronização das soluções estoque de Cu^{2+} e Zn^{2+}.

Para a padronização de solução de Cu^{2+} tomou-se uma alíquota de 20,00 mL da solução de sulfato de cobre a ser padronizada em erlenmyer. À solução adicionou-se aproximadamente 0,20 g da mistura de murexida em KCl e 5,0 mL da solução 1$molL^{-1}$ de NH_4Cl. Posteriormente, adicionou-se gota a gota amônia concentrado até o aparecimento de uma coloração amarela clara (formação do complexo indicador/metal). Titulou-se com a solução padrão de EDTA 1,53 x 10^{-3} $molL^{-1}$ até as proximidades do ponto final (solução levemente azulada). A partir dai, foi adicionado cerca de 10 mL de amônia para alcalinizar o meio e em seguida

adicionou-se mais EDTA 1,53 x 10^{-3} molL^{-1} até a permanência da coloração azul-violeta.

Para a padronização da solução de Zn^{2+}, tomou-se uma alíquota de 20,00 mL da solução de sulfato de zinco a ser padronizada em erlenmyer. Acrescentaram-se duas gotas da solução do indicador Negro de Eriocromo T e em seguida mais 20,00 mL da solução tampão NH_4OH / NH_4Cl, observando-se o aparecimento de uma coloração roxa (formação do complexo indicador/metal). Titulou-se com a solução padrão de EDTA 1,53 x 10^{-3} molL^{-1} até o aparecimento de uma coloração azul cinzento.

Esses procedimentos foram realizados em triplicata (para cada metal), obtendo-se os resultados apresentados na Tabela 28:

Tabela 28. Resultados da Titulação de Padronização dos Metais.

	Volume de EDTA 1,53 x 10^{-3} molL^{-1} gasto (mL)	
Amostra	Padronização do Cu^{2+}	Padronização Zn^{2+}
1	10,60	18,60
2	10,60	18,70
3	10,60	18,60
média	10,60	18,64

A partir dos resultados apresentados na Tabela 28, a concentração de Cu^{2+} foi calculada como sendo 8,10 x 10^{-4} molL^{-1} e a de Zn^{2+} como de 1,42 x 10^{-3} molL^{-1}

3.5.4.2 Geração do Gradiente de pH em Fluxo.

Para a obtenção das superfícies de resposta relativas às amostras num sistema de fluxo com detecção multicanal, utilizou-se um gradiente de pH gerado pela dispersão de uma solução de Na_3PO_4 0,200 molL^{-1} em uma mistura de ácido bórico 0,200 molL^{-1} e ácido cítrico 0,050 molL^{-1}. Em confluência com o gradiente de pH, foi adicionado uma linha contendo o complexante zincon 2,00x10^{-4} molL^{-1} e

a mistura de metais (amostra). O transporte dos fluidos foi realizados por tubos de polietileno e propulsionados por uma bomba peristáltica (Ismatec) a vazão constante de 0,4 mL min^{-1}. A Figura 43 mostra o esquema do sistema de fluxo utilizado. Desse modo obteve-se superfícies de resposta com 35 espectros na faixa de (425,0 nm até 763,3 nm com 341 pontos) da mistura de complexos Cu-zincon, Zn-zincon e zincon em excesso, dependendo do pH do meio, relativos a 35 tempos distintos.

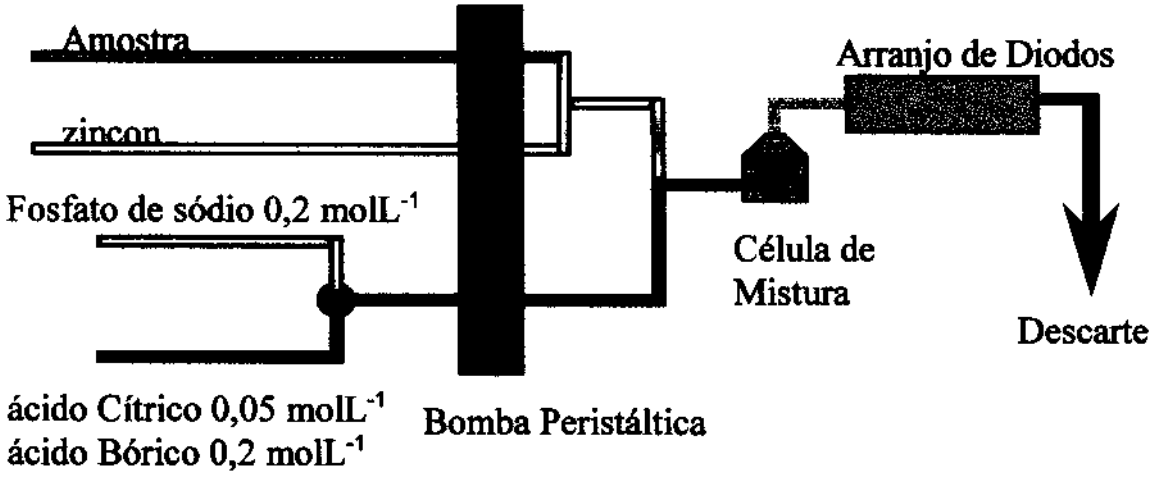


Figura 43. Esquema da análise em fluxo utilizada.

Obteve-se superfícies de resposta relativas as quatorze amostras distintas utilizando as mesmas condições de gradiente de pH, abrangendo misturas numa faixa de concentrações para o Zn^{2+} entre $2,84x10^{-5}$ e $1,14x10^{-4}$ $molL^{-1}$ e Cu^{2+} entre $9,72x10^{-6}$ e $4,86x10^{-5}$ $molL^{-1}$.

As superfícies de resposta possuem uma grande sobreposição entre os espectros dos complexos e do ligante. Desse modo a modelagem dos dados com o GRAM apresenta erros relativamente altos. Para contornar esse problema optou-se por otimizar a escolha dos comprimentos de onda e dos tempos (pH's) da superfície de resposta com o algoritmo genético. Assim pode-se encontrar

condições ótimas de separação entre as espécies presentes nas amostras, facilitando a quantificação e reduzindo o erro nas determinações.

3.5.4.3 Seleção e Validação das Variáveis do Cobre e Zinco.

Esta etapa envolveu a separação das amostras em dois grupos (seleção e validação das variáveis). A partir das amostras de cada grupo, fez-se combinações dois a dois (amostras padrão e desconhecida), e utilizou-se o erro na previsão obtida pelo GRAM como sendo a resposta a ser otimizada.

No primeiro grupo usado para a seleção de variáveis, havia sete amostras, obtendo-se um total de vinte e uma combinações para a seleção das variáveis para o Zn^{2+}. Para o cobre havia seis amostras, ou quinze combinações. No segundo grupo (validação do modelo), havia quatorze amostras para o zinco, gerando setenta novas combinações, e onze amostras para o cobre, ou quarenta combinações. O número total de amostras empregadas para a otimização do cobre é menor, pois foram suprimidas as amostras com Cu^{2+} igual a $9{,}72 \times 10^{-6}$ $molL^{-1}$, devido à sua baixa concentração.

Para ilustrar a combinação entre as amostras e indicar quais foram utilizadas para a seleção de validação do modelo, construiu-se as Tabelas 29 e 30.

A Tabela 29 apresenta as quatorze soluções preparadas com as misturas dos metais para a seleção e validação das melhores variáveis para a previsão da concentração do zinco em amostras desconhecidas, a partir de uma amostra padrão. O triângulo superior em destaque da Tabela 29, representa as noventa e uma combinações possíveis (vinte e uma combinações para seleção e setenta combinações para a validação). As amostras utilizadas para a seleção das variáveis são as posições que se encontram marcadas com a letra **M** (modelo). Por exemplo, a interseção entre as soluções 1 e 3 encontra-se marcada com a letra **M**, ou seja, tanto a solução 1 como a solução 3 foram utilizadas na etapa de seleção de variáveis. A utilização destas soluções na etapa de seleção de variáveis não impede que estas soluções sejam empregadas para a validação do

modelo, portanto a combinação entre as soluções 2 e 3 representa uma combinação possível para a validação do modelo. O mesmo raciocínio deve ser empregado para a seleção de variáveis para a determinação de cobre, na Tabela 30.

Tabela 29. Combinações utilizadas na seleção e validação do modelo de determinação de zinco.

	Sol_01	Sol_02	Sol_03	Sol_04	Sol_05	Sol_06	Sol_07	Sol_08	Sol_09	Sol_10	Sol_11	Sol_12	Sol_13	Sol_14
Sol_01			M		M		M				M		M	M
Sol_02														
Sol_03					M		M				M		M	M
Sol_04														
Sol_05							M				M		M	M
Sol_06														
Sol_07											M		M	M
Sol_08														
Sol_09														
Sol_10														
Sol_11													M	M
Sol_12														
Sol_13														M
Sol_14														

Tabela 30. Combinações utilizadas na seleção e validação do modelo de determinação de cobre.

	Sol_02	Sol_03	Sol_04	Sol_05	Sol_06	Sol_07	Sol_08	Sol_10	Sol_11	Sol_12	Sol_13
Sol_02			M		M		M	M		M	
Sol_03											
Sol_04					M		M	M		M	
Sol_05											
Sol_06							M	M		M	
Sol_07											
Sol_08								M		M	
Sol_10										M	
Sol_11											
Sol_12											
Sol_13											

3.5.4.4 Parâmetros de Configuração do Algoritmo Genético.

- Tamanho da população Inicial: 100 cromossomos
- Número de Gerações: 2000
- Probabilidade de Cruzamento: 90 %
- Probabilidade de Mutação: 1 %
- Erro máximo para a finalização do processo: 1%
- Número máximo de variáveis selecionadas: 30

3.5.5 Resultados e discussão.

Em seguida serão apresentados os resultados obtidos para a seleção de variáveis em dados bilineares. Como esclarecimento inicial deseja-se ressaltar que a aquisição dos fiagramas foi realizada em função do tempo e não em função do pH, embora o tempo e pH neste caso estejam relacionados. Entretanto apesar de não ter sido construída uma curva de calibração é possível estimar o pH próximo de determinadas faixas. Sabe-se que o pH variou entre 4,20 a 11,70 e que próximo de 5,0 ocorre a formação do complexo Cu-zincon, enquanto próximo a 8,0 existe a formação do complexo Zn-zincon.

O espectro do zincon apresenta uma dependência com o pH, de modo que para realizar-se qualquer inferência sobre o sistema, deve-se considerar separadamente o espectro do mesmo, e sua dependência com o tempo (e conseqüentemente o pH). A formação do complexo Cu-zincon ocorre entre os pH's 5,0 e 9,5, enquanto que a formação do complexo Zn-zincon ocorre em pH acima de 8,0. A Figura 44 mostra o espectro da solução de zincon puro utilizado sob ação do gradiente de pH [85,87].

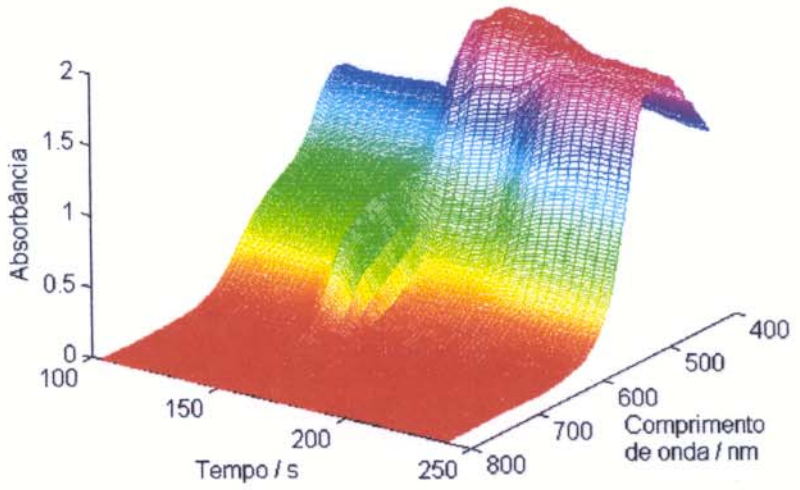


Figura 44. Superfície de resposta da solução de zincon puro sob ação do gradiente de pH gerado em fluxo.

A Figura 45, ilustra uma superfície de resposta obtida para a mistura de cobre e zinco com excesso do complexante zincon, sob ação do mesmo gradiente de pH.

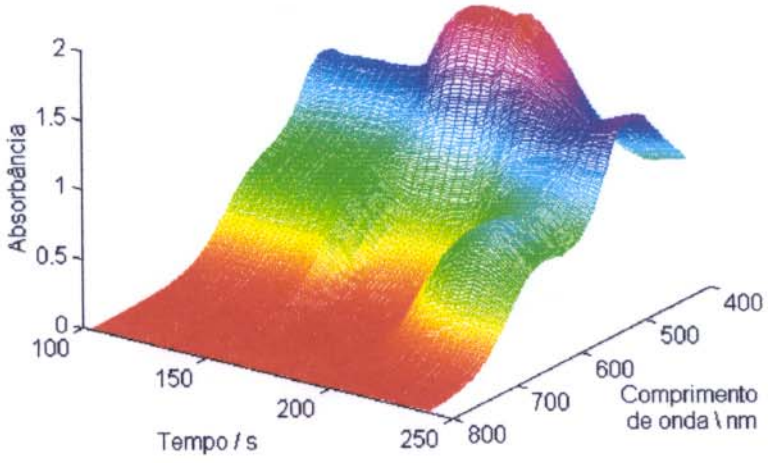


Figura 45. Superfície de resposta de amostra de Cu-zincon, Zn-zincon e zincon em excesso.

Através da análise da superfície de resposta da Figura 45 e a superfície de resposta para o zincon puro (Figura 44) pode-se observar, que nos primeiros espectros, o pH não é suficiente para haver complexação dos metais, de modo que basicamente tem-se o espectro do ligante zincon, isto pode ser verificado na literatura[18].

À medida que o pH aumenta, ou seja, à medida que se tomam os espectros, por volta do décimo espectro, observa-se uma absorbância relativa ao complexo Cu-zincon formado, em torno de 595 nm. Com o aumento do pH, há a formação do complexo Zn-zincon adicionalmente às espécies anteriores. Isso é evidenciado pelo aumento da absorbância na região em torno de 620 nm relativa à formação do complexo Zn-zincon.

Inicialmente, a modelagem dos dados com o GRAM foi realizada utilizando toda a superfície de resposta, e obteve-se os resultados apresentados na Tabela 31. Nesta Tabela é mostrada uma média dos valores de SEP e erro relativo das determinações de cobre e zinco.

Tabela 31. Resultados para zinco e cobre usando o GRAM, com todas as variáveis.

	Zinco	Cobre
SEPV/ mol L^{-1}	5,71 $x10^{-6}$	8.90 $x\ 10^{-6}$
SEPC/ mol L^{-1}	6,18 $x10^{-6}$	1,10 $x10^{-5}$
Erro relativo de calibração / %	11,92	28,00
Erro relativo de validação / %	11,33	26,13

Analisando os resultados da Tabela 31, constata-se que os erros encontrados são inadmissíveis para uma análise quantitativa de zinco e cobre. Desta forma fica claro que o GRAM tem extrema dificuldade na quantificação de espécies que possuem espectros altamente sobrepostos.

Utilizando-se o algoritmo genético, foram selecionados os espectros tomados após a permutação da válvula nos tempos: 112, 136, 148, 176, 180, 188, 192, 196 e 219 segundos e foram simultaneamente selecionados os comprimentos de onda: 755, 697, 645, 636, 597, 584 e 475 nm para a determinação do complexo de zinco. Na determinação do cobre foram selecionados os espectros tomados após a permutação da válvula nos tempos: 104, 116, 124, 128, 136, 156 e 160 segundos e foram simultaneamente selecionados os comprimentos de onda 751, 746, 730, 723, 688, 673, 658, 625, 624, 609 e 582 nm.

É importante observar que a seleção das variáveis pelo algoritmo genético é realizada de tal forma que as variáveis escolhidas minimizem a influência das outras espécies na análise. Dessa forma para a determinação do cobre, foram escolhidos os espectros registrados em tempos onde o pH não era suficiente para haver a complexação do zinco, ou seja, foram escolhidas variáveis que minimizassem esta interferência. Por outro lado, para a análise do zinco, foram

selecionados preferencialmente os espectros referentes aos valores de pH onde houvesse a formação do complexo Zn-zincon, e nos quais o complexo Cu-zincon não influenciasse no sinal observado. Os comprimentos de onda selecionados para o zinco referem-se basicamente à banda de absorção do complexo Zn-zincon, exceto por um selecionado em 475 nm que se refere à banda do ligante, cujo decaimento, proporcional à formação do complexo de zinco, também deve ser monitorado.

Os resultados obtidos utilizando-se os subconjuntos de variáveis selecionadas, proporcionou uma melhora significativa nos valores dos erros de calibração e previsão das concentrações de zinco e cobre, conforme podem ser observadas na Tabela 32.

Tabela 32. Resultados para zinco e cobre usando o GRAM, com as variáveis selecionadas.

	Zinco	Cobre
SEPV/ mol L^{-1}	$4{,}14 \times 10^{-6}$	1.14×10^{-6}
SEPC/ mol L^{-1}	1.14×10^{-6}	$8{,}71 \times 10^{-7}$
Erro relativo de calibração / %	1,76	2,24
Erro relativo de validação / %	4,40	3,04

Comparando as Tabelas 31 e 32, nota-se que realmente a seleção de variáveis melhora os resultados, otimizando desta forma a previsão da concentração das espécies de interesse.

Uma análise do conjunto de variáveis selecionadas para cada espécie, deixa evidente que para ambos os casos, a seleção das variáveis foi seletiva, minimizando a influência de outras espécies presentes na solução. Com o objetivo de comprovar estas hipóteses levantadas, realizou-se um estudo envolvendo a análise do posto (utilizando decomposição de valores singulares) das submatrizes construídas a partir dos variáveis selecionadas, chegando aos resultados apresentados na Figura 46.

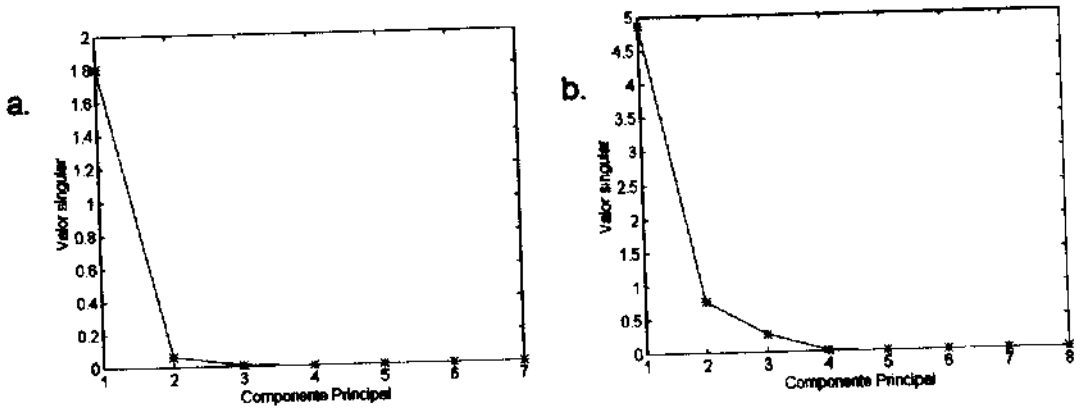


Figura 46. Análise do posto, obtida a partir da decomposição em valores singulares, da submatriz das variáveis selecionadas: **a:** para o cobre e **b:** para o zinco.

Analisando a Figura 46.a, observa-se que o posto da submatriz tomada para a quantificação do cobre é 1, portanto a variância explicada é referente a uma única espécie. Isso é explicado pelo fato de que a faixa de concentração de cobre é muito baixa quando comparada com a concentração de zincon. Desse modo, a variação de sinal relativo ao consumo de zincon é desprezível, e observa-se apenas uma espécie (Complexo Cu-zincon) responsável pelo posto unitário da submatriz selecionada. Além disso, nessa faixa de variáveis selecionadas o pH do sistema não é alto o suficiente para a formação do complexo Zn-zincon.

Pela Figura 46.b, a submatriz selecionada para a quantificação do zinco possui posto igual a 3. Isso é devido à submatriz apresentar elementos tomados em tempos onde só há o ligante, e onde há a formação de complexo Cu-zincon, e do complexo Zn-zincon.

CONCLUSÕES

O algoritmo genético mostrou-se uma ferramenta poderosa, no que tange a versatilidade e robustez do modelo. Por ser um algoritmo versátil, torna-se fácil introduzir adaptações ao programa de forma a atender a diferentes sistemas. A robustez é decorrente de um modelo consistente, o qual apresenta pequenas variações no desvio padrão dos erros, durante a previsão de novas amostras. Para modelos robustos, algumas vezes torna-se possível realizar extrapolações.

A robustez do modelo, utilizando o algoritmo genético, pode ser constatada ao compararmos os resultados apresentados para o conjunto de validação e teste, já que os erros não apresentam elevada discrepância em nenhuma das aplicações relatas neste trabalho. Outro ponto que reforça esta afirmação, é que com a seleção de variáveis em dados com ou sem o pré-tratamento, o modelo apresenta pouca variações na previsão das amostras de validação e teste.

Com relação à versatilidade, constatou-se que basicamente o mesmo algoritmo pode ser utilizado para três tipos distintos de otimizações: - Determinação do máximo de uma função (aplicação teórica); - Seleção de variáveis em espectros no infravermelho médio e próximo; - Seleção de variáveis em dados bilineares (duas dimensões) simultaneamente.

Dependendo do problema a ser otimizado, o algoritmo genético poderá consumir um elevado tempo de processamento. O tempo que o programa demorará para otimizar o sistema estará intimamente relacionado com as dimensões das matrizes de dados de entrada e do método matemático empregado para construção e avaliação do modelo. Tal característica não deve ser encarada como uma desvantagem, pois uma vez otimizado o sistema, não será mais necessário realizar novamente o processo de otimização (deste que não ocorra variações nos parâmetros otimizados inicialmente, ou de fatores que possam influenciar direta ou indiretamente em um ou mais variáveis importante do sistema). Para reduzir o tempo de processamento, tem-se proposto o

emprego da paralelização de parte do programa que requerem mais tempo para serem executadas.

A determinação de açúcares por espectroscopia no infravermelho médio / ATR mostrou que a seleção de variáveis por si só pode dispensar os pré-tratamentos de dados, já que as variáveis selecionadas procuram minimizar o efeito da relação sinal/ruído, de sobreposições de picos e de não linearidade dos dados.

A determinação de MTBE e etanol em gasolina mostrou-se viável através da técnica de espectroscopia no infravermelho próximo associada a quimiometria. O algoritmo genético mostrou-se uma ferramenta bastante interessante para ser empregada em conjunto com o filtro óptico-acústico sintonizável (AOTF), já que a AOTF possibilita a aquisição dos dados de específicos comprimentos de onda pré-selecionados. A utilização da regressão linear múltipla apresentou resultados na determinação de MTBE até melhores que o PLS, mostrando que é possível o uso de modelos matemáticos mais simples.

A determinação de Cu^{2+} e Zn^{2+} com zincon, foi o experimento que melhor ilustrou a potencialidade do algoritmo genético, já que com a seleção de variáveis houve uma melhora na previsão das concentrações, de aproximadamente 5 vezes para o zinco e 10 vezes para o cobre, quando comprado com os resultados sem a seleção de variáveis.

Em todos os experimentos realizados foi possível realizar uma atribuição química para as variáveis selecionadas, mostrando concordância entre a teoria espectroscópica e os resultados obtidos.

Entretanto, apesar do elevado potencial do algoritmo genético, bem como de outras ferramentas quimiométricas, há uma elevada resistência à introdução desses novos métodos no Brasil. Isso traz como conseqüência um atraso no desenvolvimento e aplicações de ferramentas quimiométricas que já se encontram consolidadas nos países desenvolvidos.

ESTUDOS FUTUROS

O algoritmo genético apresenta um imenso potencial em processos industriais, onde a cada dia a otimização é crucial, principalmente diante de um mercado aberto e competitivo.

Há uma clara tendência tanto de âmbito cientifico como tecnológico, da aplicação de procedimentos de otimização em sistemas complexos, onde métodos tradicionais não apresentam resultados satisfatórios.

Neste contexto, abrem-se inúmeras perspectivas para o algoritmo genético, devido à robustez e versatilidade da técnica, como os exemplos citados abaixo.

I- Técnicas Híbridas de Inteligência Artificial.

Há a possibilidade da implementação de metodologias híbridas de inteligência artificial como: Neurogenético e Neurofuzzygenético.

Uma das mais novas aplicações do algoritmo genético é a sua junção com a rede neural (Neurogenético). O algoritmo genético pode ser aplicado em conjunto com rede neural de diversas maneiras: Em otimização de topologia, o algoritmo genético é usado para otimizar a melhor topologia (número de camadas internas, número de neurônios, interconexões) para a rede neural. Para treinamento de rede neural (determinar os pesos). O algoritmo genético pode ser empregado para controle de parâmetros como, taxa de aprendizagem, função de transferências, nível de tolerância, etc.

Apesar das inúmeras possibilidades de implementação do algoritmo genético em rede neural, aparentemente não há nenhum trabalho utilizando esta metodologia em química analítica.

O sistema híbrido Neurofuzzygenético combina a vantagem do sistema Fuzzy, o qual trabalha com conhecimento explícito e que pode ser compreendido, e rede neural a qual trabalha com conhecimento implícito, o qual pode ser adquirido com o aprendizado. A lógica Fuzzy aumenta a capacidade de generalização de sistemas de rede neural permitindo realizar extrapolações se necessária além do

limite dos dados treinados. A união entre estas três técnicas aumenta a eficiência do modelo de calibração multivariada.

II – Dados de Ordem Superior.

Outra aplicação bastante promissora é a junção do algoritmo genético ao PLS-Multway, possibilitando otimizar sistemas hifenados, para a seleção de variáveis em mais de uma dimensão. Isso já foi realizado nesta dissertação para a seleção de variáveis utilizando o GRAM, mas pode ser aplicado em outras técnicas quimiométricas deste gênero.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Sharaf, M. A., Illman, D. L., Kowalski, B. R. , *Chemometrics* , John Wiley & Sons , New York, (1986).

[2] Massart, DL., Vandeginste, B. G., Deming, S. N., Michotte, Y. and Kaufman, L., *Chemometrics: a textbook*, Elsevier, New York, (1986).

[3] Morgan, E. , *Chemometrics Experimental Design* ,John Wiley & Sons , Baffins Lane, England, (1991).

[4] Workman, J. J. JR. , Kowalski, B. R., Bro, R., "Review of Chemometrics Applied to Spectroscopy; 1985 – 95, Part I ", Appl. Spectrosc. Rev., **31**, 73, (1996).

[5] Workman, J. J. JR. , Kowalski, B. R., Bro, R., "Review of Chemometrics Applied to Spectroscopy; 1985 – 95, Part II ", *Appl. Spectrosc. Rev.*, **31**, 347, (1996).

[6] Goldberg, D.E; *Genetic Algorithms in Search, Optimization, and Machine Learnig*, Addison-Wesley , Reading, (1989).

[7] Costadinnova, L. and Nedeltcheva, T.," Method for Selection of Optimum Wavelengths for Multicomponent Spectrophotometric Determinations", *Analyst*, **120**, 2217, (1995).

[8] Frenich, A. Garrido, Jouan-Rimbaud, D., Massart, DL., Kuttatharmmakul, S., Galera, M. M. and Vidal, J. L. M., "Wavelength Selection Method for Multicomponent Spectrophotometric Determinations Using Partial Least Squares" , *Analyst*, **120**, 2787, (1995).

[9] Centner, V. and Massart, DL., "Elimination of Uniformative Variables for Multivariate Calibration" , *Anal. Chem.*, **68**, 3851, (1996).

[10] Hörchner,U. and Kalivas, J. H., "Simulated-Annealing-Based Optimization Algorithms: Fundamentals and Wavelength Selection Applications", *J. Chemom.*, **9**, 283, (1995).

[11] Leardi ,R., Boggia, R. and Terrile, M., "Genetic Algorithms as a Strategy for Feature Selection", *J. Chemom.*, **6**, 267, (1992).

[12] Draper, N. , Smith, H. , *Applied Regression Analysis* , John Wiley & Sons , New York, (1981).

[13] Malinowski, E. R., *Factor Analysis in Chemistry* , John Wiley & Sons , New York, (1991).

[14] Martens, H. and Naes,T., *Multivariate Calibration*, John Wiley & Sons , New York, (1989).

[15] Geladi, P., Kowalski, B. R., "Partial Least-Squares Regression: A Tutorial", *Anal. Chim. Acta,* **185** ,1, (1986).

[16] Beebe, R.K, Kowalski, B.R, "An Introduction to Multivariate Calibration and Analysis", *Anal. Chem.*, **59**,17, 1007, (1987).

[17] Li. S, Hamilton, C. and Gemperline, P. J., " Generalized Rank Annihilation Method Using Similarity Transformations " , *Anal. Chem,* **64**,599, (1992).

[18] Rao, C. R. and Mitra, S. K., *Generalized Inverse of Matrices and it Applications* , Wiley, New York, (1971).

[19] Ruse, M., *Darwinism Defended. A Guide to the Evolution Controversies* John Wiley & Sons , New York, (1982).

[20] Lucasius, C. B. and Kateman, G. ,"Understanting and Using Genetic Algorithms Part I. Concepts, Properties and Context", *Chemom. Intell. Lab. Syst.* , **19** , 1 , (1993).

[21] Lucasius, C. B. and Kateman, G.,"Genetic algorithms for large scale optimization in chemometrics: an application", *Trends Anal. Chem.* , **10** , 254 , (1991).

[22] Holland, J. H., *Genetic Algorithms, Scientific American*, 44, (1992).

[23] Li, T. H., Lucasius, C. B. and Kateman,G., "Optimization of calibration data with the dynamic genetic algorithm", *Anal. Chim. Acta* , **268**, 123, (1992).

[24] Bruns, R. E., Scarminio, I. S., Neto, B. B., *Planejamento e Otimização de Experimentos* , Editora da Universidade Estadual de Campinas, (1995).

[25] Jang, J. S. R., Sun, C. T., Mizutani, E., *Neuro-Fuzzy and Soft Computing* , Prentice Hall Upper Saddle River, (1997).

[26] Message Passing Interface (MPI), Maui High Performance Computing Center Blaise Barney, (1996).

[27] Programming Languages and Tools : MPI, Cornell Theory Center, (1996).

[28] MPI : A message-Passing Interface Standard, University of Tennesse, Knoxville, Tenesse, (1994).

[29] Zupan, J. , Gasteiger, J. , *Neural Networks for Chemists*, VCH, Weinheim, (1993).

[30] Bos, M. and Weber, H. T. , "Comparison of training of neural networks for quantitative X-ray fluorescence spectrometry by genetic algorithm and backward error propagation", *Anal. Chim. Acta* , **247**, 97, (1991).

[31] Bangalore, A. S., Shaffer, R. E., Small, G. W. and Arnold, M. A., "Genetic Algorithm-Based Method for Selecting Wavelengths and Model Size for Use with Partial Least-Squares Regression : Application to Near-Infrared Spectroscopy", *Anal. Chem.*, **68**, 4200, (1996).

[32] Jouan-Rimbaud, D., Massart, DL., Leardi, R. and De-Noord, OE., "Genetic Algorithms as a tools for wavelength selection in multivariate calibration", *Anal. Chem.*, **67**, 4295, (1995).

[33] Fontain, E., "The problem of atom-to-atom mapping. An application of Genetic Algorithms" , *Anal. Chim. Acta* , **265**, 227, (1992).

[34] Broudiscou, A., Leardi, R., Phan-Tan-Luu,R. , "Genetic Algorithm as a tool for selection of D-optimal design", *Chemom. Intell. Lab. Syst.*, **35**, 105, (1996).

[35] Wienke, D., Lucasius, C., Ehrlich, M. and Kateman, G., "Multi-criteria target vector optimization of analytical procedures using a genetic algorithm. II. of the photometric calibration graph of dry glucose sensors for quantitative clinical analysis", *Anal. Chim. Acta*, **271**, 253, (1993).

[36] Hartnett, M. K., Bos, M., van-der-Linden, W. E. and Diamond, D., "Determination of stability constants using genetic algorithms", *Anal. Chim. Acta*, **316**, 347, (1995).

[71] Allan, J. E., "The determination of Zinc in agricultural materials by atomic absorption spectrophotometry", *Analyst*, **86**, 530, (1961).

[72] David, D. J., "Determination of Zinc and other elements in plants by atomic absorption spectrophotometry", *Analyst*, **83**, 655, (1958).

[73] Kuroka, R. and Mochisuki, T., "Continuous spectrophometry determination of copper, nickel and zinc in copper base alloys by flow injection analysis", *Talanta*, **28**, 389, (1981).

[74] Jones, G. B., "The polarographic determination of copper and zinc in plants and soils", Anal. Chim. Acta, **7**, 578, (1952).

[75] Menel, R. G. and Jackson , M. L., "Determiantion of copper and zinc in soils or plants", *Anal. Chem.*, **23**, 1861, (1951).

[76] Cowling, H. and Miller, E. J., "Determination of small amounts of zinc in plant material", *Ind. Eng. Chem. Anal. Ed.*, **13**, 145, (1941).

[77] Forster, W. A., "The Determination of copper in plant material", *Analyst*, **78**, 614, (1953).

[78] Jewsbury, A. , "The determination of Copper with sodium diethyldithiocarbamate in presence of nickel and other interfering elements", *Analyst*, **78**, 363, (1953).

[79] Claasen, A. and Bastings, L., "The photometric determination of copper by extraction as the diethyldithiocarbamate Interferences and their elimination", *Z. Anal. Chem.* , **153**, 30, (1953).

[80] Elliot, C. R. , "The determination of copper in certains high alloys steels", *Analyst*, **84**, 237, (1959).

[81] Hoste, J., "Spectrophotometric determination of copper with cuproine", *Anal. Chim. Acta*, **9**, 263, (1953).

[82] Malmstrom, B. G. and Glick, D., "Determination of Zinc in millimicrogram quantities", *Anal. Chem.*, **23**, 1699, (1953).

[83] Cholak, J., "Determination of Zinc in biological material", *Ind. Eng. Chem. Anal. Ed.*, **15**, 754, (1943).

[84] Martin, A., "Determination of Zinc with special reference to separation from cobalt and copper ", *Anal. Chem.*, **25**, 1853, (1953).

[85] Yoe, J. H. and Rush, R. M.," A new Colorimetric reagent for zinc", *Anal. Chim. Acta*, **6**, 526, (1952).

[86] Winzinzer, R. and Biro, V., "Uber Formazulkomplexe I", *Helv. Chim. Acta*, **32**, 901, (1949).

[87] Liu, R., Liu, D., Sun, A. and Liu, G. , "Simultaneous determination of Copper and Zinc in the hair of Children by pH Gradient construction in a Flow-Injection System", *Analyst*, **120**, 569, (1995).

APÊNDICE

Programa Principal do Algoritmo Genético

```
function [ pop_old ] = algo ( ncros,Num_var,pop_org,pm,pc,num_ge,lv,
conc_real,amost_desc,conc_test,Met_Sel );

% ncros ( tamanho da população de Cromossomos )
% Num_var ( número máximo de variáveis a serem selecionadas )
% pop_org  ( matriz de  espectros para a calibração )
% pm ( probabilidade de mutação )
% pc  ( probabilidade de crossover )
% num_ge ( número de gerações )
% lv  ( número de variáveis latentes )
% conc_real  ( concentração real do conjunto de calibração )
% amost_desc ( matriz com os dados dos espectros das amostras desconhecidas )
% conc_test   ( concentrações das supostas amostra desconhecidas )
% Met_Sel ( variável que define o método de calibração escolhido para usar no algoritmo
" PLS ou RLM " )

tcros  =  length ( new_gen );
inic = 0;
aux1 = pm;

%criando a primeira geração
pop_old  =  popini( ncros, tcros, Num_var );

%avaliação da primeira geração
incubadora = descod (pop_old,pop_org,lv,conc_real,amost_desc,conc_test,Met_Sel);


% Ordenando os valores de fitness em cada matriz de armazenagem
 for i = 1 : ncros-1
   for i = 1 : ncros-1
         aux = incubadora ( i,: );
         if incubadora( i,tcros1 ) > incubadora( i+1, tcros1 );
           incubadora ( i, : )  = incubadora ( i+1 , : );
           incubadora ( i+1 , : ) = aux;
         end
   end
 end

% separando os cromossomos dos seus fitness
 pop_old = incubadora ( : , 1 : tcros ) ;

% separando a matriz fitness
 fitness1 = incubadora ( : , tcros1 );
```

```
% entrando em um loop para as futuras gerações

while inic <= num_ge;
          inic = inic+1
       pm = aux1;
    if any ( fitness1 < 0.001 ) | ( inic > num_ge );
         break
    end
    if inic == 0.75*num_ge % gerando uma população nova
      pm = 0.35
    end


   % Inicializando processo de cruzamento entre cromossomos

     [ new_gen , pm] = cruz ( pop_old , pc , Num_var , pm );

   % Etapa de mutação dos cromossomos

     [ new_gen ] = mutação ( new_gen , pm );

     if inic > round (0.75*num_ge)
        [ best ] = halo ( ncros , tcros ,new_gen );
        new_gen = [ new_gen ; best];
      end


        %Purificação da raça

       [ l , c ] = size ( new_gen );
        i=1;

        while i <= l;

            if ( sum ( new_gen ( i , : ) ) == 0 | sum ( new_gen ( i , : ) ) > Num_var )
                 v1 = [];
                 v3 = [];
                 for a = 1 : i-1;
                    v1 ( a , : ) = new_gen ( a , : );
                 end

                 k=0;
                 for j = i+1 : l;
                    k = k + 1 ;
                    v3 ( k , : ) = new_gen ( j , : );
                 end
                 new_gen = [ v1 ; v3];
                 clear j k
              [ l , c ] = size ( new_gen ) ;
                 i = 0;
            end
```

```
        l = l+1;
        [ l , c ] = size ( new_gen );
        end
clear a i j k l v1 v3

% Etapa de avaliação das variáveis selecionadas

[ ncros , tcros ] = size ( new_gen ) ;
[ icre , ccre ] = size ( conc_test );
a = 0;
fitness1 = [];
        for i = 1 : ncros;
          fitness2 = [];
          y = find ( new_gen ( i , : ) );
          g = size ( y , 2 );
          if g < lv;
            lv1 = g;
          else
            lv1 = lv;
          end

        [ norg , torg ] = size ( pop_org );
          for j = 1 : g;
            for k = 1 : norg;
                var_sel ( k , j ) = pop_org ( k , y ( j ) );
                if var_sel ( k , j ) == 0
                  var_sel ( k , j ) = 1e-8;
                end
              end
          end

[ ndes , tdes ] = size ( amost_desc );

          for j1 = 1 : g ;
            for k1 = 1 : ndes;
              var_desc_sel ( k1 , j1 ) = amost_desc ( k1 , y ( j1 ) );
                if  var_desc_sel ( k1 , j1 ) == 0
                      var_desc_sel ( k1 , j1 ) = 1e-8;
                end
              end
          end

    %momento de seleção do método selecionado

     if   Met_Sel == ' pls ';

          [ p , q , w , t , u , b , ssqdif ] = pls1( var_sel ,conc_real , lv1 );
          ypred = plspred ( var_desc_sel , b , p , q , w , lv1 );
          [ lypred , cypred ] = size ( ypred );
          kk = 0;
          zz = 0;
```

```
        for ii = 1 : lypred; %inicio da subrotina para calculo do erro
          for m = 1 : ccre;
                kk  = ( conc_test ( ii , m ) - ypred ( ii , m ) )^2;
                zz  = kk + zz;
          end
          fitness ( ii , 1 ) = sqrt ( zz ) / cypred; %guardando o erro no fitness
        end

        fitness2 = (  sum ( fitness ( : , 1 ) ) /  lypred );
        fitness1( i , 1 ) = fitness2;

        if fitness1 ( i , 1 ) > 10
          fitness1 ( i , 1 ) = 20;
        end
  else
        p = pmatrix (  var_sel' , conc_real' ) ;
        prev =  p*var_desc_sel';
        prev = prev';
        [ lypred , cypred ] = size ( prev );
        kk = 0;
        zz = 0;

        for ii = 1 : lypred;  %inicio da subrotina para calculo do erro
           for m = 1:ccre;
              kk = ( conc_test ( ii , m ) - prev ( ii , m ) )^2;
                 zz  = kk + zz;
           end
           fitness ( ii , 1) = sqrt ( zz ) / cypred;
        end

        fitness2 = ( sum ( fitness ( : , 1 ) ) / lypred );
        fitness1 ( i , 1) = fitness2;
        if fitness1 ( i , 1 ) >10
          fitness1 ( i , 1 ) = 20;
        end
    end
    clear var_sel var_desc_sel fitness
  end
      new_gen = [ new_gen , fitness1 ];
   clear fitness1


%Comparação dos resultados e armazenagem
     [ii  cc ] = size ( new_gen );
       seleção = [ incubadora ; new_gen ];
 clear new_gen
```

```
[seleção] = delta(seleção , ncros ); % eliminação de indivíduos iguais

%reordenando matriz

[ l , c ] = size ( seleção );
for i = 1 : l-1
    for i = 1 : l-1
        aux = seleção ( i , : );
        If seleção ( i , tcros1 ) > seleção( i+1 , tcros1 );
        seleção ( i , : ) = seleção( i+1 , : );
            seleção( i+1 , : ) = aux;
        end
    end
    end


%corte da populações

    incubadora = seleção( 1 : ncros , : );
    clear seleção


% separando os cromossomos dos seus fitness

    pop_old = incubadora ( : , 1 : tcros) ;


% separando a matriz fitness

    fitness1 = incubadora ( : , tcros1 ) ;


clear a a1 a4 b3 d e f g h i ii j l m n n1 q r s t u var_desc_sel var_sel


end

end %fim do Programa
```