Quim. Nova, Vol. 25, No. 1, 46-52, 2002.

Artigo

APLICAÇÃO DE ALGORITMOS GENÉTICOS NA SELEÇÃO DE VARIÁVEIS EM ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO. DETERMINAÇÃO SIMULTÂNEA DE GLICOSE, MALTOSE E FRUTOSE

Paulo A. da Costa Filho e Ronei J. Poppi*
Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13083-970 Campinas - SP

Recebido em 6/12/00; aceito em 11/6/01

APLICATION OF GENETIC ALGORITHMS IN THE VARIABLE SELECTION IN MID INFRARED SPECTROSCOPY. SIMULTANEOUS DETERMINATION OF GLUCOSE, MALTOSE AND FRUCTOSE. Genetic algorithm was used for variable selection in simultaneous determination of mixtures of glucose, maltose and fructose by mid infrared spectroscopy. Different models, using partial least squares (PLS) and multiple linear regression (MLR) with and without data pre-processing, were used. Based on the results obtained, it was verified that a simpler model (multiple linear regression with variable selection by genetic algorithm) produces results comparable to more complex methods (partial least squares). The relative errors obtained for the best model was around 3% for the sugar determination, which is acceptable for this kind of determination.

Keywords: genetic algorithm; sugars determination; infrared spectroscopy.

INTRODUÇÃO

Uma das primeiras aplicações da espectroscopia no infravermelho como ferramenta analítica, foi durante o período da segunda guerra mundial[1]. Nesta ocasião, esta técnica foi usada no setor de controle de qualidade em algumas indústrias químicas alemãs. Entretanto, este tipo de análise foi rapidamente substituída pela cromatografia gasosa e cromatografia liquida de alta eficiência (HPLC), devido ao fato destas ferramentas de análise viabilizarem a realização de determinações multicomponente quantitativas e qualitativas em amostras complexas de interesse industrial de maneira mais eficiente. Estas vantagens agregaram um alto valor à cromatografia, tornando-a rapidamente uma ferramenta padrão de análise para os mais diversos tipos de análises químicas.

Contudo, sabia-se que os espectros no infravermelho armazenavam uma grande gama de informações sobre a amostra e portanto, apresentavam um elevado potencial para serem empregados nos mais diversos tipos de análises químicas e\ou físicas. Entretanto, até duas décadas atrás era praticamente impossível extrair informações quantitativas a partir dos espectros no infravermelho. Devido a este fato, a espectroscopia no infravermelho restringiu-se basicamente a aplicações qualitativas ou para reforçar hipóteses propostas sobre a estrutura química das espécies.

Nos meados dos anos oitenta, uma série de fatos nas áreas científicas e tecnológicas contribuíram para a inversão deste quadro. Dentre estes podemos destacar o desenvolvimento da microeletrônica e a popularização dos microcomputadores, que proporcionaram um significativo avanço nas análises instrumentais, possibilitando a aquisição de maneira fácil e rápida de um grande número de dados de uma mesma amostra. Conseqüentemente, o tratamento dos dados obtidos passou a exigir modelos mais complexos que iam além da tradicional calibração univariada. O problema da modelagem destes dados foi solucionado com a aplicação de técnicas quimiometricas[2]. Portanto, a quimiometria também pode ser considerada como uma das fortes razões que contribuíram para a utilização da espectroscopia como uma ferramenta de análise em aplicações qualitativas e quantitativas na química analítica.

* e-mail:ronei@iqm.unicamp.br

A crescente preocupação mundial com relação à questão ambiental é outro aspecto de grande relevância que tem incentivado o desenvolvimento e aperfeiçoamento das análises espectroscopicas no infravermelho. Estas análises, além de fornecerem os resultados de maneira mais rápida, não são destrutivas e invasivas, assim como não geram subprodutos químicos tóxicos. Devido à estas vantagens, este tipo de análise tem sido aplicada no monitoramento em linha de sistemas químicos industriais[3], na determinação de glicose[4], colesterol e triglicerídeos em plasma sanguíneo[5], no auxílio de identificação de tumores em células[6], no controle de qualidade[7], na determinação de nitrogênio em plantas[8], na indústria de polímeros[9], no estudo da composição química de solos[10], na adulteração da composição química de combustíveis[11], óleos comestíveis e alimentos[12], etc.

Como pode-se observar, hoje em dia há um vasto número de trabalhos utilizando a espectroscopia no infravermelho em análises quantitativas. Entretanto, neste momento depara-se com um novo problema: apesar da espectroscopia no infravermelho fornecer um grande número de dados, parte destes não possuem informações correlacionadas diretamente com o(s) análito(s) de interesse. Isso pode ocasionar distorções ao modelo, e conseqüentemente a conclusões errôneas nas análises.

Para minimizar este problema tem sido utilizado diversos métodos de pré-processamento[13], como escalamento, utilização de derivadas e filtragem digital para remoção de ruído. Para minimizar o ruído, vários trabalhos têm proposto o uso do filtro de média móvel[14], transformada de Fourier[15], transformada de Wavelet[16] ou Savitsky-Golay[14]. A derivada[13] vem sendo empregada freqüentemente para melhorar a definição de bandas que se encontram sobrepostas em uma mesma região espectral e para correção de linha base.

Outro procedimento que vem sendo aplicado é a seleção de variáveis, a qual permite eliminar os termos que não são relevantes na modelagem. Isso gera um sub-conjunto com o melhor número de variáveis, e que apresente maior sensibilidade e linearidade para o(s) analito(s) de interesse. Desta maneira, este procedimento minimiza ou até mesmo elimina características potenciais dos interferentes, bem como não-linearidades.

Recentemente tem-se observado uma tendência na aplicação do algoritmo genético na seleção de variáveis[17], entretanto, vale ressaltar que existem outros métodos de seleção de variáveis[18-21]. A prefe-

rência pelo algoritmo genético deve-se a sua eficiência, versatilidade e robustez [17].

A seleção de variáveis permite o uso 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 eram menor ou igual ao número de amostras[22]. Uma explicação mais detalhada sobre o algoritmo genético pode ser obtida em um artigo publicado pelo autores na revista *Química Nova*[23].

Neste trabalho foi realizada a determinação quantitativa simultânea de glicose, maltose e frutose em solução. Para tanto utilizou-se a técnica de reflexão atenuada (ATR), para a aquisição dos espectros no infravermelho médio. A reflexão total atenuada baseia-se no fenômeno da reflexão total da radiação na interface de matérias com índices de refração diferentes. Esta 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[24-28]. Este tipo de problema é comumente verificado em situações onde o solvente (água, por exemplo) comporta-se como interferente, por apresentar estiramentos característicos na mesma região de absorção da espécie de interesse, inviabilizando a análise pelos métodos tradicionais.

O interesse na quantificação destes açúcares deve-se ao fato de que estes são parte integrante e essencial do reservatório nutricional dos animais (em animais superiores é essencial a presença de glicose no sangue) e estrutural das plantas. Também possuem um grande interesse industrial, visto o grande número de aplicações destas espécies químicas nos mais diversos segmentos.

O objetivo deste trabalho é apresentar a potencialidade do algoritmo genético na seleção de variáveis em dados espectroscópicos no infravermelho, em problemas onde os espectros dos analitos estudados possuem um alto grau de similaridade. Além disso, abre a perspectiva de poder-se construir um modelo matemático mais simples, eficiente e robusto utilizando a Regressão Linear Múltipla. Também pretende-se apresentar o uso de de técnicas quimiométricas na resolução de problemas químicos que até então apresentavam-se insoluvéis ou extremamente complexos de serem resolvidos pelo métodos tradicionais.

PARTE EXPERIMENTAL

A fase preliminar à preparação das amostras envolveu um planejamento experimental[29-30] para determinar o melhor conjunto de valores de concentração para a preparação de 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. As soluções foram preparadas por pesagem dos açúcares em uma balança analitica e posterior diluição em água deonizada.

Para obter-se o número ideal de amostras para o experimento, optou-se por realizar 3 planejamentos experimentais distintos. Inicialmente foi preparado um planejamento experimental com 3 níveis e três variáveis totalizando um número total de 27 experimentos, conforme ilustrado na Figura 1.

Em seguida foi realizado um segundo planejamento na forma de um cubo menor (Figura 2) contendo 14 amostras, que se encontra no interior dos limites do primeiro cubo (Figura 1).

O terceiro planejamento envolveu a preparação de mais oito amostras, formando um terceiro cubo menor (Figura 3), o qual encontra-se no interior dos limites dos dois primeiros cubos (Figuras 1 e 2).

Para totalizar o número de 64 amostras, foram preparadas mais 15 misturas das soluções dos açúcares de forma aleatória dentro do

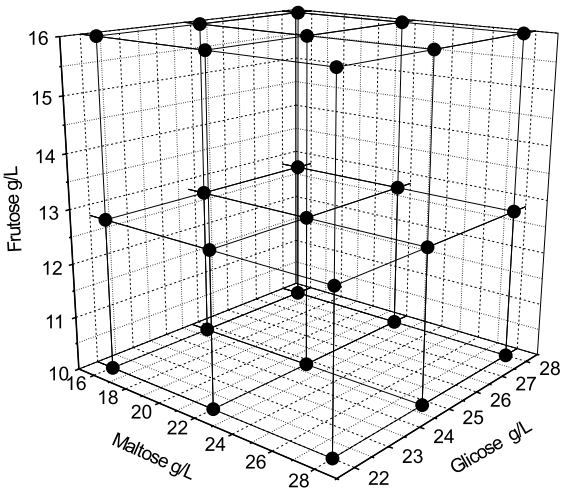

Figura 1. *Representação do planejamento experimental completo da mistura de Açúcares com três variavéis e três niveis.*

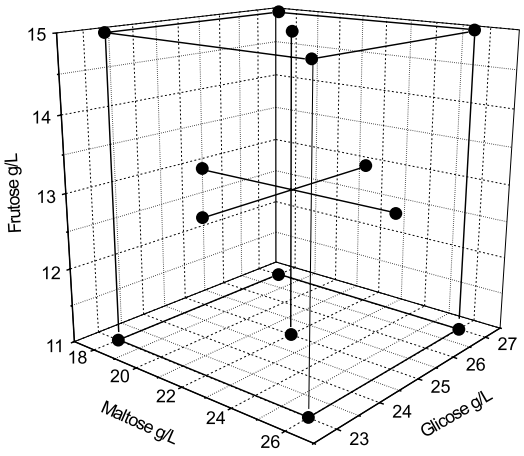

Figura 2. *Representação do planejamento experimental incompleto da mistura de Açúcares, com três variavéis e três níveis.*

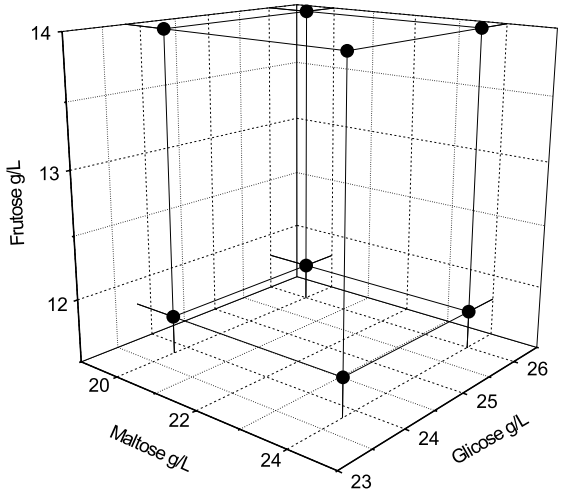

Figura 3. *Representação do planejamento experimental incompleto da mistura de Açúcares, com três variavéis e dois níveis.*

limite superior (cubo maior) e limite inferior de concentração dos 3 analitos (cubo menor).

A aquisição dos espectros foi realizada em um espectrofotômetro de infravermelho Nicolet 520 FT-IR, utilizando a técnica de reflexão total atenuada (ATR), empregando um acessório do tipo bote com um cristal de seleneto de zinco. Os espectros foram obtidos utilizando 32 leituras por espectro; resolução de 2 cm^{-1}, região de aquisição de 4000 a 400 cm^{-1}, em uma câmara purgada com nitrogênio. Foi utilizado o ar como espectro de referência.

Para o desenvolvimento dos modelos de calibração foram escolhidas 24 amostras para a construção do modelo de calibração (fase de calibração) e 19 amostras para a fase de seleção das variáveis (fase de validação). 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, composto de 17 amostras nesse caso, este tipo de problema pode ser detectado.

Parâmetros de Configuração do Algoritmo Genético

O algoritmo genético foi iniciado com uma população inicial de 100 cromossomos, com um número de gerações igual a 100, probabilidade de cruzamento de 90%, probabilidade de mutação de 1%, erro máximo para a finalização do processo de 1% e número máximo de variáveis selecionadas igual a 10. Essa mesma configuração foi executada para a seleção de variáveis utilizando o método dos mínimos quadrados (PLS)[31] e regressão linear multipla (RLM)[32]. Para o métodos dos mínimos quadrados empregou-se o número de componentes principais igual a 4.

RESULTADOS E DISCUSSÃO

Análise da Coleção de Espectros das Soluções de Açúcares e Seleção do Número de Componentes Principais

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

Realizando a análise das componentes principais[13,14] PCA (Figura 5), verifica-se que as amostras com absorbância acima de 0,7, que são as amostras 2, 26, 27 e 48, podem ser consideradas como anômalas (amostras que possuem alguma espécie de comportamento diferenciado da população da qual foram extraidas). O gráfico da primeira componente principal contra a segunda componente principal, mostra que estas amostras possuem valores de escores na segunda componente principal bem superior que as demais amostras. Isso leva a concluir que elas podem não pertencer a mesma população dos demais sessenta espectros obtidos.

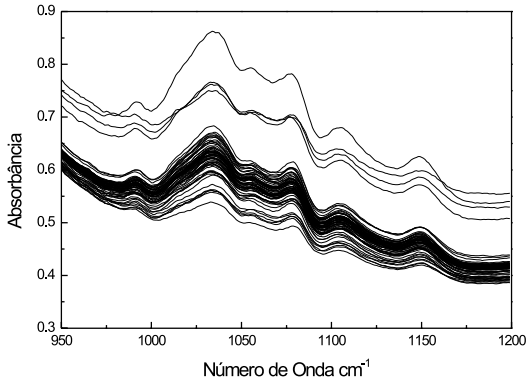

***Figura 4**. Coleção de espectros no infravermelho das 64 soluções de açúcares.*

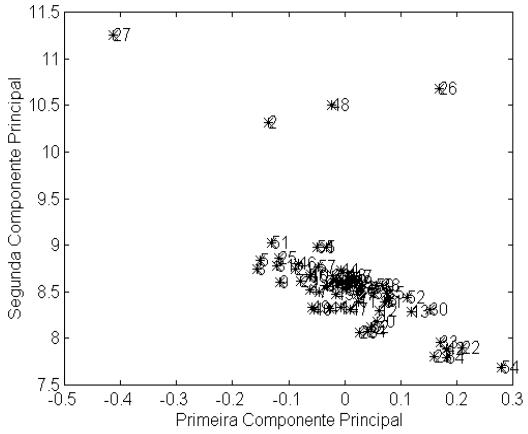

***Figura 5**. Análise das componentes principais das 64 amostras de açúcares.*

Por precaução, decidiu-se excluir estas amostras do estudo de quantificação dos açúcares, pois o comportamento diferenciado do perfil destas amostras com relação as demais sugere que estes espectros tiveram algum problema na sua aquisição, ou então podem possuir um comportamento não linear[33]. Portanto, a inclusão destes espectros pode ocasionar distorções no modelo de calibração, comprometendo seriamente os resultados de previsão do modelo.

A análise quantitativa dos açúcares envolveu a construção de seis modelos de calibração para avaliar efeitos como : influência do pré-tratamento dos dados nos resultados de previsão do modelo de calibração e importância do método de calibração utilizado (PLS ou RLM) no desempenho do modelo de calibração.

Inicialmente serão apresentados os resultados obtidos pelos modelos de calibração com a utilização dos dados sem nenhum pré-tratamento para o PLS com e sem seleção de variaveis e para a RLM com seleção de variavéis. Posteriormente, serão mostrados os resultados obtidos após a utilização de pré-tratamento dos dados para o PLS com e sem seleção de variavéis e para a RLM com seleção de variavéis. No final uma comparação dos resultados obtidos será realizada.

Modelos dos Mínimos Quadrados Parciais sem Pré-Tratamento dos Dados

Um estudo prévio utilizando-se validação cruzada[34], indicou que o número ideal das componentes principais para este conjunto de amostras seriam quatro, pois a partir desse número não existe alteração significativa no valor do erro de previsão, conforme mostrado na Figura 6.

Para que se possa comparar os resultados entre os modelos que utilizam o método dos mínimos quadrados parciais com e sem seleção de variavéis, empregou-se quatro componentes principais para todos os modelos de calibração. Optou-se por manter fixo o número de componentes principais, devido a dificuldade de determiná-los durante o processo de seleção de variáveis. Neste tipo de otimização, para obter o número de componentes principais ideais, seria necessário realizar uma validação cruzada para cada cromossomo gerado,

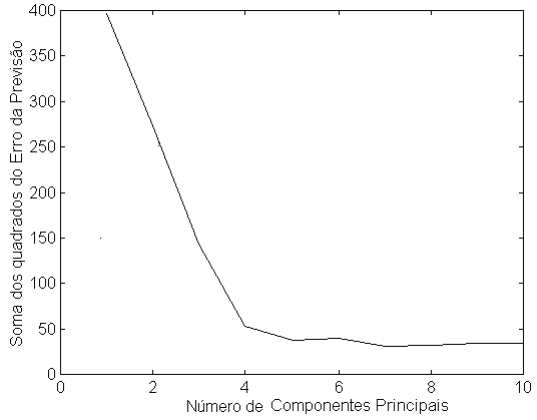

Figura 6. *Determinação do Número das Componentes Principais ideais, utilizando validação cruzada.*

tornando praticamente inviável o processo, devido ao tempo requerido de processamento.

A Figura 7 apresenta os erros relativos das determinações dos 3 açúcares analisados para o conjunto teste do modelo de calibração, utilizando a região dos espectros compreendida entre 950 a 1200 cm^{-1}. 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[35].

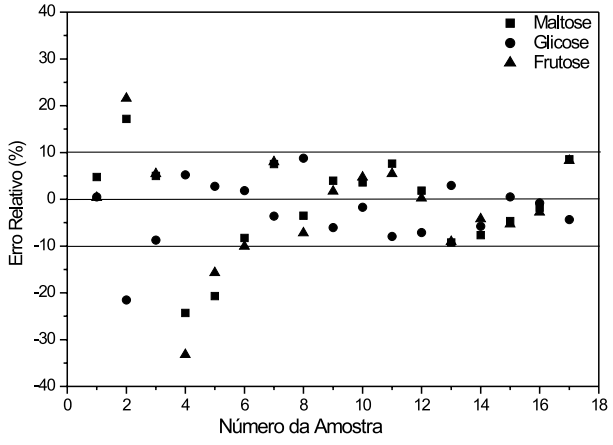

Figura 7. *Erro relativo para o conjunto teste do modelo de calibração, utilizando a região dos espectros compreendida entre 950 a 1200 cm^{-1}.*

O erro padrão de previsão (equação 1) é comumente empregado, já que representa um erro médio do modelo com as mesmas unidades da(s) propriedade(s) estimada(s), sendo mais sensível à presença de amostras com erros elevados.

O erro padrão de previsão é calculado como:

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

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

A Figura 7 evidencia que o modelo de calibração apresentou melhor desempenho para a previsão da concentração de glicose, pois apenas uma das amostras possue erro relativo superior a 10% na determinação deste analito.

Ainda pode-se observar na figura 7, que o erro relativo de previsão de maltose e frutose parecem estar correlacionados de alguma forma, já que apresentam o mesmo comportamento com relação ao erro. Curiosamente o erro de previsão da glicose possue um comportamento inverso, ou seja, erros relativos positivos nas concentrações de glicose implicam em erros negativos nas concentrações de maltose e frutose, e vice versa. Pode-se associar este comportamento a algum tipo de erro sistemático detectado ou criado pelo modelo de calibração.

Os erros padrão de previsão encontrados para o conjunto de validação foram: 2,10 gL^{-1} para a maltose, 2,67 gL^{-1} para a glicose e 1,11 gL^{-1} para a frutose. Já para o conjunto teste, os erros padrão de previsão foram: 2,24 gL^{-1} para a maltose, 1,67 gL^{-1} para a glicose e 1,38 gL^{-1} para a frutose. Neste caso, observa-se que não existe grande diferença entre os valores do SEP calculados para o conjunto de validação e teste, indicando que o modelo está robusto.

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. Neste estudo utilizou-se o espectro na faixa entre 950 a 1200 cm^{-1} , sem nenhum pré-tratamento dos dados.

No final do processo de seleção das variáveis, foi selecionado o melhor conjunto de comprimentos de onda que minimizam o erro no processo de validação do modelo de calibração. Os seguintes números de onda foram selecionadas: 1036; 1038; 1051; 1062; 1109; 1117; 1118; 1120; 1148; 1168 cm^{-1}.

A Figura 8 apresenta os erros relativos para as amostras do conjunto teste, obtidos neste caso. Após a seleção de variavéis, os erros relativos de previsão dos três analitos apresentam o mesmo tipo de comportamento sistemático, observado no modelo do PLS sem seleção de variavéis. Contudo, para este novo modelo, observou-se que a glicose apresenta o mesmo tipo de comportamento que os demais açúcares.

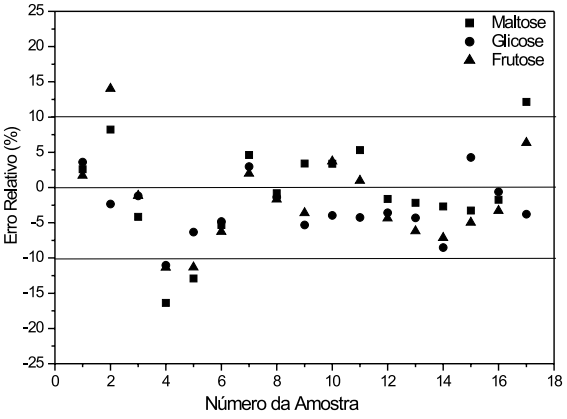

Figura 8. *Erro relativo para o conjunto teste do modelo de calibração, utilizando o algoritmo genético na seleção de variavéis.*

Este comportamento, indica que provavelmente haja uma relação direta entre os erros de previsão dos três analitos. Isso pode estar associado ao fato dos açúcares possuírem estruturas químicas parecidas. A glicose, maltose e frutose têm espectros no infravermelho similares, devido aos modos vibracionais do anel dos açúcares serem os principais responsáveis pelo perfil dos espectros[36].

Os erros padrão de previsão para o conjunto de validação foram: 0,93 gL^{-1} para a maltose, 1,06 gL^{-1} para a glicose e 0,66 gL^{-1} para a frutose. Para o conjunto teste obteve-se os seguintes erros de previsão: 1,45 gL^{-1} para a maltose, 1,26 gL^{-1} para a glicose e 0,82 gL^{-1} para a frutose.

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

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 para a modelagem. Para este modelo foram selecionadas as seguintes variáveis: 975; 996; 1036; 1039; 1068; 1071; 1100; 1103; 1148; 1164 cm^{-1}.

A Figura 9 apresenta o erro relativo para as amostras do conjunto teste. Os erros padrão de previsão encontrados para o conjunto de validação foram: 0,67 gL^{-1} para a maltose, 1,06 gL^{-1} para a glicose e 0,56 gL^{-1} para a frutose. Para o conjunto teste os erros padrão de previsão foram: 1,08 gL^{-1} para a maltose, 1,18 gL^{-1} para a glicose e 0,70 gL^{-1} para a frutose. Pode-se notar que com a RLM houve uma melhora nos resultados em relação ao PLS, onde observa-se erros relativos sempre menores que 10%, assim como SEP inferiores para os três açúcares.

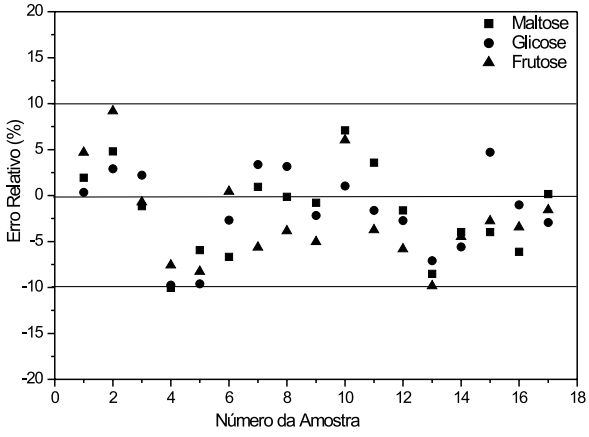

Figura 9. *Erro relativo para o conjunto teste do modelo de calibração, utilizando o algoritmo genético na seleção de variavéis na regressão linear múltipla.*

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[13,35]. Basicamente o filtro de Savisky-Golay ajusta um polinômio a uma certa janela com um número fixo de pontos do espectro. Em seguida, 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 segunda ordem.

Observando os resultados das Figura 10 verifica-se certa melhora na previsão das concentrações dos açúcares, comparado aos mesmos resultados obtidos sem o pré-tratamento dos dados e sem a seleção de variáveis. Neste caso, os erros padrão de previsão para o conjunto de validação foram 1,03 gL^{-1} para a maltose, 1,05 gL^{-1} para a glicose e 0,52 gL^{-1} para a frutose. Para o conjunto teste os erros foram 1,16 gL^{-1} para a maltose, 0,86 gL^{-1} para a glicose e 0,55 gL^{-1} para a frutose.

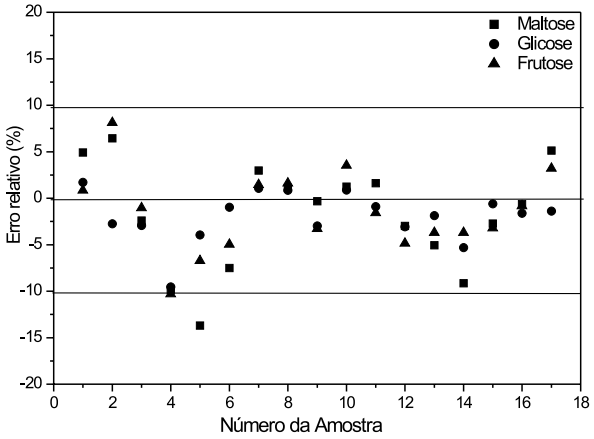

Figura 10*. Erro relativo para o conjunto teste do modelo de calibração, utilizando o método dos mínimos quadrados parciais.*

A diferença significativa observada entre os resultados sem e com pré-tratamento dos dados (antes da seleção de variáveis) reforça a necessidade da aplicação de pré-tratamento dos dados quando se utiliza um grande número de variavéis para a construção do modelo de calibração.

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

As variáveis selecionadas pelo algoritmo genético a partir de um conjunto de dados pré-tratados foram:1002; 1055; 1066; 1121; 1125; 1141; 1165 cm^{-1}. Observa-se que estas variáveis selecionadas são concordantes com as variáveis selecionadas sem o pré-tratamento dos dados, já que os números de onda selecionados encontram-se em regiões bastante próximas. Entretanto também observa-se 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 da Figura 11 e dos erros padrão de previsão que para o conjunto de validação foram : 0,77 gL^{-1} para maltose, 1,03 gL^{-1} para a glicose e 0,44 gL^{-1} para a frutose. Já para o conjunto teste os erros foram : 0,78 gL^{-1} para a maltose, 1,04 gL^{-1} para a glicose e 0,56 gL^{-1} para a frutose. Mais uma vez, observa-se que os erros obtidos após o pré-tratamento são menores e que com a seleção de variáveis os valores de SEP também diminuem.

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

A Figura 12 apresenta os erros relativos obtidos após utilizar o modelo construído a partir das seguintes variáveis selecionadas pelo algoritmo genético: 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%. Além disso, os erros padrão de previsão para o conjunto de validação foram: 0,69 gL^{-1}

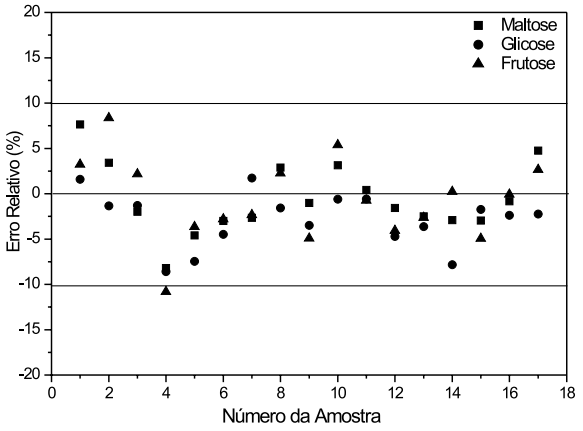

Figura 11. *Erro relativo para o conjunto teste do modelo de calibração, utilizando o algoritmo genético na seleção de variavéis no método dos mínimos quadrados parciais.*

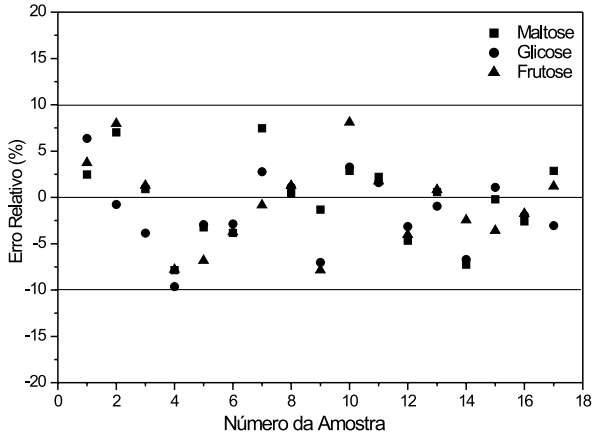

Figura 12. *Erro relativo para o conjunto teste do modelo de calibração, utilizando o algoritmo genético na seleção de variavéis para a regressão linear múltipla.*

para a maltose, 0,99 gL^{-1} para a glicose e 0,48 gL^{-1} para a frutose. Para o conjunto teste os erros foram: 0,90 gL^{-1} para a maltose, 1,08 gL^{-1} para a glicose e 0,58 para a frutose.

ANÁLISE DOS RESULTADOS OBTIDOS

Como pode-se observar pelos resultados apresentados, a seleção de números de onda proporciona uma significativa melhora nos resultados dos modelos de calibração multivariada. Além disso, é possível constatar que os resultados da regressão linear múltipla apresentaram desempenho comparável aos fornecidos pelos modelo de os mínimos quadrados parciais com seleção de variáveis, evidenciando a viabilidade do uso de um modelo matemático mais simples na quantificação da mistura de açúcares.

Para realizar uma comparação mais rigorosa dos resultados obtidos, pode-se empregar o teste F[30]. Este procedimento visa mostrar se as diferenças obtidas entre os resultados para o conjunto teste dos diversos modelos são realmente significativas. Desta forma é possível avaliar o desempenho de diferentes modelos para um mesmo conjunto de dados.

O teste F foi usado da seguinte maneira:

$$F(n, j) = \left(\frac{SEP_{Padrão}}{SEP_{AG\_XX}} \right)^2 \quad (2)$$

onde:

- SEP_{AG_XX} representa o valor obtido para o erro padrão de previsão dos modelos que utilizaram o algoritmo genético para a seleção de variáveis.
- $SEP_{Padrão}$ representa o valor obtido para o erro padrão de previsão dos modelos que utilizaram somente o PLS (sem a seleção de variáveis).
- "n" e "j" indicam o número de amostras utilizadas para calcular o SEP do conjunto teste, utilizando o algoritmo genético (AG) e para o método padrão de referência de calibração, respectivamente.

A Tabela 1 apresenta os valores do teste F para os modelos de calibração dos mínimos quadrados parciais e regressão linear multipla sem pré-tratamento dos dados. Para este conjunto de dados, com 95 % de confiança, o valor crítico para o teste F é 2.27.

Tabela 1. Valores do teste F para os modelos de calibração sem pré-tratamento dos dados.

Modelo	Maltose	Glicose	Frutose
PLS	2.39	1.76	2.83
RLM	4.30	2.00	3.89

Após a realização do teste F com 95% de confiança para os resultados de SEP do conjunto teste dos diversos modelos construídos, constatou-se que apenas para a glicose, quando não se realiza nenhum pré-tratamento, a seleção de variáveis não conferiu uma melhora significativa nos resultados, conforme se pode observar na Tabela 1. O teste F reitera a expectativa que a aplicação do algoritmo genético em dados sem pré-tratamento pode resultar em uma melhora nos resultados.

Posteriormente, o teste F foi aplicado aos dados pré-tratados, e os resultados são mostrados na Tabela 2. O teste F mostrou que os modelos construídos com e sem seleção de variáveis dos dados pré-tratados não apresentaram diferença significativa.

Tabela 2. Valores do teste F para os modelos de calibração com pré-tratamento dos dados.

Modelo	Maltose	Glicose	Frutose
PLS	2.21	0.68	0.96
RLM	1.66	0.63	0.90

Os resultados obtidos pela regressão linear múltipla com a seleção de variáveis são semelhantes aos com seleção envolvendo o método dos mínimos quadrados parciais. Também pode-se constatar que os resultados da RLM são similares quando os dados são pré-processados ou não. Isso mostra que os resultados podem independer da sofisticação do modelo matemático e portanto, reforçando a tese de que modelos mais simples e sem pré-tratamento podem fornecer resultados tão bons quando os mais complexos. Contudo, os leitores mais céticos podem argumetar que o algoritmo genetico por si só é mais complexo que o PLS, o que não justificaria sua aplicação. Entretanto, é necessário lembrar que a utilização do algoritmo genético é mais simples e pode ser realizada sem a interveção ou ajuda do operador, o que confere uma grande vantagem para ser utilizado quando exista pouco, ou nenhum, conhecimento sobre técnicas quimiometricas. Por outro lado, para a utilização de modelos como o dos mínimos quadrados parciais é necesário um conhecimento mínimo do método.

Na Figura 13 estão indicados os números de onda selecionados pelo algoritmo genético para a regressão linear múltipla, quando se utiliza a primeira derivada dos espectros nos cálculos.

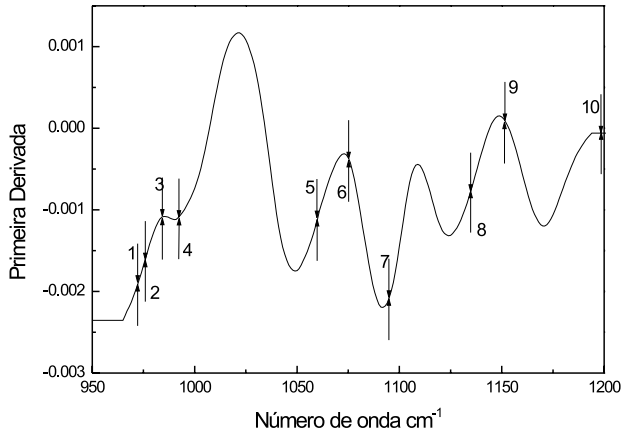

***Figura 13**. Primeira derivada do espectro da mistura de açúcares, com as variáveis selecionadas pelo algoritmo genético.*

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[33,37]. Assim, a utilização de espectroscopia no infravermelho pode tornar-se interessante para determinações "on-line" em processo industrial.

A Tabela 3 apresenta uma tentativa de atribuição dos números de onda selecionados para a regressão linear múltipla, relacionando-os com os respectivos açúcares. 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 [33,37].

Tabela 3. Atribuição dos números de onda selecionados.

	Número de onda Selecionado (cm^{-1})	Estiramento das Espécies puras (cm^{-1})	Atribuição	
1	972	972	C-C	(Frutose)
2	976	972	C-C	(Frutose)
3	985	984	C-C	(Glicose)
4	992	984	C-C	(Glicose)
5	1060	1056,1065	C-O-C	(Frutose, Maltose)
6	1075	1080	C-O-C, C-O	(Maltose)
7	1095	1092	C-O	(Maltose)
8	1135	1136	C-O	(Glicose)
9	1151	1148	C-O	(Glicose)
10	1199	1192	O-C-H, C-C-H, C-O-H	(Frutose)

CONCLUSÃO

O algoritmo genético mostrou-se uma ferramenta poderosa, no que tange à robustez dos modelos. A robustez é decorrente de um modelo consistente, o qual apresenta pequenas variações no desvio padrão dos erros, durante a previsão de novas amostras. Esta é uma importante característica, pois para modelos robustos, algumas vezes torna-se possível realizar extrapolações.

A robustez do modelo, utilizando o algoritmo genético, pode ser constatada ao comparar-se os resultados apresentados para o conjunto de validação e teste, já que os erros não apresentam elevada discrepância em nenhum dos modelos relatados 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.

A determinação de açúcares por espectroscopia no infravermelho médio com reflexão total atenuada 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. Sendo esta uma característica muito importante, pois torna o uso de ferramentas matemáticas mais simples como a regressão linear, além de dispensar o pré-tratamento dos dados .

REFERÊNCIAS

1. Coates, J. P.; *Appl. Spectrosc. Rev.* **1996**, *31* , 179.
2. Massart, D. L.; Vandeginste, B. G.; Deming, S. N.; Michotte, Y.; Kaufman, L.; *Chemometrics: a textbook*; Elsevier, New York, 1986.
3. Hammond, R. P.; *Proc. Control Qual.* **1997**, *9*, 117.
4. Hazen, K. H.; Arnold, M. A.; Small, G. W.; *Anal. Chim. Acta* **1998**, *371*, 255.
5. Heise, H. M.; Marbach, R.; Bittner, A.; *J. Near Infrared Spectrosc*. **1996**, *6*, 361.
6. Haaland, D. M.; Jones, D. T. H.; Thomas, E. V.; *Appl. Spectrosc*. **1997**, *51*, 340.
7. Boulou, J. C.; *Analusis* **1998**, *26*, M46.
8. Mello, C.; Poppi, R. J.; de Andrade, J. C.; Cantarella, H.; *Analyst* **1999**, *124*, 1669.
9. Urban, M. W.; Allison, C .L.; Johnson, G. L.; DiStefano, F.; *Appl. Spectrosc.* **1999**, *53*, 1520.
10. Messerschmidt I.; Cuelbas, C. J.; Poppi, R. J.; de Andrade, J. C.; de Abreu, C. A.; Davanzo, C. U.; *J. Chemom.* **1999**, *13*, 265.
11. Guchardi, R.; da Costa Filho, P. A.; Poppi, R. J.; Pasquini, C.; *J. Near Infrared Spectrosc.* **1998**, *6*, 333.
12. Kangming, M.; van de Voort, F. R.; Ismail, A. A.; Zhuo, H.; Cheng, B.; *J. Am. Oil Chem. Soc.* **2000**, *77*, 681.
13. Martens, H.; Naes,T.; *Multivariate Calibration*; Wiley; New York, 1989.
14. Malinowski, E. R.; *Factor Analysis in Chemistry*; Wiley; New York, 1991.
15. Cerqueira, E. O.; Poppi, R. J.; Kubota, L. T.; Mello, C.; *Quim. Nova* **2000**, *23*, 690.
16. Walczak, B.; Massart, D. L.; *Chemom. Intell. Lab. Syst.* **1997**, *36*, 81.
17. Goldberg, D.E; *Genetic Algorithms in Search, Optimization, and Machine Learnig*; Addison-Wesley; Reading, 1989.
18. Costadinnova, L. ; Nedeltcheva, T.; *Analyst* 1995,*120*, 2217.
19. Frenich, A. G. ; Jouan-Rimbaud, D.; Massart, D.L.; Kuttatharmmakul, S.; Galera, M. M.; Vidal, J. L. M.; *Analyst* **1995**,*120*, 2787.
20. Centner, V.; Massart, DL.; *Anal. Chem*. **1996**, *68*, 3851.
21. Hörchner,U.; Kalivas, J. H.; *J. Chemom.* **1995**, *9*, 283.
22. Leardi, R.; Boggia, R.; Terrile, M.; *J. Chemom.* **1992**, *6*, 267.
23. Costa Filho, P. A. da; Poppi, R. J.; *Quim. Nova* **1999**, *22*, 405.
24. Mirabella Jr., F. M.; *Appl. Spectrosc. Rev.* **1985**, *21* ,45.
25. Miller, M. P.; *Appl. Spectrosc. Rev.* **1987**, *25*, 329.
26. Göbel, R.; Krska, R.; Kellner, R.; Seitz, R. W.; Tomellini, S. A.; *Appl. Spectrosc.* **1994**, *48* , 678.
27. Pike, P. R.; Sworan, P. A.; Cabaniss, S. E.; *Anal. Chim. Acta* **1993**, *280*, 253.
28. Bayada, A.; Lawrance, G. A.; Maeder, M. ; Molloy, K. J.; *Appl. Spectrosc.* **1995**, *49*, 1789.
29. Morgan, E.; *Chemometrics Experimental Design*; Wiley, Baffins Lane,1991.
30. Bruns, R. E.; Scarminio, I. S.; Neto, B. B.; *Planejamento e Otimização de Experimentos* ; Editora da Universidade Estadual de Campinas, Campinas, 1995.
31. Geladi, P.; Kowalski, B. R.; *Anal. Chim. Acta* **1986**,*185* ,1.
32. Draper, N.; Smith, H.; *Applied Regression Analysis*; Wiley, New York,1981.
33. Mirouze, F. L.; Boulou, J. C.; Dupuy, N.; Meurens, M.; Huvenne, J. P.; Legrand, P.; *Appl. Spectrosc.* **1991**,*47* , 1187.
34. Geladi, P.; Kowalski; *Anal. Chim. Acta* **1986**, *185*, 19.
35. Willians, P.; *Near-Infrared Technology in The Agricultural and Food Industries;* American Association of Cereal Chemists, St. Paul, Minnesota, USA, 1990.
36. Cadet, F.; Bertrand, D.; Robert, P.; Maillot, J.; Dieudonné, J.; Rouch, C.; *Appl. Spectrosc*. **1991**, *45* , 166.
37. Bellon-Maurel, V.; Vallat, C.; Goffinet, D.; *Appl. Spectrosc.* **1995**, *49* , 556.