Modelos Baseados em Espectros Raman para Avaliação de Toxicidade de Misturas Etanol-Metanol

Rafael E. de Góes, Bárbara R. Heidemann, Lúcia Valéria R. de Arruda, Marcia Muller, José L. Fabris
Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial
Universidade Tecnológica Federal do Paraná
Curitiba, Brasil
rgoes@utfpr.edu.br

***Resumo*—Este trabalho apresenta os resultados preliminares obtidos usando a técnica de reconhecimento de padrões para a determinação de toxicidade da mistura etanol-metanol a partir do espectro de espalhamento Raman. Foram usadas no trabalho, primeiramente, uma Rede Neural Artificial do tipo *Multi Layer Perceptron* e a Análise de Componente Principal. Para misturas binárias sem reação entre os constituintes, mantendo-se as etapas de pré-processamento e usando ambos os métodos de maneira isolada, é possível discriminar a toxicidade. Para amostras contendo outros compostos, a combinação desses métodos pode ser utilizada para determinação em tempo real de toxicidade.**

Palavras chave*—*Espectroscopia Raman, Reconhecimento de Padrões, Mistura Etanol-Metanol.

I. INTRODUÇÃO

O etanol, também conhecido como álcool etílico, compõe grande parte das bebidas alcoólicas. Na produção de cachaças, especialmente no caso de alambiques artesanais, o processo de fermentação é feito de maneira pouco controlada, podendo-se encontrar traços de metanol [1]. Essa contaminação pode ser decorrente do próprio processo de fermentação via presença não controlada de leveduras. As duas substâncias são incolores e de odor similar, mas a ingestão acidental de metanol pode causar acidose metabólica, cegueira e, conforme o volume ingerido em relação à massa corporal, levar a morte [2].

A análise comumente utilizada na indústria alcooleira para detecção de metanol é a cromatografia gasosa, porém, esse método é caro e consome tempo [3]. Assim este trabalho estuda métodos alternativos para a identificação da presença de metanol em misturas com etanol. Quando a radiação eletromagnética interage com as moléculas de etanol e metanol, um dos processos que pode resultar é o espalhamento Raman. Neste caso parte da energia da radiação incidente é espalhada em outros comprimentos de onda, produzindo um espectro característico que é função da estrutura molecular do analito. Como resultado, podem ser observadas bandas laterais ao comprimento de onda da radiação de excitação. Para o caso de líquidos, pouca ou nenhuma preparação da amostra é necessária.

Quanto maior a região espectral medida, maior a capacidade de diferenciação entre as substâncias. Este tipo de abordagem é referida na literatura como "*full spectral processing*", isto é, processamento pleno do espectro [4]. Uma das vantagens do uso do espalhamento Raman para esta identificação é o fato de que se pode obter uma região espectral de assinatura, correspondente aos modos de vibração molecular da substância, na região visível do espectro eletromagnético. Isso permite usar componentes óticos mais simples em relação a espectroscopia no infravermelho - região onde as transições moleculares de absorção e emissão de baixa energia aparecem. Como desvantagem deste método pode ser citada a baixa intensidade do sinal, ocasionando uma pobre relação sinal ruído. A Fig. 1 apresenta os espectros típicos de espalhamento Raman do etanol e do metanol. O espectro é dado em unidades arbitrárias, ou contagens (*counts*), do sistema de medição, em função da diferença de energia, em cm^{-1}, entre a radiação incidente e a radiação espalhada. Como o comprimento de onda, e consequentemente a energia, de excitação é fixo, o eixo das absissas é chamado de Deslocamento Raman (*Raman shift*).

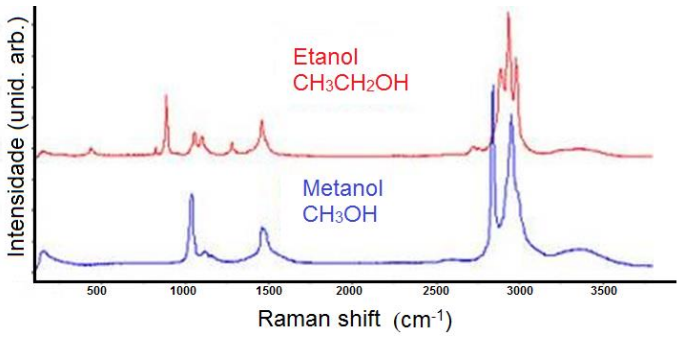

Fig. 1. Espectros Raman para o etanol e metanol. Adaptado de www.horiba.com .

A identificação da toxicidade utilizando a técnica ótica apresentada neste trabalho é realizada com base no reconhecimento de padrões no espectro de espalhamento Raman das amostras. Para isto são utilizadas duas técnicas conhecida, as Redes neurais artificiais e a Análise de componente principal, na montagem do classificador. O sistema a ser desenvolvido é simples e tem como objetivo a classificação de amostras em tóxicas e não tóxicas. A seguir, são apresentados os procedimentos adotados para medição e pré-processamento de sinal, bem como os resultados preliminares do classificador obtidos para uma mistura binária de etanol e metanol.

Este trabalho conta com apoio da CAPES, CNPq, FINEP e Fundação Araucária.

II. METODOLOGIA

Num sistema de reconhecimento de padrões típico, como mostrado na Fig. 2, várias etapas são executadas até o resultado da classificação [5].

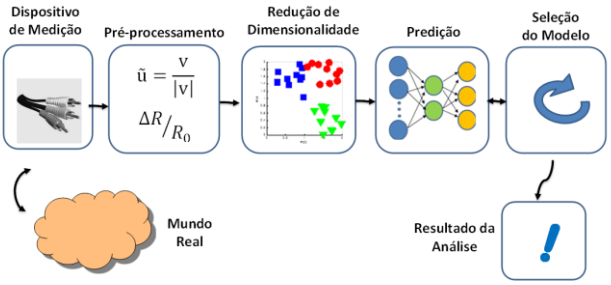

Fig. 2. Processo de reconhecimento de padrões. Adaptado de [5].

A seguir são apresentados, para o problema em questão, os detalhes de cada etapa realizada para se atingir o objetivo proposto. Para isso, foi utilizado o aplicativo *Matlab® r2008* e seus *toolboxes* de Bioinformática e Reconhecimento de Padrões da Mathworks.

A. *O Mundo Real: formação do espaço amostral*

Representando o "Mundo Real" da Fig. 2, são consideradas as amostras de líquido cujas propriedades óticas, no caso, o espectro de espalhamento Raman, serão medidas. Foram utilizadas 11 amostras com diferentes concentrações de metanol diluídas em etanol puro. Para o preparo dessas amostras, foram utilizadas duas buretas, a fim de aferir os volumes nas diferentes concentrações. O volume de metanol foi inserido no de etanol, conforme a concentração pretendida. As amostras preparadas são apresentadas na primeira coluna da Tabela 1.

B. *Dispositivo de Medição*

Na etapa seguinte, já sendo considerada parte do dispositivo de medição, estas amostras são colocadas numa cubeta de quartzo e excitadas com luz azul, proveniente de um Laser de Argônio (modelo Innova 70 da Coherent), sintonizado em 488 nm e potência de 300 mW. A Fig. 3 apresenta um diagrama do aparato experimental.

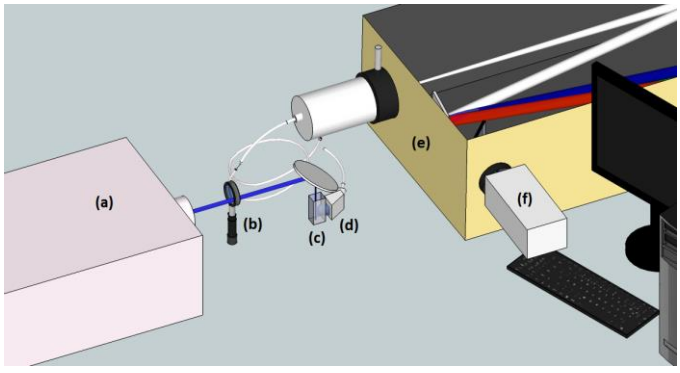

Fig. 3. Diagrama do aparato experimental no qual são evidenciados: o Laser de Argônio (a), a lente convergente (b), a cubeta (c), a lente GRIN (d), o espectrômetro(e) e o CCD (f).

A amostra é excitada pelo feixe de laser que passa por uma lente convergente e o espalhamento é coletado perpendicularmente à excitação através de uma lente de índice de refração gradual (GRIN) acoplada a uma fibra ótica (acoplador Horiba 220F) que leva a radiação até o espectrômetro (Horiba iHR550). Na fenda de entrada há um banco ótico para focalização. Este equipamento tem uma rede de difração de 1200 linhas/mm e a radiação dispersada internamente é medida por um sensor CCD (Horiba Synapse) cuja resolução do Conversar Analógico Digital (ADC) é de 16 bits. O sensor contém 1024 x 256 pixels e, para melhorar a relação sinal ruído, é refrigerado termoeletricamente a -75 ^{o}C.

Os sinais gerados pelo equipamento formam tabelas de intensidade de espalhamento em função do comprimento de onda. A configuração da fenda de entrada do equipamento resultou numa resolução de 0,05 nm, o que corresponde a aproximadamente 2 cm^{-1}. A região espectral escolhida resultou em arquivos de 1024 pontos por amostra com um espaçamento espectral de 0,035 nm entre cada par de pontos. Os espectros de espalhamento Raman medidos para o etanol e o metanol puros são apresentados na Fig. 4

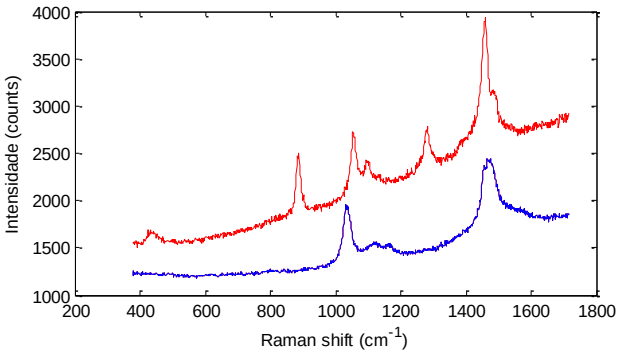

Fig. 4. Espectros medidos para o etanol (vermelho) e metanol (azul) puros. O eixo vertical é dado em contagens (counts) do ADC.

C. *Pré-processamento dos dados*

A primeira etapa de processamento consistiu em remover o sinal de fluorescência superposto ao espectro Raman das amostras. Esse sinal forma uma linha base e pode ser proveniente de componentes óticos do sistema, ou algum contaminante na amostra. Como a fluorescência apresenta uma banda larga em relação às bandas de espalhamento Raman, a técnica escolhida foi a aplicação da função *msbackadj* do toolbox de bioinformática do *Matlab®*. Essa função faz uma correção da linha base de um sinal com picos estreitos. Em seguida, para que cada vetor de treinamento tenha um peso igual, todas as amostras foram normalizadas de modo que a área sob o espectro, que corresponde a energia de toda radiação espalhada na região espectral de interesse, com intensidade I em função da frequência ν, fosse igualada a 1, conforme (1).

$$I(\nu) = \frac{I(\nu)}{\sqrt{\sum[I(\nu)^2]}} \qquad (1)$$

Posteriormente foi aplicado um filtro convolucional Savitzki-Golay. Neste caso, considerando que as características de interesse do espectro devem ser mantidas com poucas distorções, isto é, com remoção de ruído porém sem o alargamento ou atenuação dos picos [4], o melhor resultado foi obtido com um filtro de 7ª. ordem e uma janela de 101 valores. Os resultados após a remoção da linha base, normalização e aplicação do filtro convolucional são apresentados na Fig. 5. Um dos efeitos que podem ser notados é que a aplicação deste tipo de filtro é análoga ao aumento da fenda de entrada do espectrômetro e resulta em um espectro com menor resolução.

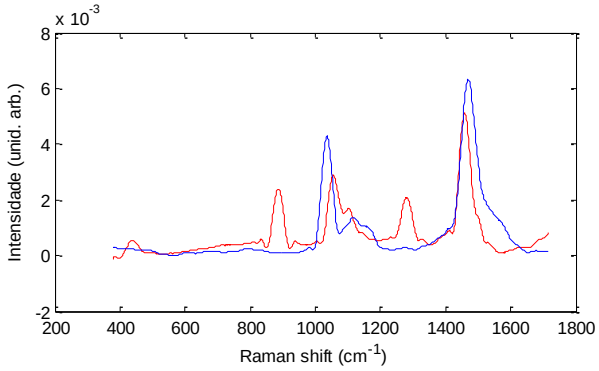

Fig. 5. Espectros do etanol (vermelho) e do metanol (azul) puros após a etapa de pré-processamento. Nesta etapa o eixo vertical é dado em unidades arbitrárias.

D. Predição e Redução de dimensionalidade

Caso toda a janela espectral apresentada na Fig. 1 fosse considerada, o espectro teria 3052 dimensões. Entretanto, dado que um espaço de características com 1024 dimensões ainda é alto para a avaliação proposta no presente artigo, foram adotadas duas estratégias de reconhecimento de padrão separadamente de modo a construir um modelo capaz de realizar a classificação das amostras.

A primeira técnica desenvolvida é baseada em Redes Neurais Artificiais (RNA), a qual é recomendada para avaliação de sistemas de grandes dimensões. Para isto foi treinada uma RNA de 1024 entradas para realizar a classificação. A segunda técnica desenvolvida foi a realização de uma Análise de Componente Principal (PCA) para reduzir a dimensionalidade do espaço de características de entrada para depois, com os vetores de treinamento projetados nesse espaço reduzido, determinar o melhor método de separação. As duas técnicas, conforme reportado na literatura, podem ainda ser combinadas para solução de problemas complexos, já que o desempenho da RNA é melhorado quando treinada com uma representação mínima do espaço de características (resultado da PCA).[6]

Devido ao fato de as medidas formarem um espaço amostral limitado a 11 amostras distintas, optou-se por sintetizar matematicamente diversas concentrações para uso na etapa de predição e redução de dimensionalidade e usar as medidas reais apenas na etapa de teste. O espectro do etanol puro adicionado a um ruído aleatório foi misturado em diferentes proporções ao do metanol puro. Este procedimento é válido, uma vez que as duas substâncias empregadas não reagem entre si.

Foram geradas 202 amostras. Cada amostra é um vetor de 1024 características. Como o treinamento é supervisionado para o caso da RNA, para cada amostra foi assinalado um vetor alvo cujo critério é "**perigoso**" (P), atribuindo-se o valor 1 à saída e "**não perigoso**" (não P), atribuindo-se o valor 0 à saída. O valor alvo também permite a visualização da separabilidade no PCA. O limiar de toxicidade foi definido arbitrariamente em 0,15 v/v. A Fig. 6 mostra a sobreposição dos espectros que compõem o conjunto de amostras usadas para o treinamento da RNA e para a execução do PCA. As áreas hachuradas representam regiões espectrais onde há variação das intensidades em função das concentrações.

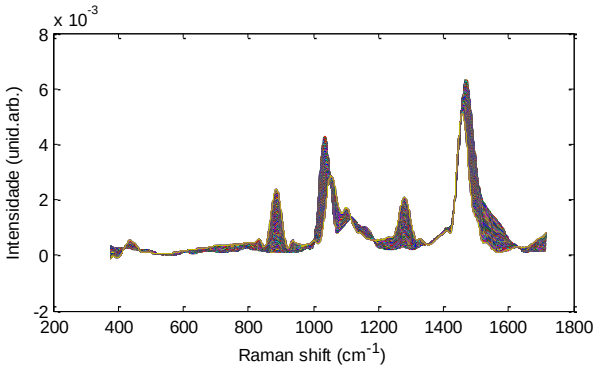

Fig. 6. Espaço de caracteristicas usado para o treinamento da RNA

1) Treinamento da Rede Neural Artificial do tipo MLP

Para a estrutura de RNA apresentada na Fig. 7 foi realizado o treinamento com 10, 100 e 300 neurônios na camada escondida. Os algoritmos utilizados foram padrão do *Matlab*®. A divisão dos dados entre treinamento, validação e teste foi feita de modo aleatório. Para o algoritmo de treinamento foi usado o "*Scaled Conjugated Gradient*" com avaliação de desempenho baseada no Erro Quadrático Médio.

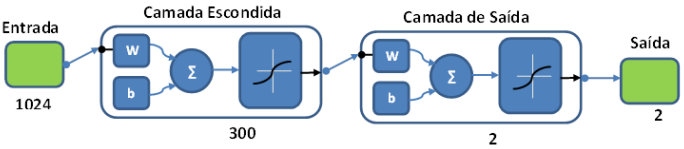

Fig. 7. Diagrama da RNA treinada. Foram feitos testes com 10, 100 e 300 neurônios na camada escondida. [Adaptado de Mathworks]

2) Análise de Componente Principal

Num segundo momento foi aplicada a Análise de Componente Principal (PCA) às amostras pré-processadas. Inicialmente, para que as dimensões tenham o mesmo peso na análise, cada uma das 1024 dimensões da matriz A foi subtraída de sua média e dividida pelo seu desvio padrão nas 202 amostras, conforme (2).

$$B_i = \frac{A_i - \bar{A}}{\sigma_A} \qquad (2)$$

A partir destes vetores padronizados, a matriz B de dimensão 1024x202, foi realizada a análise de componente principal. Para isso foi usada a função *princomp* do toolbox de estatística do *Matlab*®

Essa análise resultou em uma matriz de rotação C de 1024x1024 dimensões. Os novos vetores projetados, a matriz P de 202x1024 dimensões, é obtida multiplicando-se os vetores de entrada B pela matriz C conforme (3). As componentes principais são um subconjunto deste resultado.

$$P = B^T \times C \tag{3}$$

III. Resultados e Discussão

Para verificar a capacidade de generalização da RNA os vetores de características das 11 amostras reais, conforme a Fig. 8, foram apresentados na entrada. Os resultados da classificação, na forma de um valor numérico entre 0 e 1 para cada uma das saídas, são sumarizados na Tabela 1. São apresentados os resultados obtidos com a RNA contendo diferente número de neurônios na camada escondida. As células marcadas em laranja indicam erros de classificação.

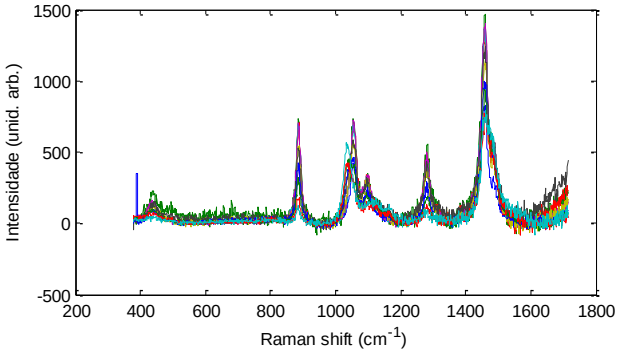

Fig. 8. Espectros reais com a remoção da linha base

TABELA 1. Resultados da Classificação pela RNA

Concentr. Metanol (v/v)	Resultados na saída da RNA-MLP em função do número de neurônios na camada escondida					
	10 neurônios		*100 neurônios*		*300 neurônios*	
	P	***não P***	***P***	***não P***	***P***	***não P***
0,01	0,00	0,92	0,83	0,07	0,46	0,84
0,02	0,00	0,91	0,00	1,00	0,18	1,00
0,03	0,00	0,98	0,02	0,92	0,10	0,99
0,04	0,00	0,93	0,06	0,59	0,06	1,00
0,05	0,00	0,97	0,11	0,42	0,11	0,99
0,10	0,00	0,98	0,00	0,91	0,28	0,97
0,25	0,00	0,98	0,00	1,00	0,72	1,00
0,40	0,00	0,95	0,94	0,02	0,80	0,45
0,55	0,00	0,94	0,99	0,00	0,98	0,07
0,70	0,09	0,82	1,00	0,00	0,99	0,00
0,85	0,76	0,38	1,00	0,00	1,00	0,00

A Fig. 9 apresenta um gráfico das duas primeiras componentes principais. As amostras perigosas são representadas como quadrados azuis e as não perigosas como triângulos verdes. Apesar do ruído incluído deliberadamente na síntese matemática das amostras, grande parte da informação está contida na primeira componente principal e o discriminador pode ser apenas uma reta da componente principal PC1 e cujo valor de separação é -40.

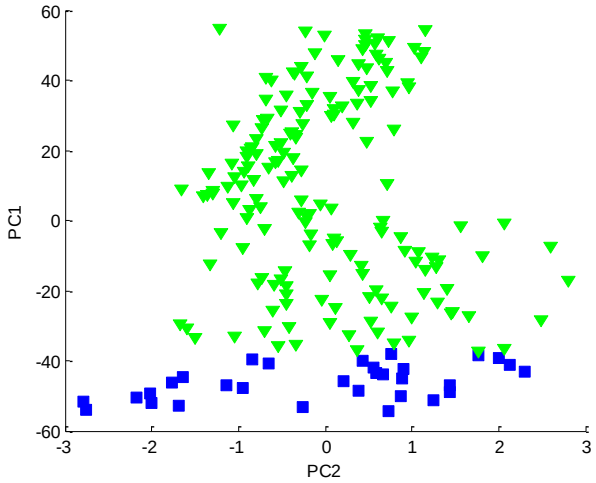

Fig. 9. Componentes principais PC1 e PC2. Amostras perigosas (quadrados azuis) e amostras não perigosas (triângulos verdes).

Para verificar a capacidade de separação da rotação aplicada, os vetores de características das amostras reais mostrados na Fig. 8 passaram pela etapa de pré-processamento, padronização e foram multiplicadas por esta matriz. O resultado é apresentado na Fig. 10.

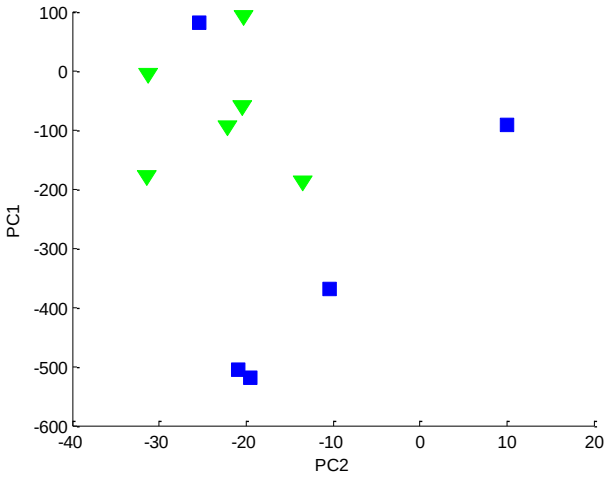

Fig. 10. Espaço de características das amostras reais rotacionado para a maior variância. . Amostras perigosas (quadrados azuis) e amostras não perigosas (triângulos verdes).

Para a análise inicial utilizando a RNA a Tabela 1 indica que a quantidade de neurônios na camada escondida deve ser grande o suficiente para capturar a informação contida nos

espectros. Para um número pequeno, ocorrem erros de classificação. Para o caso de 100 neurônios, apesar de os valores serem próximos aos de 300 neurônios, aparece certa indecisão na saída com valores próximos para a indicação de "**perigoso**" e "**não perigoso**". Os valores notadamente incorretos são destacados na cor laranja. Percebe-se, portanto, que caso não seja criada uma região de indecisão mais ampla, algumas concentrações são classificadas de maneira incorreta.

Para o caso do PCA, pela representação gráfica das duas componentes principais percebe-se que as amostras classificadas como perigosas são facilmente separáveis das amostras não perigosas somente pela primeira componente principal PC1.

Um aspecto importante a ser ressaltado é que espectros medidos experimentalmente devem, assim como os vetores usados na entrada do PCA, ser padronizados pela média e desvio padrão dos vetores utilizados para calcular a matriz de rotação antes de multiplicá-los pela matriz de coeficientes obtida pelo PCA.

Para ambos os casos, um dos vetores de teste não foi classificado corretamente. Examinando com mais cuidado a origem deste ponto, percebe-se que ele vem de uma das medidas com superposição significativa de fluorescência ao espalhamento Raman. Os espectros são apresentados na Fig. 11 e a origem mais provável da fluorescência é a contaminação da amostra.

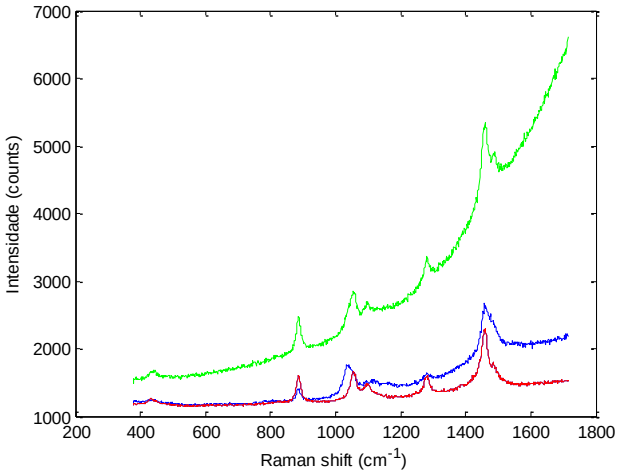

Fig. 11. Espectros de teste antes do pré-processamento. A medida em verde, mais distante da linha base é a amostra com concentração de 0,25 v/v.

IV. CONCLUSÕES

No contexto da classificação da toxicidade da amostra etanol-metanol, dada esta análise inicial, que corresponde ao processamento de espectros de espalhamento Raman, conclui-se que, para um treinamento supervisionado de uma RNA do tipo MLP, é necessária uma quantidade compatível de neurônios na camada escondida para capturar a informação contida no espectro. O uso do PCA permitiu, por meio da obtenção de uma matriz de rotação capaz de maximizar a variância, averiguar que as amostras representam de fato apenas uma combinação linear das substâncias, isto é, uma mistura binária sem reação.

Conclui-se, também, que a determinação de um espaço amostral representativo é crucial. Como representativo, deve-se ter um número de amostras equilibrado entre "**perigoso**" (P) e "**não perigoso**" (não P). A capacidade de generalização depende, também, da inclusão de erros sistemáticos no espaço amostral tais como a fluorescência da amostra e o ruído impulsivo gerado por raios cósmicos atingindo o CCD. Em misturas com mais de dois componentes significativos, e com reações entre si, a separação pode ser não linear e a escolha de uma RNA combinada com PCA ou um discriminante mais elaborado como LDA ou Máquinas de Vetor Suporte pode ser empregada.

A maneira mais eficiente de diminuir o efeito da fluorescência seria o aumento do comprimento de onda do laser de excitação. A disponibilidade de lasers de semicondutor bem como redes de difração e CCDs de baixo custo e tamanho reduzido pode fazer com que o princípio apresentado no presente artigo possa ser empregado para a construção de um medidor portátil para a avaliação de toxicidade de cachaças artesanais [7].

O trabalho descrito no presente trabalho propõe-se a direcionar a implementação do reconhecimento de padrões para misturas reais. Os próximos passos a serem realizados compreendem a análise de ambiente químico mais complexo como ocorre nas cachaças. Neste caso, a redução de dimensionalidade com o PCA e posterior treinamento de uma RNA tende a ser a abordagem mais promissora.

REFERÊNCIAS

[1] A. J. Paine, and A. D. Dayan, "Defining a tolerable concentration of methanol in alcoholic drinks", Human & Experimental Toxicology, vol. 20, pp. 563-568, 2001.

[2] A. Vale, "Alcohols and Glycols", Specific Substances, Medicine 40:2,Elsevier, pp. 89-93, 2011.

[3] L. M. Reid, C. P. O´Donnell, G. Downey, "Recent technological advances for the determination of food authenticity", Trends in Food & Science Technology, vol 17, pp.344-353, 2006.

[4] J. Ferraro, "Introductory Raman Spectroscopy, Second Edition". Academic Press, 2002.

[5] R. O. Duda, P. Hart, D. Stork, "Pattern Classification". Second Edition Wiley, 2001.

[6] J. M. Saleh, and B. S. Hoyle, "Improved Neural Network Performance Using Principal Component Analysis on Matlab", Int. Journal of the Computer, the Internet and Management, vol. 16:2 pp1-8, 2008.

[7] M.A. Young et al. "Surface-enhanced Raman spectroscopy with a laser pointer lightsource and miniature spectrometer". Canadian Journal of Chemistry, vol. 82 pp 1435-1441, 2004.