UNIVERSIDADE FEDERAL DA BAHIA

INSTITUTO DE CIÊNCIAS DA SAÚDE

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

ELDERLEI DE JESUS PITA SOUZA

APLICAÇÃO DA ESPECTROSCOPIA RAMAN E ESTATÍSTICA MULTIVARIADA NO ESTUDO QUANTITATIVO DE MOLÉCULAS DE INTERESSE BIOTECNOLÓGICO

Salvador
2017

ELDERLEI DE JESUS PITA SOUZA

APLICAÇÃO DA ESPECTROSCOPIA RAMAN E ESTATÍSTICA MULTIVARIADA NO ESTUDO QUANTITATIVO DE MOLÉCULAS DE INTERESSE BIOTECNOLÓGICO

Dissertação apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade Federal da Bahia, como requisito para obtenção do grau de Mestre em Biotecnologia.

Orientador: Prof. Dr. Elias Ramos de Souza
Co-orientador: Prof. Dr. Paulo Fernando de Almeida

Salvador
2017

Souza, Elderlei de Jesus Pita
Aplicação da Espectroscopia Raman e da Estatística Multivariada no estudo quantitativo de moléculas de interesse biotecnológico / Elderlei de Jesus Pita Souza. -- Salvador, 2017.
68 f. : il

Orientador: Elias Ramos de Souza.
Coorientador: Paulo Fernando de Almeida.
Dissertação (Mestrado - Programa de Pós-graduação em Biotecnologia) -- Universidade Federal da Bahia, Instituto de Ciências da Saúde, 2017.

1. Espectroscopia Raman. 2. Estatística Multivariada. 3. PCA. 4. Regressão. 5. Soluções aquosas. I. Souza, Elias Ramos de. II. Almeida, Paulo Fernando de. III. Título.

CDU 606

ELDERLEI DE JESUS PITA SOUZA

Aplicação da Espectroscopia Raman e Estatística Multivariada no Estudo Quantitativo de Moléculas de Interesse Biotecnológico

Dissertação apresentada como requisito para obtenção do grau de Mestre em Biotecnologia pelo Instituto de Ciências da Saúde da Universidade Federal da Bahia.

Aprovada em 08 de setembro de 2017.

BANCA EXAMINADORA:

Dr. Elias Ramos de Souza – Orientador _____
Doutor em Biofísica pela Universidade Federal do Rio de Janeiro, UFRJ, Brasil.
Instituto Federal de Educação, Ciência e Tecnologia da Bahia.

Dr. Paulo Fernando de Almeida _____
Doutor em Engenharia de Alimentos pela Universidade Estadual de Campinas, UNICAMP, Brasil.
Universidade Federal da Bahia.

Dr. José Mário Araújo _____
Doutor em Engenharia Elétrica pela Universidade Federal da Bahia, UFBA, Brasil.
Instituto Federal de Educação, Ciência e Tecnologia da Bahia.

Dedico este trabalho a Deus, à minha família e a todos aqueles que acreditaram em mim.

AGRADECIMENTOS

Primeiramente, minha eterna gratidão a Deus, aquele que me deu o fôlego de vida e me permitiu chegar até aqui. Muitos foram os momentos em que não acreditei que conseguiria concluir este trabalho. Mas Ele, a cada instante, me conduziu e me deu palavras de estímulo, colocando pessoas ao longo de minha caminhada.

Agradeço aos meus pais, Neto e Mirian, por terem sido usados por Deus para me educarem e me preparar para o mundo; por toda preocupação com meu bem-estar, pela compreensão e todo sacrifício realizado até hoje, para que eu me tornasse o que sou.

Aos meus irmãos, Elderlan e Ebe, pelo apoio e estímulo.

À Juliana, minha esposa, a qual me foi presenteada por Deus e desempenhou papel fundamental na reta final deste trabalho.

Aos meus orientadores, Professor Dr. Elias Ramos e Professor Dr. Paulo Almeida, por terem ido além da orientação acadêmica, compreendendo e colaborando na medida do possível para que eu conseguisse desenvolver este trabalho e formar-me como indivíduo.

Ao Professor Dr. Denis Gilbert, pela confiança que me foi depositada, ao liberar o meu acesso irrestrito ao espectrômetro Raman do LFNA. Este trabalho não teria êxito sem esta atitude nobre.

Aos colegas do LABEM, pelo tempo e experiências compartilhados.

Aos colegas do Instituto de Física, com quem tive diversos momentos de discussão, aprendizado e descontração ao longo desta trajetória.

Aos meus irmãos em Cristo, os quais sempre me estimularam a continuar e compreenderam minhas ausências nas atividades.

Aos meus colegas de trabalho, na PETROBRÁS/FAFEN-BA, pelas diversas trocas de turno realizadas, permitindo que eu conseguisse realizar experimentos e testes, assistir aulas e conciliar todos os outros compromissos.

Aos Professores: Dr. Eduardo do Nascimento, pelas importantes conversas sobre Estatística; Dr. Eduardo Simas, pela ajuda em Processamento de Sinais; Dr. José Mario, por ter aceitado o desafio de participar de última hora da minha banca.

Ao Programa de Pós-graduação em Biotecnologia, pela oportunidade de concluir a pesquisa.

A tantos outros profissionais que passaram por mim durante o período deste trabalho, sem os quais não seria possível chegar ao final desta jornada: os motoristas da frota de turno, que por tantas vezes me ajudaram nas trocas de turno e nos dias em que saí direto da Universidade; aos porteiros e vigilantes do Instituto de Física e do Instituto de Ciências da Saúde, com os quais por inúmeras vezes compartilhei o eco dos corredores nos finais de semana, feriados e madrugadas, nos experimentos que pareciam intermináveis.

Enfim: nenhum homem é uma ilha. Não se faz ciência sozinho. A finalização deste trabalho é resultado de parcerias, em todos os níveis.

SOUZA, Elderlei de Jesus Pita. Aplicação da espectroscopia Raman e da Estatística Multivariada no estudo quantitativo de moléculas de interesse biotecnológico. 68 f. il. 2017. Dissertação (Mestrado) – Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, 2017.

RESUMO

A análise, monitoramento e detecção de substâncias presentes em meios de produção, assim como a determinação de suas concentrações são fatores de grande importância em processos biotecnológicos, permitindo, por exemplo, estudo de otimização de processos metabólicos de microorganismos. Contudo, o nível de complexidade e a grande incerteza, associados aos resultados de alguns métodos, limitam seu uso e reduzem o grau de confiabilidade dos mesmos. Neste cenário, a espectroscopia de espalhamento Raman com base em suas diversas vantagens como a capacidade de obtenção de espectros de amostras em qualquer estado físico e condição de temperatura e pressão, associado à ideia de “impressão digital” espectral das substâncias, apresenta-se como proposta de técnica para as demandas mencionadas. No entanto, devido à sua natureza de técnica semi-quantitativa, requer ferramentas matemáticas adequadas para o correto tratamento e interpretação de seus dados. O uso de técnicas estatísticas multivariadas, como a Análise de Componentes Principais (PCA) e a Regressão Linear Multivariada (MLR) permitem o uso dos dados espectrais na sua totalidade, obtendo-se o máximo de informações neles contidas. O presente trabalho aplica estes métodos a dados oriundos de espectros Raman obtidos de diversas soluções aquosas de nitrato de sódio, glicerol e raminose (substâncias de interesse biotecnológico), em diferentes concentrações¸ relacionando as amplitudes de cada um destes espectros às suas proporções presentes nas misturas. Assim, foram criados modelos de regressão para a calibração destes dados, utilizando as intensidades espectrais como preditores e as respectivas concentrações como respostas, sendo realizados testes de predição e validação destes mesmos modelos. Também foi realizado o pré-processamento matemático destes dados através do PCA, identificando as variáveis de maior relevância e filtrando parte do ruído presente nos espectros. Foram também realizadas avaliações qualitativas dos mesmos espectros, discutindo-se suas principais características. A análise dos resultados obtidos confirmou a capacidade do método em identificar a presença das substâncias em questão nas misturas testadas, além de determinar suas respectivas concentrações através de seus espectros Raman. A Análise de Componentes Principais também mostrou-se eficiente no tratamento dos dados, possibilitando, inclusive, a identificação de padrões espectrais entre as amostras, nem sempre perceptíveis sem o adequado tratamento matemático.

Palavras-chave: Espectroscopia Raman, Estatística Multivariada, PCA, Regressão, Soluções Aquosas.

SOUZA, Elderlei de Jesus Pita. Application of Raman Spectroscopy and Multivariate Statistics at quantitative study of molecules of biotechnological interesting. 68 f. il. 2017. Master Thesis – Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, 2017.

ABSTRACT

The analysis, monitoring and detection of substances in production media, as well as the determination of its concentrations are factors of major at biotechnological processes, allowing, for example, optimization studies of metabolic processes of microorganisms. However, the complexity level and the large inherent uncertainty to some methods limit their use and reduce their reliabilty level. In this scenario, Raman scattering spectroscopy based on its several advantages such as the ability to obtain spectra of samples in any physical state and temperature and pressure conditions, associated to the idea of spectral “fingerprint” of the substances, is presented as a proposal of for the mentioned demands. However, because its semi-quantitative technique characteristic, it requires suitable mathematical tools for the correct treatment and interpretation of its data. The use of Multivariate Statistical techniques, such as Principal Component Analysis (PCA) and Multivariate Linear Regression (MLR) allow the use of the spectral data in their entirety, reaching the major the information contained in them. The present work applies these methods to data from Raman spectra obtained from various aqueous solutions of sodium nitrate, glycerol and rhaminosis (substances of biotechnological interest), in different concentrations¸ relating the intensities of each of these spectra to their proportions in the mixtures. Thus, regression models were created for the calibration of these data, using the spectral intensities as predictors and the respective concentrations as responses, being carried out prediction and validation tests on these same models. It was also performed the mathematical preprocessing of these data with use of the PCA, identifying the variables of greater relevance and filtering part of the existing noise in the spectra. Qualitative evaluations of the same spectra were also performed, discussing its main characteristics. The analysis of the results confirmed the ability of the method at detection of the mentioned substances in the tested mixtures, in addition to determining their respective concentrations through their Raman spectra. The Principal Component Analysis also proved to be efficient in data processing, allowing even the identification of spectral patterns among the samples, which are not always perceptible without suitable mathematical treatment.

Keywords: Raman Spectroscopy, Multivariate Statistics, PCA, Regression, Aqueous Solutions.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

cm^{-1}	Centímetros recíprocos
MLR	Multivariate linear regression
MSM	Meio Salino Mineral
nm	Nanômetros
PC	Componente Principal
PCA	Principal Component Analysis
s	segundos

SUMÁRIO

1 INTRODUÇÃO

Em processos biotecnológicos, a quantificação de substâncias é de suma importância para o controle, compreensão e otimização de resultados dos sistemas em estudo ou em regime de produção das mesmas. Contudo, muitas vezes a mensuração ou mesmo a identificação da presença ou não destas substâncias e compostos são difíceis, ou incertas, ou mesmo exigem um grau de especificidade dos equipamentos envolvidos ao ponto de o custo para tal inviabilizar o acesso à informação, como o caso da análise cromatográfica. Então, recorre-se a métodos indiretos de avaliação, como a utilização do Índice de Emulsificação (E24), para se avaliar a presença ou atuação de surfactantes — que pode ser falho, devido a emulsificação também poder ocorrer com outras substâncias presentes no sistema, além do erro de paralaxe do avaliador (CENTENO DA ROSA et al., 2010). Ou, muitas vezes o procedimento para quantificação é composto por tantas etapas ao ponto que a propagação de erros torna-se inevitável, colocando em xeque a confiabilidade dos resultados, além de critérios subjetivos de medição, como citado acima, impactando mais na possibilidade de erro. Exemplos disso são os procedimentos de determinação da concentração de Raminolipídeos num meio de produção, como descritos por Wang e colaboradores (2007) e Patel e Desai (1997), onde são executadas 8 e 6 etapas, respectivamente, para se chegar a um resultado. Situações como estas tornam-se críticas e determinantes em condições onde a sensibilidade dos resultados possa sofrer interferência. Substâncias como o nitrato de sódio, o glicerol – amplamente utilizados como fontes metabólicas de microorganismos – e a raminose – uma parte de uma substância produzida por alguns destes microorganismos –, são amplamente utilizadas em pesquisas as quais dependem de métodos quantitativos e de identificação como os citados acima. Surge, assim, a necessidade da validação de outros métodos com confiabilidade mínima para estas aplicações.

1.1 ESPECTROSCOPIA RAMAN E ESTATÍSTICA MULTIVARIADA

O efeito Raman corresponde ao espalhamento inelástico da radiação eletromagnética monocromática ao interagir com as moléculas (SALA, 2008). A Espectroscopia de espalhamento Raman, ou simplesmente Espectroscopia Raman, é o estudo deste fenômeno, evidenciando os níveis de energia existentes na molécula que promove o espalhamento. O espectro Raman de uma determinada substância é considerado como uma "impressão digital" da mesma (*fingerprint*), uma vez que cada elemento químico apresenta um único padrão de sinal Raman, ou espectro. Por este motivo, é cada vez maior o número de publicações sobre o espectro Raman das mais diversas substâncias (BASCHENKO; MARCHENKO, 2011).

A técnica tem sido amplamente utilizada no campo de caracterização de materiais, devido a diversas vantagens, dentre elas: é uma técnica não-destrutiva, de aplicação relativamente rápida, não requer pré-tratamento da amostra a ser analisada, é aplicável às amostras numa grande gama de estados físicos (sólidos, líquidos, gases e vapores, sejam a altas ou baixas temperaturas) além de não exigir grandes quantidades das amostras. Ainda, a intensidade do sinal é proporcional à intensidade da radiação e à intensidade do fenômeno promovido pela molécula. Por estes motivos, é possível identificar a presença de substâncias e micro-organismos, assim como alterações moleculares e ainda caracterizá-las (ERIX, 2011; RYDER; CONNOR; GLYNN, 2000; SMITH; DENT, 2005).

Devido a esta grande versatilidade, a espectroscopia Raman apresenta-se como uma ótima alternativa a todas as técnicas atualmente conhecidas e utilizadas na determinação e caracterização de substâncias, uma vez que muitas das técnicas atualmente utilizadas exigem alguma forma de pré-tratamento ou preparo da amostra, onerando significativo tempo, em algumas delas, além do custo de reagentes para tal finalidade e da grande especificidade de alguns equipamentos, como na espectrofotometria e na cromatografia (ABDEL-MAWGOUD et al., 2011). Contudo, a espectroscopia Raman é uma técnica semi-quantitativa, sendo

necessário a utilização de métodos numéricos e estatísticos para a obtenção de interpretação matemática.

Nesta realidade surge o papel da **Estatística Multivariada**, a qual permite a manipulação, ajuste e interpretação de grandes conjuntos de dados (informações) em condições onde os sistemas ou modelos são simultaneamente descritos por (ou relacionam-se simultaneamente com) mais de uma variável. Muitas informações contidas em dados numéricos de descrição de fenômenos e processos só se tornam evidentes ou interpretáveis mediante o adequado tratamento e processamento de múltiplas variáveis. A identificação das variáveis mais relevantes para um sistema; a remoção ou atenuação de sinais ruidosos; a avaliação da confiabilidade de um resultado de um experimento – todas estas situações são comuns e exigem alguma atenção da parte daqueles que manipulam dados numéricos. Inge Koch, em sua obra "*Analysis of Multivariate and High-dimensional data*" (2013), enfatiza que conjuntos de dados muito grandes (em termos de dimensões e quantidade de amostras), a estrutura essencial pode vir a ser mascarada pelo ruído, sendo importante a redução dos dados originais, de modo a viabilizar a preservação da informação principal e a remoção de dados ruidosos, aleatórios ou de pouca relevância, os quais poderiam comprometer a análise. Dentre as diversas técnicas matemáticas utilizadas na Estatística, pode-se destacar a **Análise de Componentes Principais**, ou PCA (do inglês *Principal Component Analysis*) – técnica esta que torna possível, dentre outros pontos, a redução da dimensão de grandes conjuntos de dados, inclusive dados espectroscópicos. Uma outra técnica bastante conhecida e utilizada é a Regressão Multivariada, a qual permite parametrizar dados, de maneira a se determinar ou estimar resultados em função de seus preditores.

Assim, este trabalho apresenta a aplicação da espectroscopia Raman, associada à estatística multivariada, como proposta de método de avaliação e quantificação de substâncias em processos produtivos em biotecnologia, o que permitiria identificá-las em meios de produção e determinar suas respectivas proporções, além de acompanhar o consumo das mesmas e a produção de metabólitos por parte dos microorganismos envolvidos, e assim tornar-se uma alternativa aos métodos de quantificação comumente utilizados.

Este trabalho apresenta-se organizado da seguinte forma: no Capítulo 2, há uma breve explanação teórica sobre o campo da espectroscopia, o efeito Raman e a espectroscopia Raman. Também se discute os principais conceitos da Estatística Multivariada e algumas de suas técnicas, como a Análise de Componentes Principais e a Análise de Regressão. No Capítulo 3, descrevem-se todos os procedimentos realizados e equipamentos utilizados para este trabalho, garantindo a sua reprodutibilidade. Em seguida, no Capítulo 4, são discutidos os resultados dos testes executados sobre os dados matemáticos dos espectros, tanto em nível qualitativo, quanto em nível quantitativo, apresentando-se algumas das conclusões obtidas, com eventuais aprofundamentos teóricos sobre alguns tópicos que venham a se justificar. Finalizando, nos Capítulos 5 e 6, tem-se as considerações finais sobre a pesquisa realizada, e as obras utilizadas como referência para o desenvolvimento de todo o trabalho.

2 REVISÃO BIBLIOGRÁFICA

2.1 EFEITO RAMAN, ESPECTROSCOPIA E ESPECTROSCOPIA RAMAN

A luz, ao interagir com a matéria, pode ser espalhada, absorvida ou mesmo pode atravessar a matéria, não interagindo com a mesma. No caso da interação, tal estudo é atribuído ao campo da espectroscopia, sendo tanto no campo da Teoria Eletromagnética Clássica, tratando a luz como radiação eletromagnética, quanto no campo quântico, segundo a teoria corpuscular da luz.

Sabe-se que a energia de um fóton é proporcional à sua frequência, segundo a lei de Planck. De acordo com a teoria da dispersão da luz de Kramer-Heisenberg, um fóton de frequência *v*, ao interagir com uma molécula de uma determinada substância, “salta” de um estado inicial de energia para um estado excitado, com nível de energia superior, chamado de “estado virtual”. Este estado virtual é causado justamente pela interação do feixe de laser com as moléculas, causando a sua polarização. Por este motivo, não é um estado real e instável, para o fóton e para a molécula, além de sua curta duração (da ordem de 10^{-17} s). Rapidamente, a molécula retorna para um de seus estados iniciais, emitindo novamente o fóton, podendo este apresentar o mesmo nível de energia inicial, ou níveis diferentes. As três possíveis situações são (RANDHAWA, 2003):

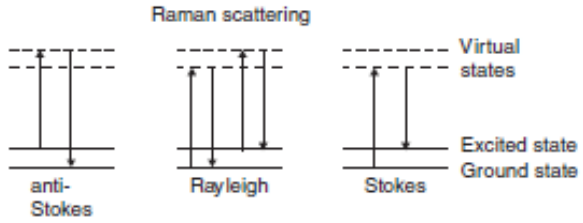
a) O fóton liberado possui a mesma frequência do fóton absorvido. Tal fenômeno de emissão é consequência da interação do primeiro fóton com o núcleo da molécula, ocasionando um choque elástico, sendo denominado espalhamento Rayleigh.

b) O fóton liberado apresenta-se com frequência menor que a frequência do fóton absorvido (i.e., a energia no fóton liberado é menor), devido à interação do fóton incidente com a nuvem eletrônica da molécula, ocorrendo um espalhamento inelástico com absorção de parte da energia inicial deste fóton. Denomina-se espalhamento Raman Stokes.

c) O fóton liberado pela molécula que retorna para um de seus estados iniciais, apresenta uma frequência maior do que o fóton absorvido e, consequentemente, uma energia maior. É causado também pelo choque inelástico do fóton com a nuvem de elétrons, sendo que neste caso a molécula cede mais energia para o fóton reemitido. Tal fenômeno denomina-se espalhamento Raman Anti-Stokes.

Um diagrama explicitando a diferença entre os três espalhamentos pode ser visto na Figura 11. O espalhamento Raman é relativamente fraco, em torno de 1 em 10^5-10^8 dos fótons que são espalhados. Este fenômeno foi previsto por Smekal em 1923 e observado experimentalmente em 1928 por Sir Chandrasekhra Venkata Raman e K. S. Krishnan. Ao contrário do espalhamento Rayleigh, o efeito ou espalhamento Raman não apresenta mudança na côr da luz espalhada, além de ser possível obter o fenômeno através de uma radiação monocromática.

Figura 1- Representação gráfica dos espalhamentos Rayleigh e Raman, com os respectivos estados energéticos



Fonte: (JACOBSSON; JOHANSSON, 2009)

A espectroscopia Raman corresponde ao estudo deste efeito, através da medida da intensidade desta radiação espalhada pelas moléculas das mais diversas substâncias, como função da frequência ou comprimento de onda ou número de onda.

A espectroscopia, no geral, estuda a interação da radiação eletromagnética com a matéria, sendo um dos seus principais objetivos a determinação dos níveis de energia de átomos ou moléculas (SALA, 2008). Os espectros fornecem as transições, que são as diferenças de energia entre os níveis. A região espectral onde

estas são observadas depende do tipo de nível envolvido, podendo ser eletrônico, vibracional ou rotacional.

A interação da radiação eletromagnética com o movimento vibracional dos núcleos pode originar o espectro vibracional no infravermelho ou o espalhamento Raman. No caso do espalhamento Raman, ocorre o espalhamento inelástico da radiação eletromagnética monocromática que interage com as moléculas (JACOBSSON; JOHANSSON, 2009; SALA, 2008).

A espectroscopia Raman é conhecida como uma “técnica de impressão digital”, uma vez que todos os elementos e substâncias químicas fornecem um único padrão de sinal Raman, ou espectro. Por este motivo, é cada vez maior o número de publicações sobre o espectro Raman das mais diversas substâncias (BASCHENKO; MARCHENKO, 2011).

A técnica tem sido amplamente utilizada no campo da caracterização de materiais, devido a diversas vantagens, dentre elas: é uma técnica não-destrutiva; de aplicação relativamente rápida; não requer pré-tratamento da amostra a ser analisada; permite a implementação sobre amostras em pequenas quantidades, em virtude do diâmetro do laser utilizado; permite o estudo em substâncias em meio aquoso, pois a água possui atividade Raman fraca, havendo pouca interferência das vibrações de suas moléculas sobre os espectros principais; é possível obter o espectro das amostras mesmo através do vidro de tubos de ensaio e outros frascos fechados, sendo uma vantagem para substâncias higroscópicas e altamente reativas em contato com o ar (FERRARO; NAKAMOTO; BROWN, 2003; SMITH; DENT, 2005).

Assim, com o uso da espectroscopia Raman, é possível identificar a presença de contaminantes em misturas, como utilizado em estudos forenses (MARCELO et al., 2015; RYDER; CONNOR; GLYNN, 2000). Na área biomédica, como outro exemplo de aplicação, tem sido utilizada para identificação de microorganismos, caracterização de alterações biomoleculares e estudos de processos de adsorção (CADUSCH et al., 2013).

Na espectroscopia Raman, a intensidade do sinal Raman é proporcional à quarta potência da frequência da radiação incidente e também proporcional à intensidade desta mesma radiação (JACOBSSON; JOHANSSON, 2009; RANDHAWA, 2003). Assim, com o uso de métodos estatísticos adequados, torna-se possível correlacionar a intensidade de determinados picos dos espectros com a concentração de determinadas substâncias presentes na amostra em análise, quando necessário.

2.2 ANÁLISE MULTIVARIADA, QUIMIOMETRIA, ANÁLISE DE COMPONENTES PRINCIPAIS E REGRESSÃO

Com os avanços da tecnologia, é cada vez mais comum a aquisição de grandes conjuntos de dados numéricos que venham a descrever matematicamente alguma amostra, independente de sua natureza, oriundos de técnicas analíticas. E muitos destes dados descrevem a mesma amostra, sob variáveis distintas, sendo então necessário o tratamento simultâneo destas variáveis e observações. Dados desta natureza são denominados como *multivariados*. Ao conjunto de técnicas estatísticas aplicáveis a dados multivariados dá-se o nome de **Estatística Multivariada ou Análise Multivariada**. Esta área pode apresentar diferentes enfoques, a depender tanto do tipo de dado a ser tratado quanto do objetivo da análise. Estas aplicações podem ser classificadas em dois grupos, segundo Sueli Mingoti (2005): técnicas exploratórias de simplificação e técnicas de inferência estatística. No primeiro caso é possível reorganizar os conjuntos de dados, reduzi-los e mesmo transformá-los em outros conjuntos de dados, de forma a facilitar a análise e tratamento dos mesmos; em geral, dispensam o pré-tratamento destas observações. No segundo caso, o objetivo é a validação dos dados, quantificação da relação entre estes e a estimação de parâmetros.

A **Quimiometria** pode ser entendida como um conjunto de técnicas ou métodos de Análise Multivariada, utilizadas para correlacionar estatisticamente alterações (características) em conjuntos de dados na área de Química Analítica. O Jornal *Chemometrics and Intelligent Laboratory Systems* define o termo como “a disciplina que utiliza métodos matemáticos e estatísticos para o design ou seleção

de procedimentos otimizados e experimentos, e para prover o máximo de informação através da análise de dados químicos"(tal definição pode ser encontrada no "Guia para Autores", na página do Jornal, com instruções sobre a natureza do conteúdo dos artigos aceitos para publicação)[1].

Desta forma, é uma área vista como um caminho a ser utilizado para a obtenção das interpretações necessárias de dados matemáticos de experimentos realizados, ajudando no esclarecimento das informações obtidas. São diversas as técnicas utilizadas, testadas e divulgadas como viáveis à utilização. Uma delas, bastante aplicada no campo de prospecção de dados multivariados é a **Análise de Componentes Principais** (PCA, de *Principal Components Analysis*) (ESBENSEN; GELADI, 2009; KEITHLEY; HEIEN; WIGHTMAN, 2009; WOLD; ESBENSEN; GELADI, 1987). É uma técnica estatística utilizada para extrair informações relevantes de grandes conjuntos de dados. Com a sua aplicação, é possível reduzir a dimensionalidade de um conjunto de dados onde há um grande número de variáveis inter-relacionadas. Isto é feito de forma que o máximo de variância presente nos dados seja mantido. Entende-se como variância a informação contida nos dados matemáticos em pauta. Assim, a técnica possibilita a redução do tamanho do conjunto de dados, mantendo ainda assim o máximo possível de informação pertinente. Essa redução se dá pela obtenção de um novo e reduzido conjunto de variáveis não correlacionadas, chamadas **componentes principais**, dispostas em ordem decrescente de variância relativo às variáveis originais. Como resultado, os dados originais podem ser representados por um número menor de variáveis, além da possibilidade de serem reconstruídos como combinação linear destas novas variáveis. Isto é bastante útil em conjuntos de dados espectroscópicos, facilitando a análise e classificação dos dados, separação de variáveis importantes e identificação de padrões entre as amostras, mesmo quando não são facilmente percebidas por análise superficial dos mesmos dados (ADAM; SHERRATT; ZHOLOBENKO, 2008; ARRUDA et al., 2003; DA SILVA, 2008; IM et al., 2009; RIBEIRO, 2012). Desta maneira, a Análise de Componentes Principais seria uma

1 Ver em: http://www.elsevier.com/journals/chemometrics-and-intelligent-laboratory-systems/0169-7439/guide-for-authors#2002

aplicação inserida no primeiro grupo de classificação citado anteriormente. Sua aplicação, atualmente, se dá em diversos campos; inclusive na Química Analítica.

Quando aplicada no campo da Química Analítica, em especial com dados espectroscópicos, a técnica de PCA apresenta diversas vantagens, dentre elas: a possibilidade de se manipular simultaneamente todos os dados que compõem a informação total, sem a necessidade de se escolher, a priori, subconjuntos de dados. Uma consequência do tratamento destes dados com o PCA é a identificação das variáveis que melhor descrevem os dados em estudo, através dos valores das novas variáveis presentes nos PC's, denominadas *loadings*. Neste momento, através de alguns critérios, escolhe-se o número de PC's a compor o novo conjunto de dados a ser manipulado (com novas variáveis e com menor dimensão). Isso permite, na espectroscopia, filtrar o sinal principal nos espectros, eliminando o ruído e erros aleatórios dos mesmos (KEITHLEY; HEIEN; WIGHTMAN, 2009). Através da análise das novas variáveis, num segundo momento, é possível também reduzir o tamanho do conjunto de dados original, retirando variáveis com menor expressão (menor peso na descrição dos dados).

Tratando-se de grandes conjuntos de dados, todas estas etapas e procedimentos tornam-se inviáveis de execução, sem a utilização de recursos computacionais. Ramos e colaboradores (1986) e Keithley e colaboradores (2009) enfatizam esta necessidade em seus artigos.

Uma das principais aplicações da matemática na Química Analítica é a construção de modelos os quais consigam explicar a relação entre diferentes tipos de informação. Quando parte destas informações são explicadas como resposta de outras variáveis, numa relação entre variáveis independentes versus variáveis dependentes, define-se estes modelos matemáticos como modelos de regressão, onde as variáveis independentes são também denominadas preditores e as variáveis dependentes denominadas respostas. No campo da Estatística Multivariada, o estudo de modelos de regressão através de múltiplos preditores é denominado **Análise de Regressão Múltipla**. Para o caso particular de uma relação linear entre diversos preditores e diversas respostas, há a **Regressão Linear Multivariada** (MLR - *Multivariate linear regression*), o qual será o modelo de

regressão a ser utilizado neste trabalho (HIDALGO; GOODMAN, 2013; NOGUEIRA, 2007). Por tratar-se de conjuntos de dados de grandes dimensões, a organização dos mesmos sob a forma matricial torna-se mais conveniente para a manipulação e processamento. Utilizando a notação matricial, o modelo de regressão multivariada pode ser expresso como

$$\mathbf{Y}=\mathbf{X}.\boldsymbol{\beta}+\boldsymbol{\varepsilon} \quad (1)$$

Sendo:

Y – matriz *n* x *m*, de n amostras e m respostas

X – matriz *n* x *p*, de n amostras e p variáveis

β – matriz *p* x *m*, de estimadores de parâmetros para as *p* variáveis

ε – matriz de erro *n* x *m*

Desta forma, para o presente trabalho, é possível processar simultaneamente a predição de concentrações de mais de uma substância, nos mesmos dados espectrais.

O procedimento de **calibração** consiste na obtenção de um modelo matemático que permita realizar a correlação entre duas grandezas, possibilitando a predição de respostas (valores numéricos) a um determinado sistema em função de outros parâmetros que venham a descrevê-lo. Um modelo matemático tem como objetivo facilitar a compreensão do sistema observado, seja ele de natureza física, química ou biológica.

A calibração pode ser conduzida tanto a partir de dados univariados, quanto a partir de dados multivariados, quando o processo envolve simultaneamente várias respostas, ou várias variáveis, adentrando no campo da Análise Multivariada. Neste último caso, segundo Ramos e colaboradores (1986), pode ser classificada em quatro grupos, dos quais destacaremos dois:

a) Calibração indireta, quando faz-se necessário estimar estatisticamente o modelo descritor e este envolve dados de várias amostras e sinais multirespostas (como espectros), ou necessita de resultados de um outro método para a estimação de parâmetros;

b) Calibração direta, quando o modelo a ser utilizado é conhecido.

Aplicada à espectroscopia, as etapas de um procedimento de calibração consistem em:

a) Organizar um conjunto de dados, composto de espectros das substâncias de interesse, cujas concentrações são conhecidas, de modo a servir como padrão de referência para os demais espectros a se analisar (*trainning set*);

b) Ajustar estes dados, eliminando ruídos e interferências nos espectros, além das informações redundantes, através de métodos estatísticos;

c) Realizar a regressão, determinando-se a relação entre as concentrações e estes espectros, estimando parâmetros de relação entre;

d) Realizar a predição das concentrações das amostras desconhecidas, com base no modelo de regressão obtido.

Percebe-se que as expressões “calibração” e “regressão” são comumente confundidas, devido principalmente pela diferença de interpretação entre os profissionais de diferentes campos de atuação. A primeira expressão, em geral, é utilizada pelos profissionais das áreas aplicadas, como a Química Analítica, e Ciências da saúde; já a última refere-se a uma denotação mais formal, normalmente utilizada na Estatística e áreas correlatas. Conclui-se, ainda, que a calibração consiste na primeira das várias etapas da Análise de Regressão, justamente a etapa da criação do modelo matemático ajustado. Neste trabalho, será esta a interpretação adotada. A Análise de Regressão consiste em um campo muito mais amplo onde, além da criação do modelo matemático, se avalia a sua coerência e limitações.

Para uma compreensão mais aprofundada das etapas e requisitos para a elaboração de um modelo quimiométrico, desde o dimensionamento adequado da quantidade de amostras até a etapa de predição de dados, passando pelas etapas de calibração e validação, recomenda-se a leitura de Kramer (1998).

3 METODOLOGIA

3.1 PREPARO DAS AMOSTRAS

Foram escolhidos como substâncias para a realização dos estudos o nitrato de sódio, o glicerol, a água destilada e a raminose. No Laboratório de Biotecnologia e Ecologia de Microorganismos (LABEM), localizado no Instituto de Ciências da Saúde da UFBA, foram preparadas diversas soluções a partir destas substâncias citadas, como segue na Tabela 1 abaixo:

Tabela 1 - Amostras preparadas para o estudo espectroscópico

Substância	Concentração
Água destilada	
Glicerol	0,5%
Glicerol	1,0%
Glicerol	2,5%
Glicerol	5,0%
Glicerol	10,0%
Glicerol	20,0%
Glicerol	25,0%
Glicerol	50,0%
Glicerol	75,0%
Glicerol	PA
Nitrato de sódio	0,5g/l
Nitrato de sódio	1,0g/l
Nitrato de sódio	2,0g/l
Nitrato de sódio	2,5g/l
Nitrato de sódio	5,0g/l
Nitrato de sódio	10,0g/l
Nitrato de sódio	PA
Nitrato + Glicerol	10 g/l + 10%
Nitrato + Glicerol	5 g/l + 5%
Nitrato + Glicerol	2,5 g/l + 2,5%
Raminose	0,25g/l
Raminose	0,5g/l
Raminose	1,0g/l
Raminose	2,0g/l
Raminose	PA
Nitrato + Raminose	0,5g/l + 0.5g/l
Nitrato + Raminose	0,5g/l + 1g/l
Nitrato + Raminose	1,0g/l + 0,5g/l
Nitrato + Raminose	1,0g/l + 1,0g/l
Nitrato + Raminose	2,5g/l + 0,5g/l
Nitrato + Raminose	5,0g/l + 1,0g/l

As substâncias escolhidas e suas respectivas concentrações são justificadas por suas utilizações em meios de cultura amplamente utilizados na literatura, com o Meio Salino Mineral (MSM), aplicado no cultivo de alguns microorganismos, contendo, dentre outras substâncias, o nitrato de sódio como fonte de Nitrogênio e o Glicerol, como fonte de Carbono. A raminose é uma molécula polar, constituinte de uma outra molécula anfipática chamada de Raminolipídio, frequentemente utilizada para a quantificação indireta desta última, a qual é de grande interesse científico e industrial, por suas possíveis aplicações (ABDEL-MAWGOUD; LÉPINE; DÉZIEL, 2010; CENTENO DA ROSA et al., 2010; CHRZANOWSKI; ŁAWNICZAK; CZACZYK, 2011; LEITERMANN et al., 2008; SILVA et al., 2010).

3.2 DEFINIÇÃO DOS PARÂMETROS DE LEITURA E OBTENÇÃO DOS ESPECTROS RAMAN

No Laboratório Multiusuário de Espectroscopia Eletrônica da UFBA – LAMUME –, localizado no Laboratório de Física Nuclear Aplicada do Instituto de Física da UFBA, foram realizadas as leituras dos espectros Raman das amostras selecionadas e preparadas para o projeto. A alíquota de cada amostra foi da ordem de 200 microlitros, para as soluções, enquanto que para as amostras sólidas a quantidade foi livre. Foi utilizado o espectrômetro Raman, modelo NRS-5100 da Jasco Incorporated USA, constituído de laser de comprimento de onda de 532 nm (cor verde), 6 mW de potência, grade de difração de 1800 linhas por milímetro e lente objetiva de 20x. Foram preliminarmente realizadas leituras dos espectros em diversas configurações dos parâmetros disponíveis para ajuste do espectrômetro, para a definição do ajuste mais adequado em termos de qualidade dos espectros a se trabalhar. As duas configurações escolhidas e utilizadas foram:

a) Espectros como resultado de 10 aquisições (leituras da amostra) com duração de 15 segundos, cada leitura, no intervalo de 100 cm^{-1} a 3000 cm^{-1};

b) Espectros como resultado de 15 aquisições com duração de 10 segundos, cada leitura, no intervalo de 100 cm^{-1} a 3000 cm^{-1}.

Ou seja: cada espectro obtido é resultado da média aritmética de 10 espectros que foram capturados ao longo de 15 s, cada uma das 10 capturas, para a primeira

configuração. Na segunda configuração, a interpretação é a mesma, apenas invertendo-se o número de leituras com o tempo de exposição para cada leitura. Em ambas as configurações, o tempo total de exposição de cada amostra para a obtenção do espectro final foi de 150 segundos. Também, em ambas as configurações, não há a apresentação por parte do software do erro atribuído às medidas, ou o nível de dispersão das 10 ou 15 medidas. Sendo assim, será considerado como nula a dispersão das medidas e o valor apresentado em cada posição de deslocamento Raman será tratado como o *Valor Esperado* (ou seja, será considerado que não houve erro na medida), além do fato que o equipamento encontrava-se devidamente calibrado.

Para cada amostra, foram obtidos de 3 a 5 espectros sob o mesmo procedimento, onde a quantidade de espectros coletados variou em função da qualidade dos espectros obtidos (avaliação visual, em função dos ruídos no sinal). É importante registrar que a cada espectro obtido, a alíquota da amostra era substituída, com o objetivo de se evitar os efeitos da evaporação da alíquota por aquecimento (maiores detalhes em “Resultados e Discussão”). Assim, foram obtidos os espectros Raman de todas as substâncias utilizadas no projeto, tanto no seu estado puro, na medida do possível, quanto os espectros de suas diversas diluições e combinações de misturas, totalizando 32 amostras e 146 espectros Raman destas mesmas amostras acima citadas.

3.3 ORGANIZAÇÃO DOS DADOS ESPECTROSCÓPICOS

Cada espectro obtido foi convertido através de software específico do equipamento em dados numéricos no formato de coordenadas cartesianas (Raman shift *vs*. intensidade) e salvo em arquivos cujos formatos pudessem ser manipulados por outros softwares, como o MATLAB, Excel, R, Octave, Statística, etc. Após obtidos estes espectros das soluções selecionadas e preparadas, estes dados numéricos foram organizados e agrupados em matrizes de forma a se realizar os estudos e tratamentos matemáticos planejados.

3.4 APLICAÇÃO DA ANÁLISE DE COMPONENTES PRINCIPAIS E REGRESSÃO MULTIVARIADA

De posse dos dados numéricos dos espectros Raman das amostras das substâncias já citadas, realizou-se a etapa exploratória destes. Alguns destes espectros foram selecionados para utilização como padrão de calibração, através da estimação de parâmetros de equações de regressão, de modo a se determinar a concentração de outras amostras com base nestas calibrações (predição de resultados). Para tal, foram criadas 8 matrizes, contendo como informações os dados espectroscópicos de determinadas amostras, sendo que cada matriz reponde por uma combinação diferente de dados de calibração.

As oito matrizes foram organizadas como segue na Tabela 2.

Tabela 2 - Matrizes preparadas para criação de modelos de calibração e as respectivas amostras presentes.

Matriz	Tipos de amostra
1	Água Nitrato de sódio
2	Água Raminose
3	Água Glicerol
4	Água Nitrato de sódio + Raminose
5	Água Nitrato de sódio Raminose
6	Água Nitrato de sódio Raminose Nitrato de sódio + Raminose
7	Água Nitrato de sódio Raminose Nitrato de sódio + Raminose Glicerol
8	Água Nitrato de sódio Glicerol

Estas foram assim concebidas de forma a se avaliar a influência da natureza dos dados utilizados nos padrões de calibração com finalidade de predição de resultados.

Cada uma das oito matrizes foi submetida à aplicação da Análise de Componentes Principais, utilizando a matriz de covariância dos dados originais centrados na média.

Na Análise de Componentes Principais, cada linha do conjunto de dados da matriz corresponde aos dados de uma única amostra, enquanto cada coluna da matriz corresponde a uma variável da amostra. Neste estudo, as matrizes submetidas ao PCA foram organizadas de forma que cada amostra no conjunto de dados representasse um único espectro Raman, enquanto cada variável corresponderia a uma posição de deslocamento Raman (em cm^{-1}). Como todos os espectros foram obtidos no mesmo intervalo de deslocamento Raman (de 100 cm^{-1} a 3000 cm^{-1}), todas as matrizes do estudo possuem 2901 colunas, ou seja, todas as amostras possuem 2901 variáveis, cada uma.

Após a aplicação do PCA, foi realizada a avaliação do número de componentes principais (PC’s) a ser considerado, descartando as componentes de menor variância de acordo com o critério estabelecido e consequentemente reduzindo a dimensão da nova matriz e, em seguida, reconstruindo a matriz original, com base apenas nas PC's selecionadas. A partir de então, através da matriz transformada, foi criado o modelo de regressão linear multivariado, relacionando os dados espectrais das amostras com as respectivas concentrações, utilizando o ajuste por Mínimos Quadrados múltiplos, obtendo-se os parâmetros estimadores das equações de regressão. Em seguida, foram realizados testes de validação do modelo de predição obtido, determinando-se a concentração de amostras já conhecidas, através de seus espectros Raman.

4 RESULTADOS E DISCUSSÃO

4.1 ANÁLISE QUALITATIVA DOS ESPECTROS

4.1.1 Independência dos espectros

Os gráficos da Figura 11 demonstram a ideia de impressão digital molecular para a espectroscopia Raman. O espectro da mistura entre glicerol e nitrato de sódio é exatamente a composição dos espectros das substâncias isoladas. Em algumas situações, quando, por exemplo, as diluições são muito grandes, pode ocorrer o mascaramento de alguns picos dos espectros, só sendo possível identificá-los através de tratamento matemático dos dados.

4.1.2 Análise da influência das concentrações das amostras na intensidade dos espectros

A Figuras 3 e 4 apresentam os espectros Raman de amostras de glicerol e nitrato de sódio em três e cinco concentrações diferentes, respectivamente. Aqui é possível demonstrar a semelhança de comportamento entre a espectroscopia de espalhamento Raman e a espectroscopia de absorção do Infravermelho, no que compete à lei de Beer-Lambert, onde a intensidade dos espectros é proporcional à concentração das amostras (RANDHAWA, 2003; SWINEHART, 1962). Isto está nítido no grupo de amostras das diluições de nitrato de sódio (0,5; 1,0; 2,0; 2;5 e 5,0 g/l) e nas amostras de glicerol (a 2,5%; 25% e 50%), todas estas com espectros adquiridos sob os mesmos parâmetros de leitura, com 10 aquisições de 15s (cada espectro representado nos gráficos corresponde à média aritmética dos 5 espectros obtidos e cada um destes 5 espectros é resultado da média das 10 aquisições já citadas).

4.1.3 Análise da influência do tempo de exposição das leituras sobre os espectros obtidos

Também, visualmente, é possível identificar a influência do tempo de exposição da amostra ao laser utilizado na obtenção dos espectros, aumentando a intensidade do sinal Raman dos espectros, assim como a aplicação da *Lei dos Grandes*

Números, quando o aumento do número de aquisições (amostras) resulta na diminuição dos erros aleatórios (ruídos de background) nos espectros. As Figuras 5 a 8 demonstram tais observações, comparando espectros obtidos das mesmas amostras (com mesmas concentrações), mas com tempos e quantidades de aquisições diferentes (10 aquisições de 15 s cada *versus* 15 aquisições de 10 s, cada). Aqui, é possível demonstrar que o tempo de coleta individual de cada espectro é realmente importante, uma vez que o tempo total de exposição para leitura é o mesmo nos dois casos (150 segundos). Os espectros resultantes das coletas com menos amostras e maior tempo (simplificaremos a notação para 10/15s) apresentam valores maiores na intensidade, porém com muito mais perturbações na linha de base dos espectros, evidenciando a natureza ruidosa dos mesmos. Já nos espectros resultantes das coletas com mais aquisições e menor tempo (onde simplificaremos a notação para 15/10s), observa-se o comportamento diametralmente oposto, quando as intensidades dos espectros das mesmas amostras (nas mesmas concentrações, lembrando) são sensível e visivelmente menores, além do nível de ruído dos mesmos espectros ser consideravelmente menor. Por estes motivos, decidiu-se adotar como padrão de leitura para as etapas seguintes do trabalho a configuração 15/10s.

Um outro fenômeno que influenciou nesta decisão foi a observação de que um tempo maior de exposição da amostra ao laser do espectrômetro tinha como consequência um ligeiro aquecimento da solução da amostra, levando à evaporação da mesma. Em alguns casos, foi possível perceber que ocorria a elevação da intensidade dos picos dos espectros (consequência de elevação de concentração devido à evaporação), quando a alíquota da amostra não era substituída. (Gráficos não apresentados aqui).

4.1.4 Análise da influência da diluição das amostras em água

Os gráficos da Figura 10 apresentam o efeito da diluição das amostras em baixas concentrações. Os espectros das diluições de raminose praticamente desaparecem, se comparados com o espectro da raminose pura (Figura 9) apresentando majoritariamente o espectro da água, o solvente da mistura (Figura

10). Isto se deve, possivelmente, à pequena proporção do soluto, em relação à água. Tal condição requer um estudo à parte.

Observa-se, também, no gráfico da Figura 4, que os espectros de nitrato de sódio em baixas concentrações apresentam nitidamente alguns dos picos do espectro da água. Contudo, a proporcionalidade é evidenciada mais especificamente no pico da região de 1.100 cm^{-1}, como já foi discutido em seção anterior. Isso leva a concluir que o espectro Raman do nitrato de sódio apresenta pequenas intensidades em seus picos, para estas concentrações.

Kang e colaboradores (2009) e Xu e colaboradores (2013) discutem sobre os efeitos da água no espectro Raman de substâncias que utilizam a mesma como solvente. Nos artigos, demonstram que a interação com a água promove o deslocamento de picos espectrais, assim como em algumas situações, leva a picos presentes nas substâncias desaparecerem, seja pela pequena intensidade destes, diante do espectro da água, seja pela possível quebra de estruturas moleculares no meio aquoso, devido às interações das ligações de hidrogênio, podendo, ainda, ocorrer o mascaramento de alguns picos em virtude das bandas da água. Assim, faz-se necessário um pré-processamento matemático dos dados espectrais para a sua utilização.

Figura 2 - Espectros de Nitrato de sódio e Glicerol em diferentes concentrações, evidenciando a ideia de identidade espectroscópica das substâncias

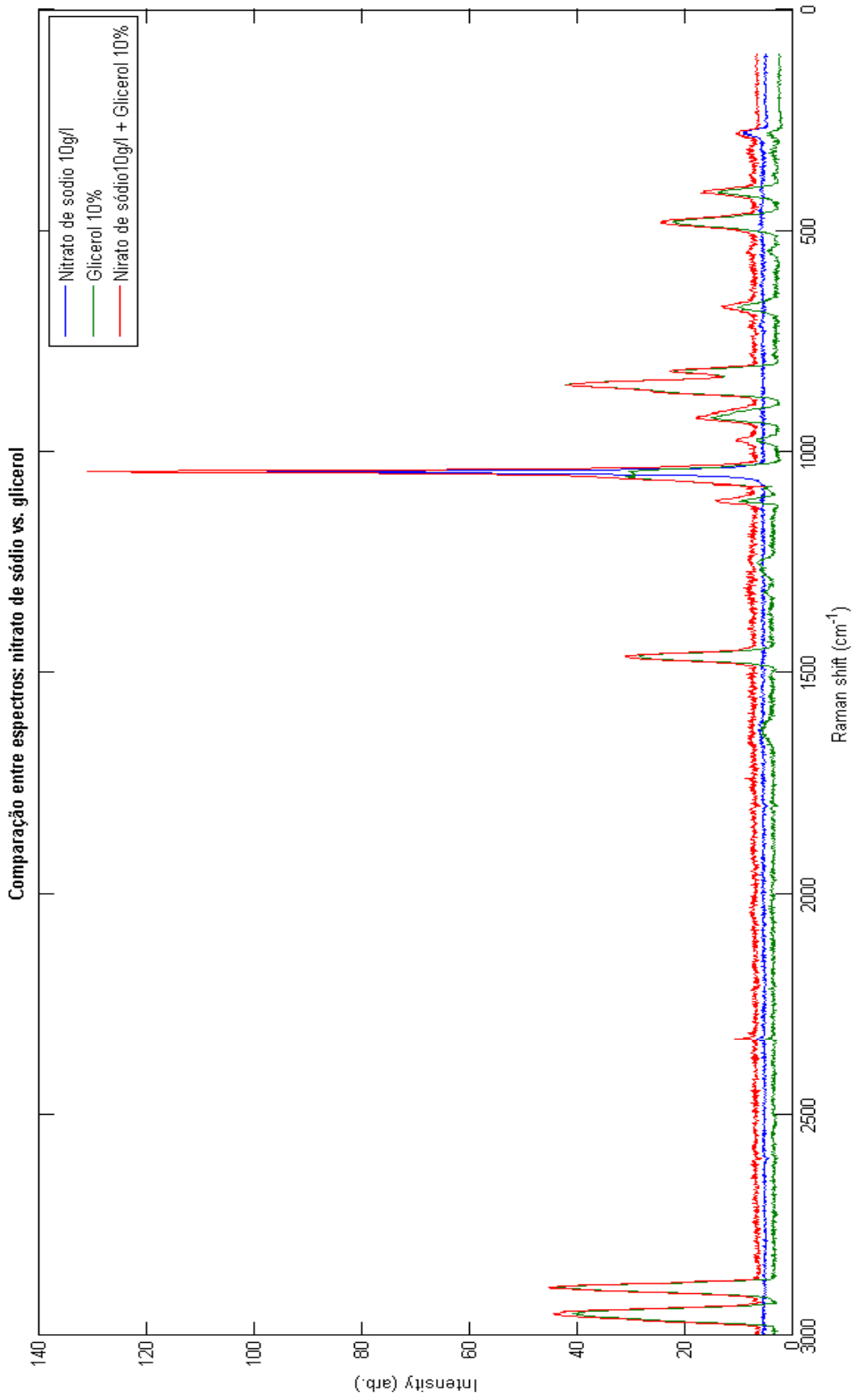


Figura 3 - Soluções de glicerol em diferentes concentrações.

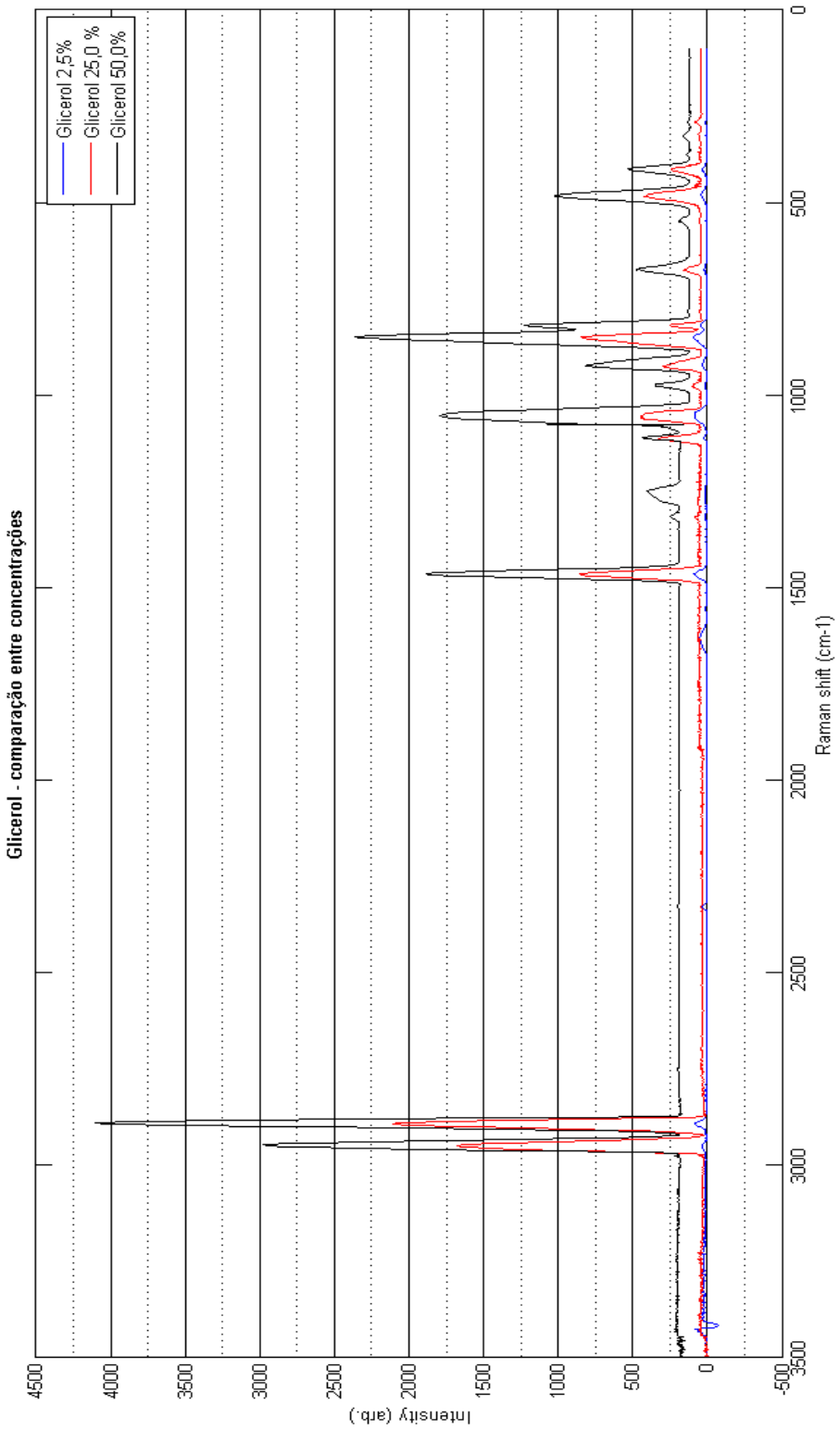


Figura 4 - Soluções de nitrato de sódio em diferentes concentrações.

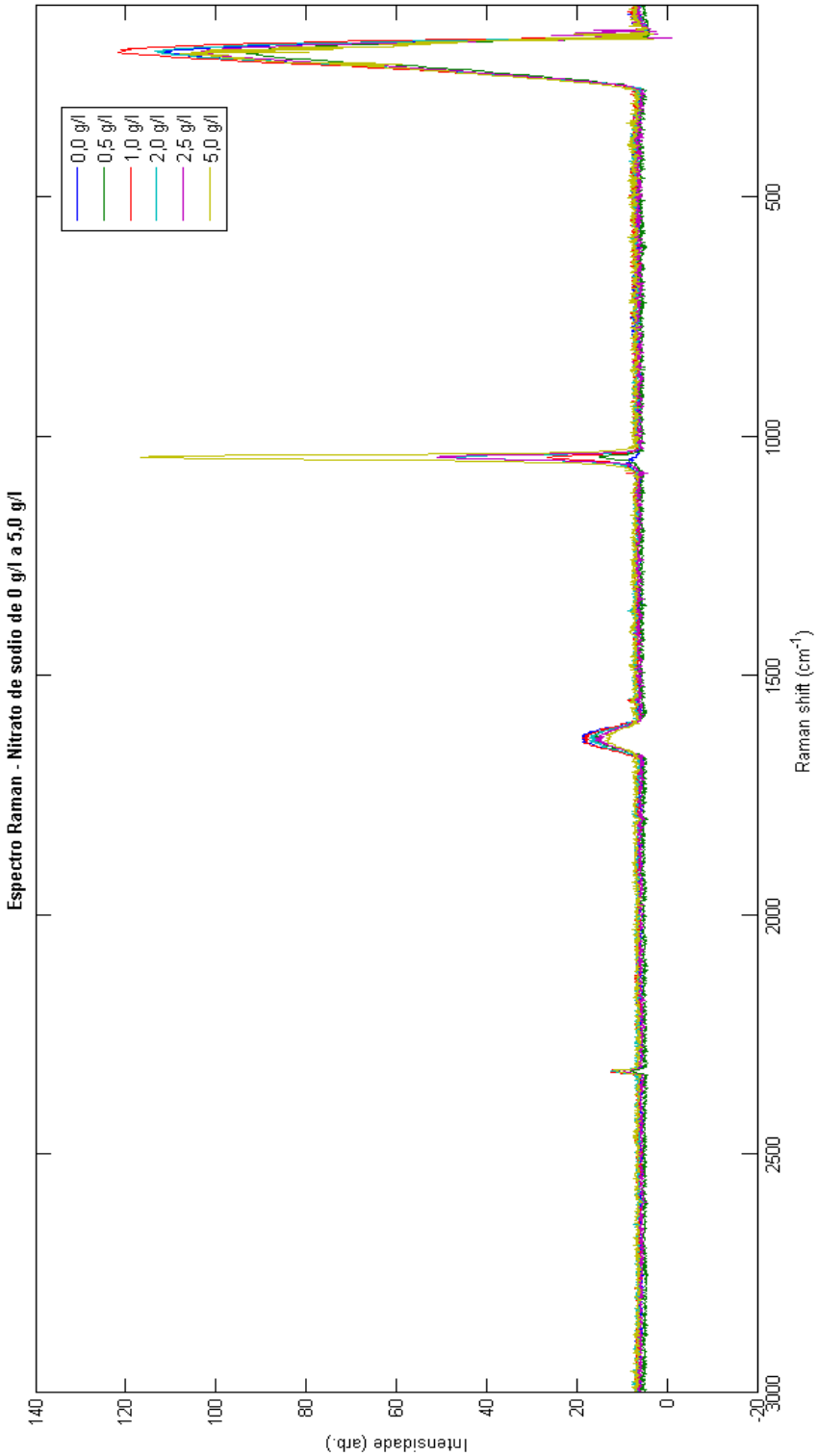


Figura 5 - Glicerol a 2,5%. Relação entre o tempo de exposição da amostra para a obtenção do espectro e a intensidade dos mesmos. Relação entre o número de amostras (número de aquisições) e o nível de ruido presente nos espectros.

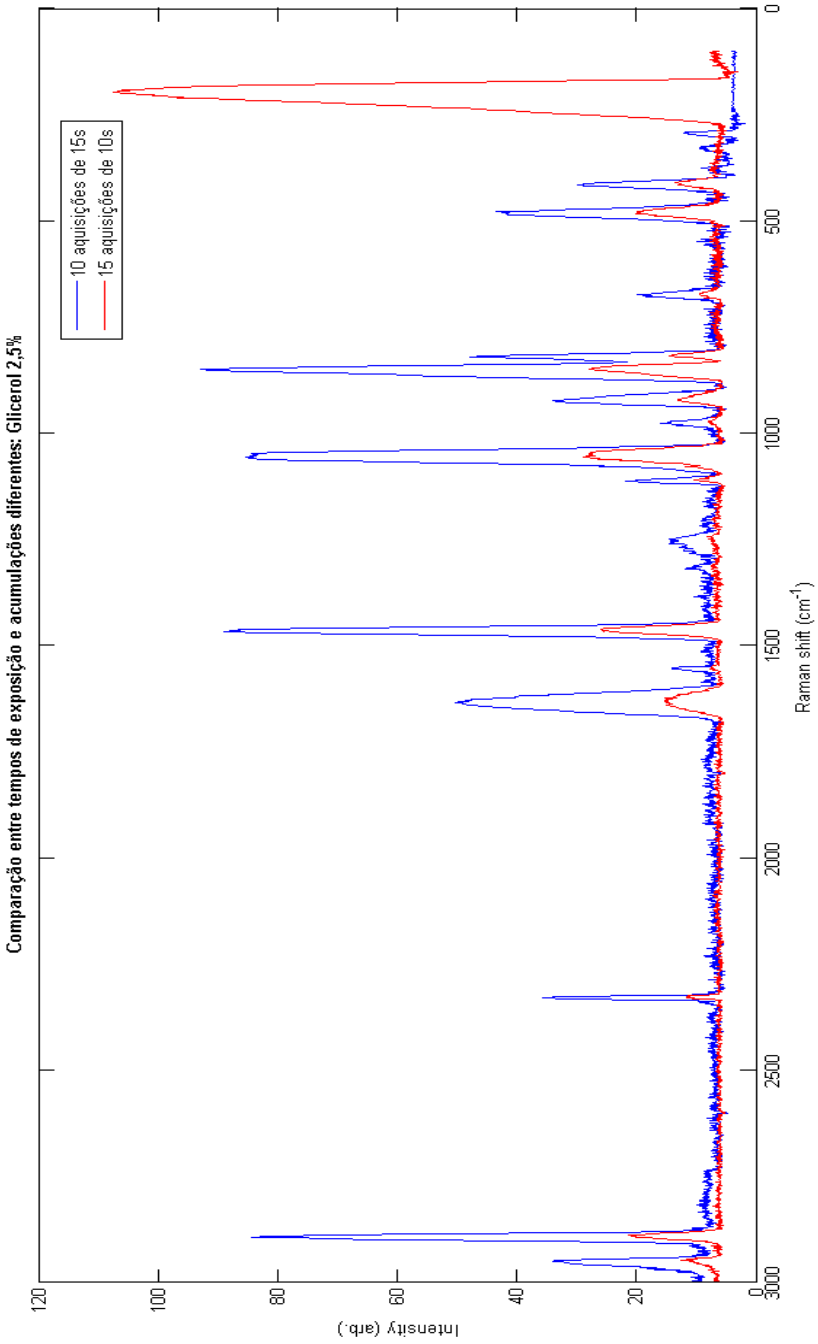


Figura 6 - Glicerol a 10,0%. Mesma avaliação sobre tempo de exposição e número de aquisições.

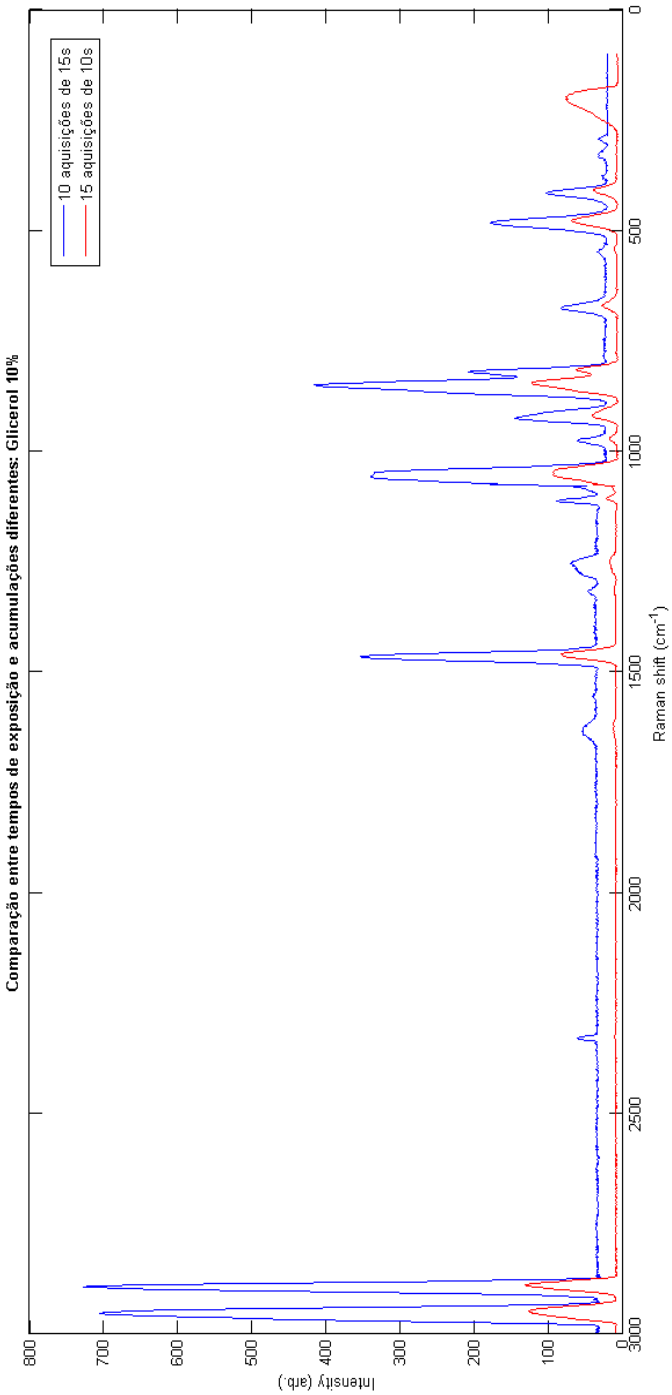


Figura 7 - Nitrato de sódio a 1,0 g/l. Mesma avaliação sobre tempo de exposição e número de aquisições.

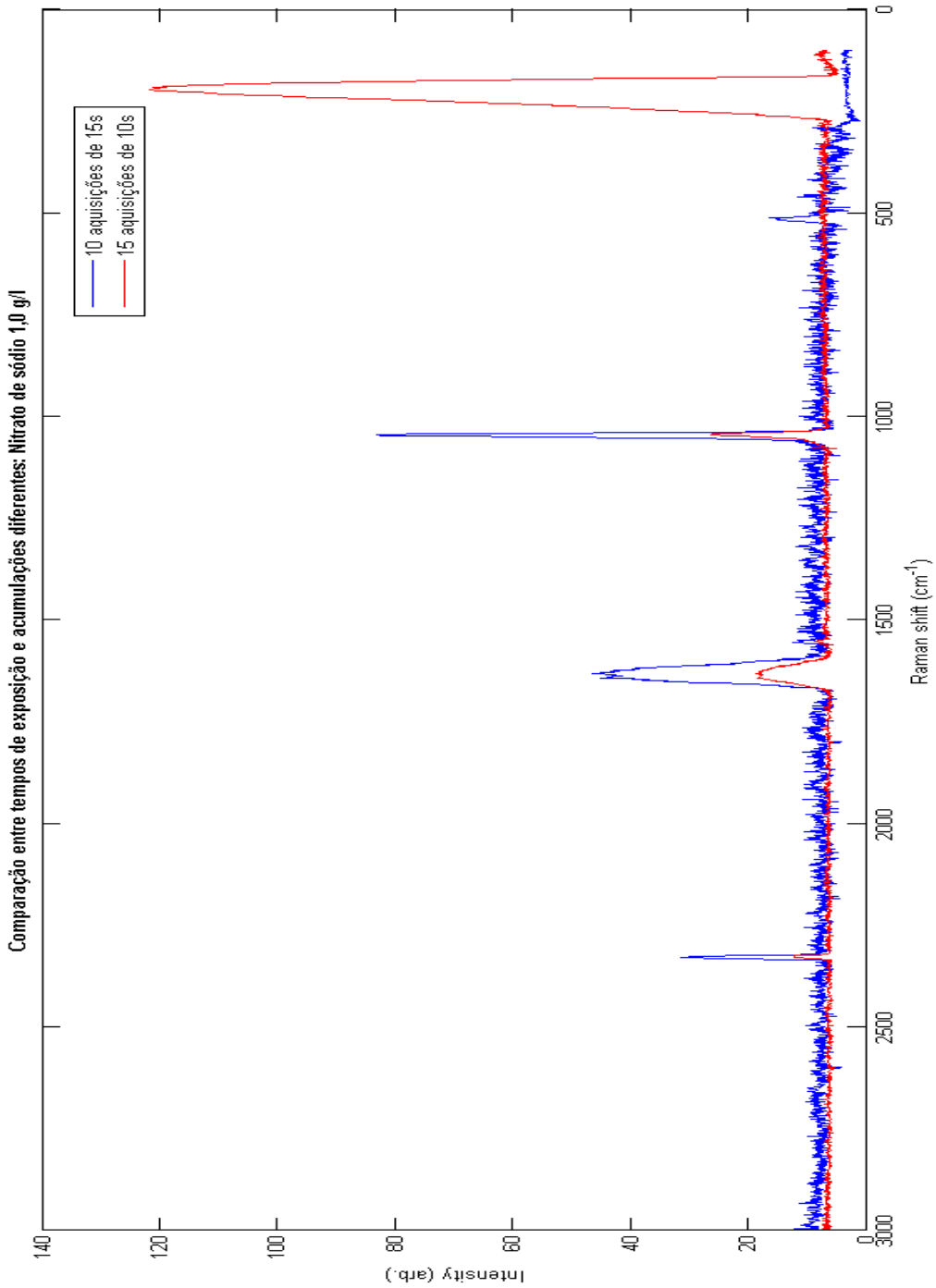


Figura 8 - Nitrato de sódio a 10,0 g/l. Mesma avaliação sobre tempo de exposição e número de aquisições.

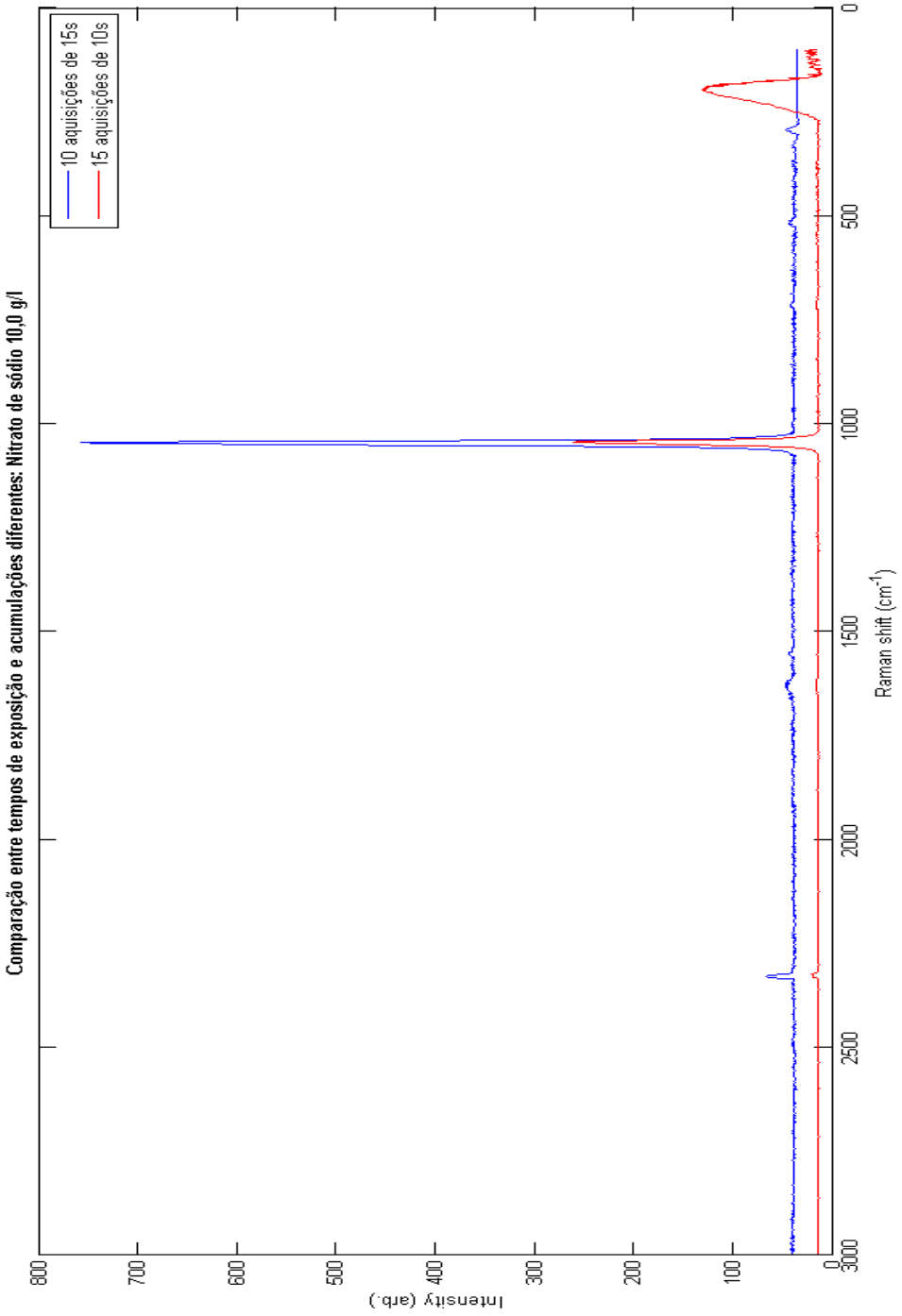


Figura 9 - Espectro Raman da Raminose pura, faixa de 100 a 3500 cm^{-1}

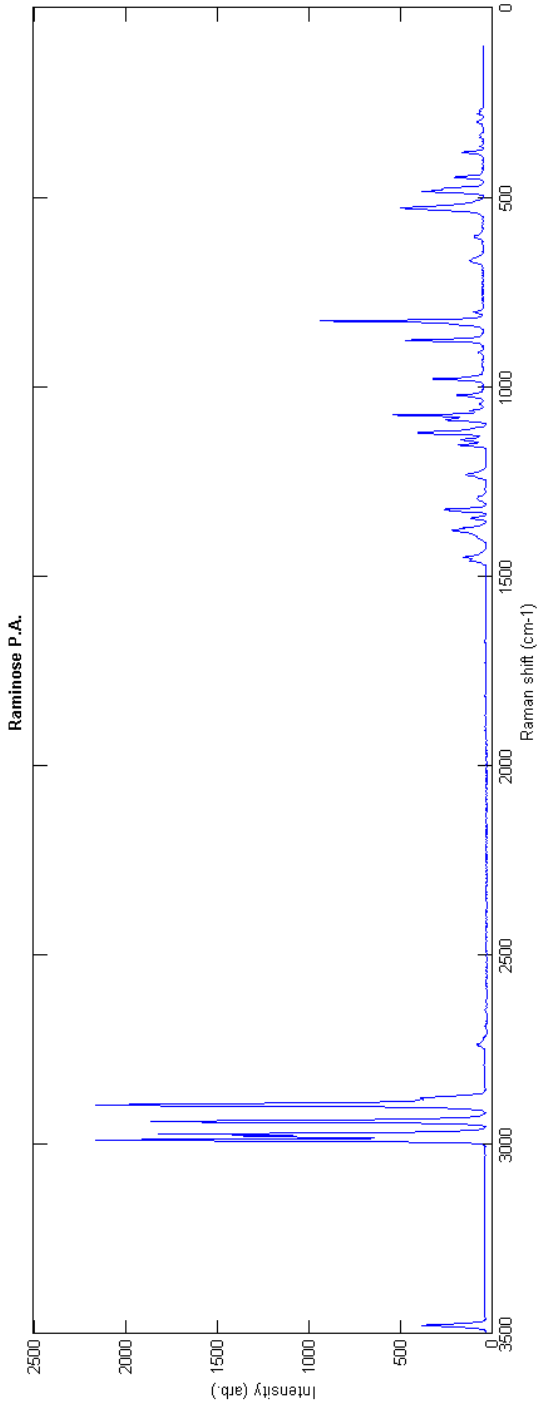
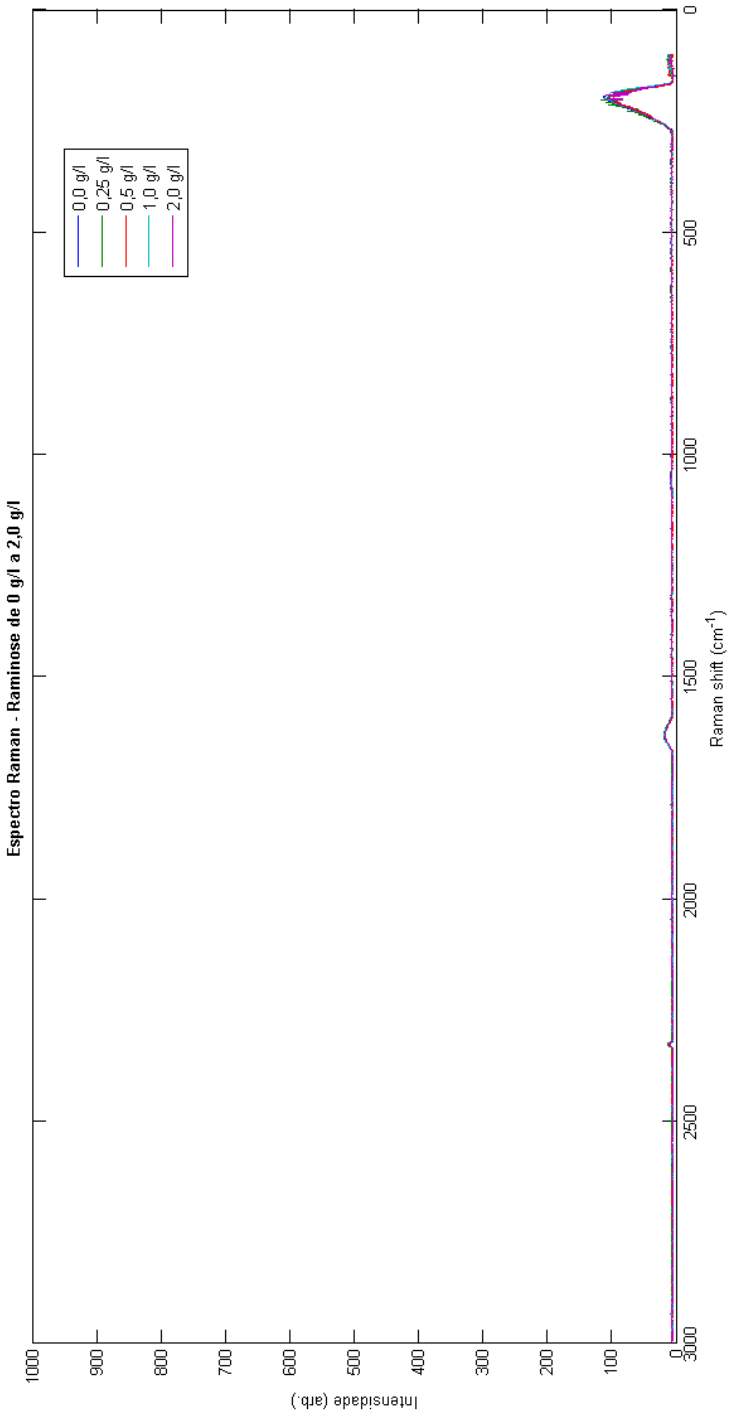


Figura 10 - Espectro Raman de 4 soluções de Raminose diluída em água e mais o espectro da própria amostra de água sobrepostos.



4.2 ANÁLISE EXPLORATÓRIA DOS DADOS: ANÁLISE DE COMPONENTES PRINCIPAIS E REGRESSÃO MULTIVARIADA

4.2.1 Avaliação qualitativa dos dados

Nesta seção, a Matriz 07 será considerada como referência para a discussão, por ser composta pelo maior número de amostras do trabalho[2]. As demais matrizes serão utilizadas para efeito de comparação de resultados parciais.

A Matriz 07, composta pelas amostras listadas na Tabela 3, possuía 92 linhas e 2901 colunas, sendo cada coluna uma variável descritora (um valor de intensidade de uma determinada posição espectral no Raman shift) e cada linha a representação de uma amostra (um espectro). Dados em matrizes como esta e as demais utilizadas neste trabalho são denominados por Koch (2013) como “dados de grandes dimensões” (do inglês HDD – *high-dimensional data*), onde um dos grupos de classificação destacado é aquele em que o número de dimensões (variáveis) é grande, além de ser maior do que o número de amostras do conjunto de dados. Ele chama a atenção para o cuidado que se deve tomar com a escolha do número de componentes a se manipular, uma vez que o número máximo permitido, neste caso, é o valor da dimensão n, que descreve a quantidade de amostras, uma vez que $n<d$. Assim, para a Matriz 07, o maior número possível de componentes a se escolher seria 92.

A Figura 11 apresenta os gráficos das 4 primeiras componentes principais (também chamadas de *loadings*) do conjunto de dados da Matriz 07. Cada ponto no eixo horizontal corresponde a uma das variáveis (coeficientes) da Componente Principal em questão. Observando o perfil extremamente ruidoso da PC 04, é possível comprovar a hipótese de que as principais informações dos espectros encontram-se nas primeiras Componentes, com maiores autovalores.

Analisando as projeções dos escores em função das componentes 01 e 02, nos dados da Matriz 07 (Figura 12), observa-se que algumas amostras se destacam das demais no gráfico, as quais são: nitrato de sódio a 10,0 g/l; glicerol a 5% (v/v);

2Considerando o conjunto de espectros coletados com os mesmos parâmetros.

glicerol a 10% (v/v) e glicerol a 20% (v/v). As amostras de nitrato de sódio a 5,0 g/l e glicerol a 2,5% (v/v) também se distinguem das demais, mas com grande proximidade de um grande *cluster* contendo as outas 68 amostras (em destaque na Figura 12 (b)). Observando o gráfico (b), percebe-se que ainda há alguma separação entre as amostras, contudo menos destacadas, se comparado às seis já citadas. Possivelmente este comportamento seja consequência do alto grau de diluição das amostras que compõem este cluster, onde o espectro da água venha a dominar sobre as particularidades espectrais das demais substâncias envolvidas.

Tabela 3 - Amostras presentes na Matriz de dados 07

Substância	Concentração
Água destilada	
Glicerol	0,5%
Glicerol	1,0%
Glicerol	2,5%
Glicerol	5,0%
Glicerol	10,0%
Glicerol	20,0%
Nitrato de sódio	0,5g/l
Nitrato de sódio	1,0g/l
Nitrato de sódio	2,0g/l
Nitrato de sódio	2,5g/l
Nitrato de sódio	5,0g/l
Nitrato de sódio	10,0g/l
Nitrato + Glicerol	10 g/l + 10%
Nitrato + Glicerol	5 g/l + 5%
Nitrato + Glicerol	2,5 g/l + 2,5%
Raminose	0,25g/l
Raminose	0,5g/l
Raminose	1,0g/l
Raminose	2,0g/l
Nitrato Nitrato de sódio + Raminose	0,5g/l + 0.5g/l
Nitrato Nitrato de sódio + Raminose	0,5g/l + 1g/l
Nitrato Nitrato de sódio + Raminose	1,0g/l + 0,5g/l
Nitrato Nitrato de sódio + Raminose	1,0g/l + 1,0g/l
Nitrato Nitrato de sódio + Raminose	2,5g/l + 0,5g/l
Nitrato Nitrato de sódio + Raminose	5,0g/l + 1,0g/l

Numa aplicação em separado do PCA sobre os dados das misturas de Raminose + Nitrato de sódio (Matriz 04) já é possível identificar padrões entre as amostras participantes (Figura 13). Já na aplicação do PCA sobre os dados isolados das soluções de Raminose e Nitrato de sódio (sem misturas), na Matriz 05,

observou-se o mesmo padrão de agrupamento de amostras apresentado na Matriz 07, como pode ser visualizado nos score plots da Figura 14, com destaque apenas às amostras de maior concentração no conjunto de dados (amostras de Nitrato de sódio a 5,0 e 10,0 g/l, basicamente), tanto para o gráfico PC1 *versus* PC2 (Figura 14 (a)), quanto para o gráfico de PC2 *versus* PC3 (Figura 14(c)). O mesmo ocorreu na aplicação sobre a Matriz 06, contendo tanto as soluções puras, quanto as misturas – apenas as soluções constituídas de maior concentração de soluto conseguiram evidenciar-se diante do conjunto total de amostras (Figura 15).

O estudo das projeções dos escores das componentes principais permitem uma avaliação sobre a natureza das amostras envolvidas, uma vez que é possível ordená-las por nível de influência das mesmas sobre todo o conjunto de dados, assim, como também permite uma avaliação qualitativa do conjunto de dados, distinguindo os conjuntos por similaridade de resultados, além de permitir a visualização de agrupamentos entre os mesmos, o que nem sempre é facilmente identificado através dos dados originais.

Tabela 4 - Resumo dos percentuais de variância explicada em cada matriz, nos casos contendo amostras de água e sem amostras de água.

	Com agua			Sem agua		
Matriz	PC1	PC2	PC3	PC1	PC2	PC3
4	69,77	9,25	5,58	64,30	10,99	7,00
5	71,78	16,58	2,86	72,30	16,02	2,89
6	60,14	26,59	2,86	61,16	25,50	2,89
7	85,23	7,95	3,97	85,40	7,96	3,79

Para avaliação da influência das amostras de água sobre o comportamento dos conjuntos de dados na Análise de Componentes Principais, foi aplicada, também a técnica sobre todas as matrizes retirando-se as amostras de água dos conjuntos. A Tabela 4 apresenta um resumo comparativo do comportamento do percentual de variância das três primeiras componentes das Matrizes 04, 05, 06 e 07, sendo tratadas em duas condições: com amostras de água na composição dos dados; e sem as amostras de água na composição dos dados. O que se observou foi que, efetivamente, apenas a Matriz 04 recebe influência de dominância das amostras de água, pois o nível de variância da primeira componente foi reduzido de 69,77% para 64,30%, com a retirada destas amostras da matriz, além de ocorrer uma pequena

redistribuição entre as três primeiras componentes. Nas outras três matrizes, houve um aumento do nível de variância na primeira componente, significando que tais amostras apresentavam relevância, mas sem comportamento de dominância no conjunto.

Figura 11 - Loadings da Matriz de dados 07, 4 primeiros Componentes (PC's).

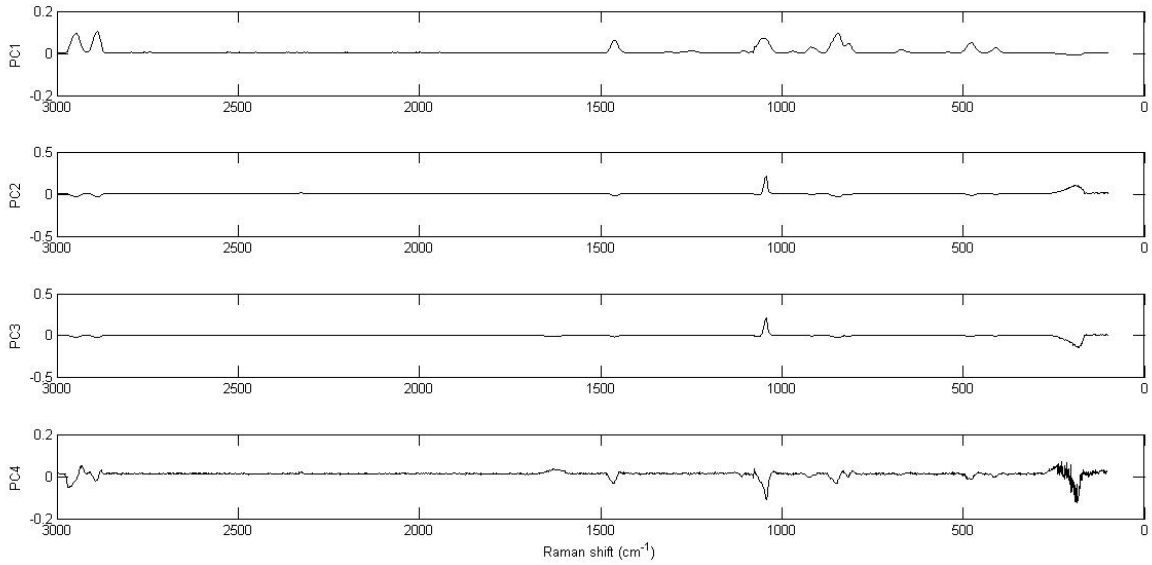


Figura 12 - a) Score plot dos PC's 01 e 02; (b) pontos em destaque no gráfico acima.

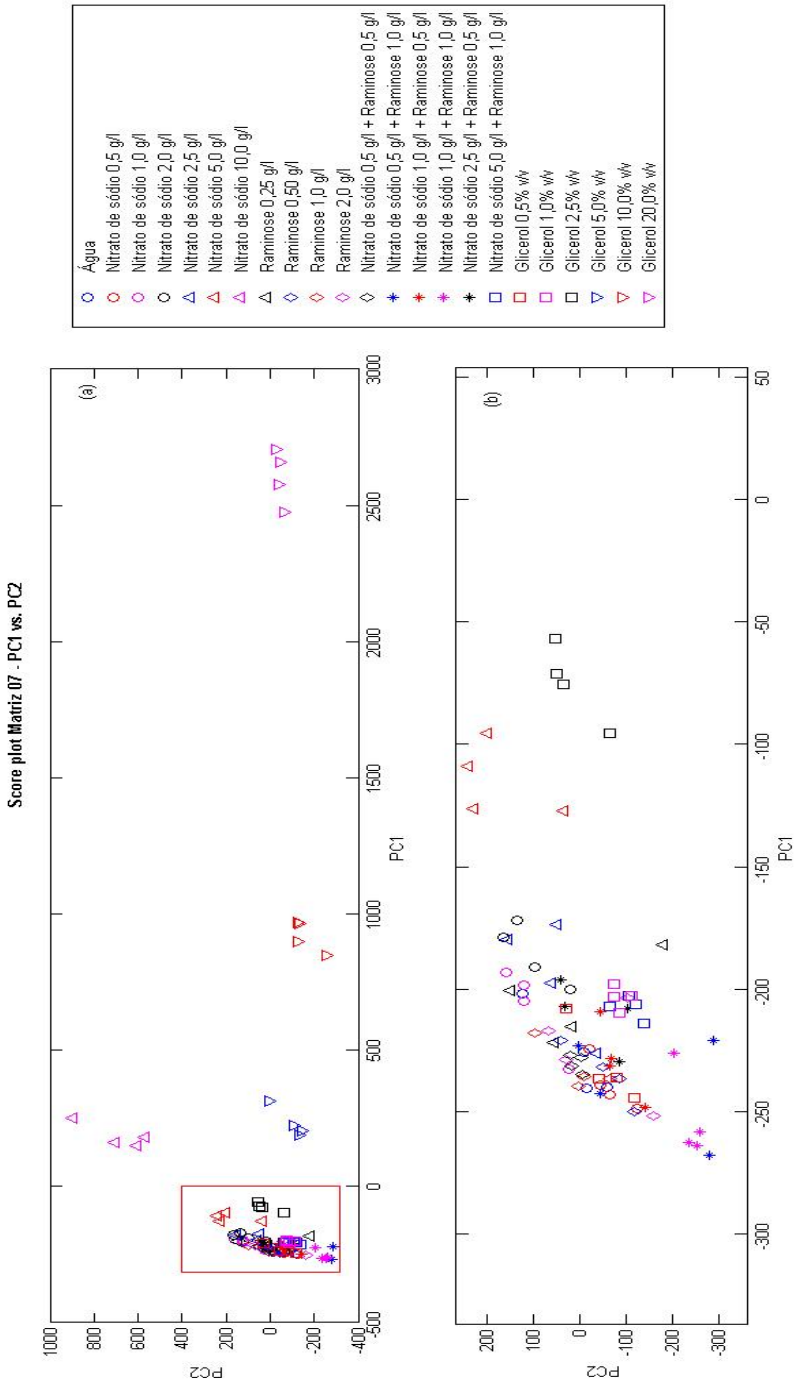


Figura 13 - Score plots dos PC's 01, 02 e 03 da Matriz 04.

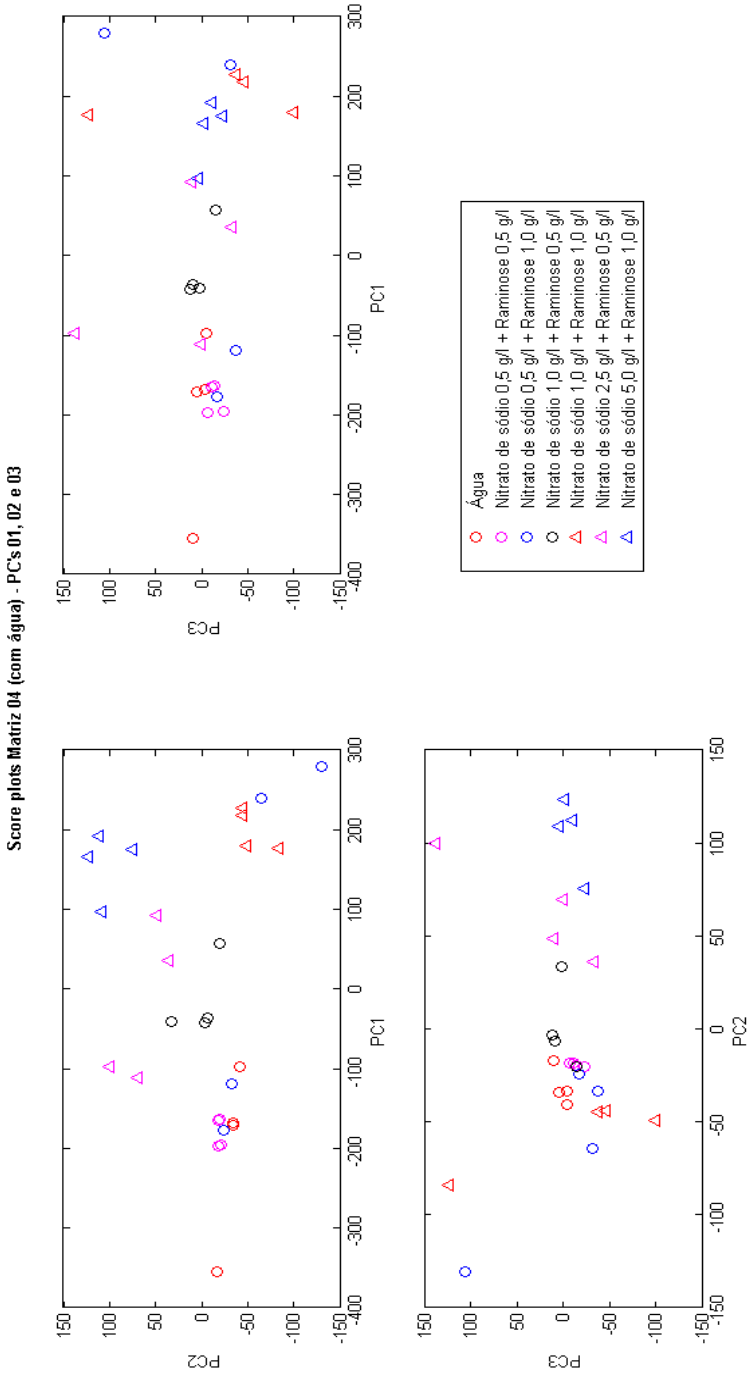


Figura 14 - Score plots da Matriz 05.

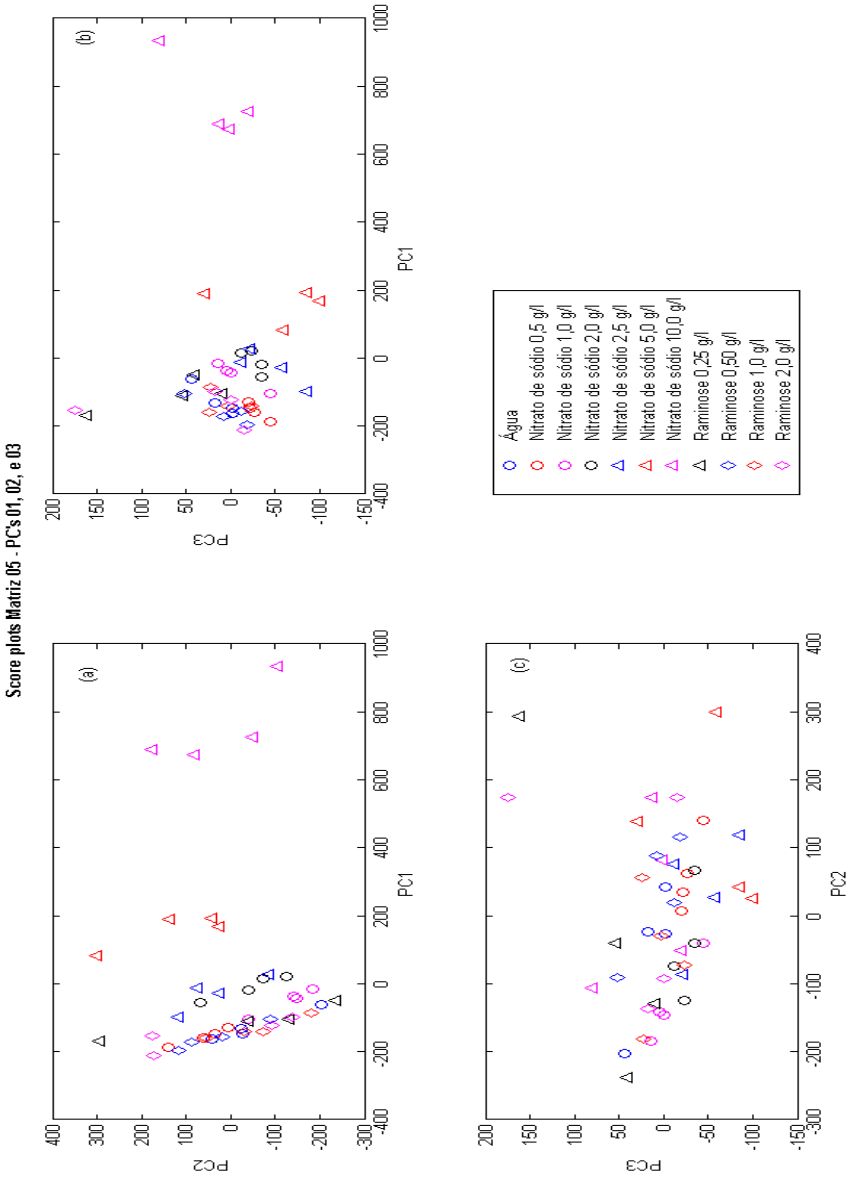
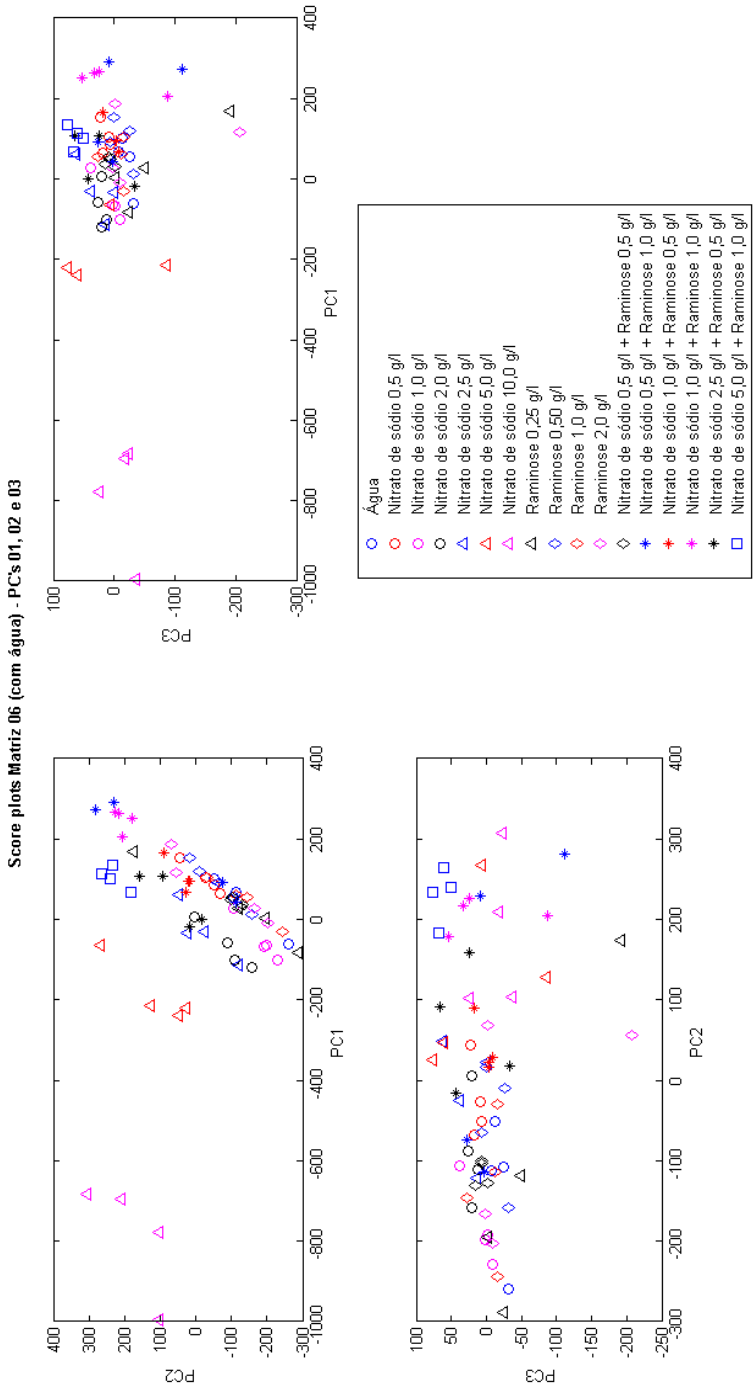


Figura 15 - Score plots da Matriz 06.



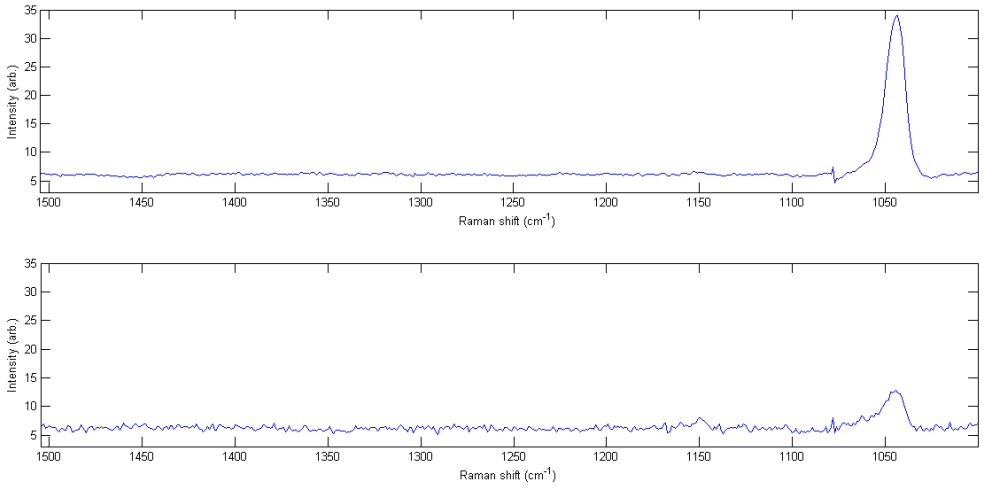
4.2.2 Análise e predição de dados

Para a predição dos valores da concentração das soluções envolvidas neste projeto, foram utilizados os dados numéricos dos espectros para a estimação de parâmetros de regressão multivariada. Contudo, os dados utilizados para tal estimação foram primeiramente filtrados através da técnica de PCA. A determinação do número de Componentes que melhor descreve o conjunto de dados com o mínimo de informação tem como uma das consequências práticas a redução do ruído atribuído aos sinais, uma vez que muito da informação redundante e secundária é eliminado com as Componentes de menor variância. Segundo Keithley, Heien e Wightman (2009), as componentes principais podem ser pensadas como vetores em um sistema de coordenadas abstrato, descrevendo fontes de variância de um conjunto de dados. Assim, após a criação da nova matriz transformada, reconstrói-se a matriz de dimensões originais, mas desta vez em função exclusiva das componentes selecionadas, de acordo com a operação

$$A = S \cdot L^T , \qquad (2)$$

sendo **A** a nova matriz de dados reconstruída, **S** a matriz de escores e **L** a matriz de coeficientes dos autovetores.

Figura 16 - Tôpo: trecho do espectro da solução de Nitrato de sódio 0,5 g/l + Raminose 0,5 g/l reconstruído com apenas 2PC's. Base: o a mesma região espectral sem qualquer tratamento matemático.



A Figura 16 acima mostra um trecho do espectro da mistura de Nitrato de sódio 0,5 g/l + Raminose 0,5 g/l, reconstruído com apenas 2 PC's, em comparação com o mesmo trecho do espectro original, sem tratamento por PCA. Aqui é possível demonstrar a capacidade de filtragem destas informações desnecessárias citadas. Percebe-se, ainda, que com a identificação das componentes mais relevantes, consegue-se destacar alguns picos que antes encontravam-se mascarados pelo sinal ruidoso, como neste caso. Este tipo de comportamento, em especial, requer atenção, uma vez que pode influenciar em resultados de outras técnicas que venham a ser utilizadas em seguida.

Assim, após a etapa de reconstrução dos dados dos espectros contidos nas matrizes, foram realizadas as demais etapas da calibração multivariada destes dados, com o objetivo de utilizar as informações contidas nos espectros para calcular as concentrações das substâncias neles presentes.

Na Regressão Multivariada dos dados, as variáveis independentes representadas consistem nos dados espectrais e as concentrações das substâncias envolvidas são as variáveis dependentes do sistema. A estimativa dos parâmetros foi realizada pelo método dos Mínimos Quadrados Múltiplo, obtendo-se um modelo de regressão da forma

$$C = A \cdot B \tag{3}$$

onde $\mathbf{C}_{nxm}$ (c_{j1}, c_{j2}, …, c_{nm}) corresponde à matriz das concentrações das *p* substâncias presentes nas *n* amostras que tiveram seus espectros coletados, **A** (a_{j1}, a_{j2}, …,a_{ni}) é a matriz contendo os dados espectrais das *n* amostras de calibração, após o processamento por PCA, e **B** corresponde à matriz com os parâmetros de regressão estimados para cada variável a_i.

4.2.2.1. Análise e validação do modelo de calibração da Matriz 07

Como resultado da Análise de Componentes, identificou-se que a primeira PC responde sozinha por 85,23% da variância total dos dados e que as três primeiras componentes respondem por 97,15% da mesma variância total.

A Figura 17 apresenta o *Scree plot* dos dados da Matriz 07. O Scree plot é uma ferramenta gráfica utilizada para explicar o nível de variância explicada por parte de cada componente. Esta representação pode ser feita tanto com os valores absolutos dos autovalores correspondentes a cada componente principal, quanto através de valores percentuais aos quais cada autovalor corresponde em relação à variância total. Desta forma, o scree plot é utilizado para a avaliação do número de componentes principais a se utilizar na redução da dimensão dos dados originais através da criação da nova matriz transformada. Avaliando os gráficos, percebe-se que as duas primeiras componentes principais respondem por mais de 90% da variância total explicada, sendo que apenas a primeira PC já responde por 85,23% da informação total do conjunto de dados. A terceira Componente representa 3,97% desta variância total. Um critério comumente utilizado para a definição do número de componentes a se reter é identificar o valor no qual a curva tende a se estabilizar, ou como usualmente se refere, identifica-se a “região do cotovelo” do gráfico, definindo este ponto como o número limite. Por este critério, conclui-se que apenas 2 PC’s são suficientes para a descrição adequada dos dados originais. Assim, a nova matriz que poderia descrever os dados espectroscópicos da Matriz 07 possui apenas 02 linhas e 2901 colunas.

Contudo, para o objetivo principal deste trabalho, é de maior interesse utilizar a própria matriz original reconstruída a partir das PC’s selecionadas, pois esta já apresenta os espectros com menor interferência de ruídos, após reconstruída, uma vez que este tipo de informação é representado em maiores proporções nas PC’s de maior ordem, as quais foram mantidas. Assim, os dados originais foram reconstruídos através do produto entre a matriz de escores com apenas dois componentes e a matriz dos coeficientes, também com apenas dois componentes.

O resultado das predições foi avaliado através da análise de resíduos, baseado no quadrado da diferença entre os valores medidos e os valores esperados, o qual Da Silva (2008) define como E*statística Q*. O resumo dos resultados obtidos pode ser visto abaixo, na Tabela 5.

Tabela 5 - Resumo do teste de validação cruzada do modelo de calibração da Matriz 07

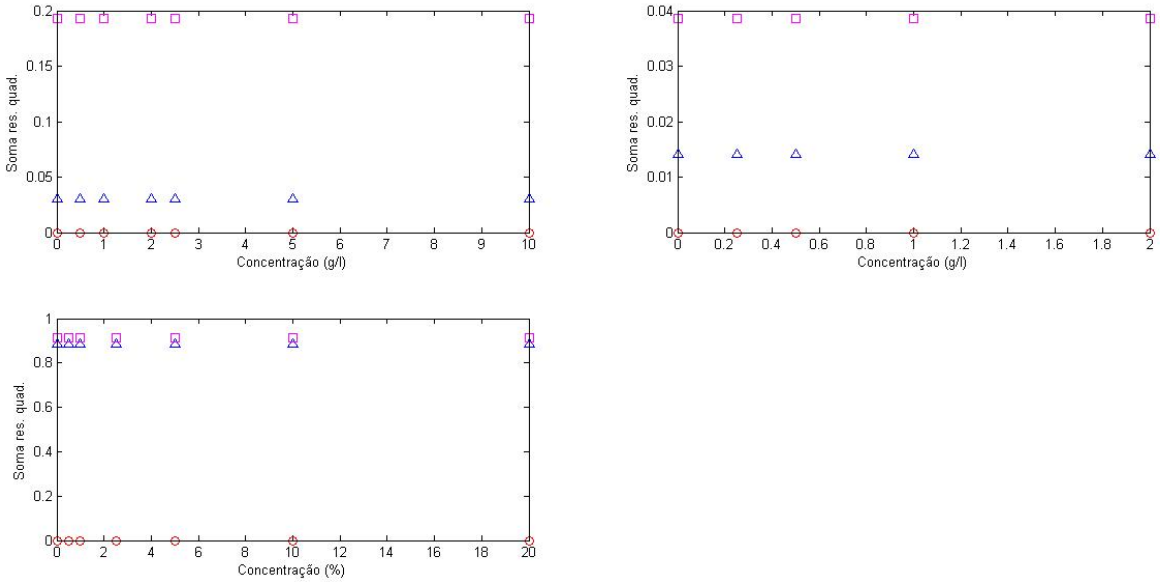
Análise residual – Matriz 07			
Nitrato de sódio			
	Regressão com dados originais	**Regressão com 2PC**	**Regressão com 3PC**
Soma Quad Res	1,024E-26	67,289	47,839
Média Quad Res	1,113E-28	0,731	0,520
Raminose			
	Regressão com dados originais	**Regressão com 2PC**	**Regressão com 3PC**
Soma Quad Res	3,393E-27	27,506	22,635
Média Quad Res	3,689E-29	0,299	0,246
Glicerol			
	Regressão com dados originais	**Regressão com 2PC**	**Regressão com 3PC**
Soma Quad Res	1,497E-26	27,600	26,981
Média Quad Res	1,627E-28	0,300	0,293

Percebe-se claramente que a predição utilizando como calibração todos os dados espectrais, sem qualquer tratamento, apresenta uma melhor performance, o que é natural, por se tratar de uma validação cruzada (utilizou-se os mesmos dados da calibração para o teste de validação). Em contraponto, avaliando o resultado dos resíduos para os preditores utilizados com PCA, pode-se considerar os valores como aceitáveis, sem uma avaliação estatística mais aprofundada. Estratificando a análise, verificando o comportamento para cada concentração individualmente, percebe-se que não houve desvios entre os resultados por categoria de concentração, mantendo-se um padrão, como pode ser visto nos gráficos da Figura 17.

Tabela 6 - Amostras utilizadas para novo teste de validação com a Matriz 07

Substância	Concentração
Água destilada	-
Glicerol	2,5%
Nitrato + Raminose	0,5g/l + 0.5g/l
Nitrato + Raminose	0,5g/l + 1g/l
Nitrato + Raminose	1,0g/l + 0,5g/l
Nitrato + Raminose	1,0g/l + 1,0g/l
Nitrato + Raminose	2,5g/l + 0,5g/l
Nitrato + Raminose	5,0g/l + 1,0g/l

Figura 17 - Gráficos da média dos quadrados dos desvios para cada concentração testada na predição da Matriz 07.



Os gráficos do topo correspondem aos resíduos das amostras de nitrato de sódio e raminose, respectivamente. O gráfico da base descreve os resíduos do glicerol. Para os três gráficos, há pontos representando os resíduos das três predições: Calibração com os dados originais (círculos vermelhos); calibração com os dados reconstruídos com apenas 2 PC's (triângulos azuis); e calibração baseada nos dados reconstruídos a partir das 3 primeiras PC's (quadrados na cor magenta). Observa-se que a regressão com 3 PC's foi a que apresentou pior resultado, dentre as três opções, e que no caso das amostras de glicerol, não houve diferença considerável entre os resultados para as calibrações com PCA.

Para o teste de validação com dados externos retirou-se algumas amostras anteriormente presentes na Matriz 07 (que haviam sido utilizadas inicialmente na calibração) e reservou-as, com o intuito de utilizar na predição. Assim, reaplicou-se a Análise de componentes principais sobre a nova matriz e obteve-se um novo conjunto de preditores. As amostras reservadas para o teste de validação pode ser visualizadas na Tabela 7.

Tabela 7 - Amostras utilizadas para novo teste de validação com a Matriz 07

Substância	Concentração
Água destilada	-
Glicerol	2,5%
Nitrato + Raminose	0,5g/l + 0.5g/l
Nitrato + Raminose	0,5g/l + 1g/l
Nitrato + Raminose	1,0g/l + 0,5g/l
Nitrato + Raminose	1,0g/l + 1,0g/l
Nitrato + Raminose	2,5g/l + 0,5g/l
Nitrato + Raminose	5,0g/l + 1,0g/l

O resumo da análise residual da nova validação pode ser visto abaixo, na Tabela 8. Os resultados de um teste de validação com dados externos aos dados de calibração apresentam um comportamento mais coerente entre os três métodos de calibração. Observa-se que, apesar de mais elevados, comparados à validação cruzada, os valores residuais apresentam-se numa faixa de valores mais próximos entre si, com exceção à regressão com dados originais para o glicerol, que ainda foi nitidamente melhor. Isso se deve, possivelmente, pela natureza dos dados utilizados na nova matriz de calibração: todos os dados do novo *trainning set* eram oriundos de espectros de soluções puras (não haviam espectros de misturas nesta nova matriz; todas elas foram reservadas para a calibração). Este resultado, então, é bastante animador, pois além de apresentar coerência na predição, ainda reforçou a ideia já discutida acima sobre a identidade espectral das substâncias, mesmo que em misturas.

Tabela 8 - Resumo da análise residual do novo teste de validação da Matriz 07

Análise residual – validação Matriz 07			
Nitrato de sódio			
	Regressão com dados originais	**Regressão com 2PC**	**Regressão com 3PC**
Soma Quad Res	3,132E+01	35,129	30,879
Média Quad Res	9,788E-01	1,098	0,965
Raminose			
	Regressão com dados originais	**Regressão com 2PC**	**Regressão com 3PC**
Soma Quad Res	6,998E+00	10,257	9,830
Média Quad Res	2,187E-01	0,321	0,307
Glicerol			
	Regressão com dados originais	**Regressão com 2PC**	**Regressão com 3PC**
Soma Quad Res	1,178	7,724	11,114
Média Quad Res	0,037	0,241	0,347

4.2.2.2. Análise e validação do modelo de calibração da Matriz 04

A Matriz 04, avaliada, foi composta de 28 amostras entre água e soluções aquosas de nitrato de sódio com raminose, em diversas concentrações. Semelhante ao tratamento realizado com a Matriz 07, a regressão foi obtida tanto através dos dados originais dos espectros, quanto através dos dados modificados pela Análise de Componentes Principais, para efeito de comparação. Na aplicação do PCA sobre as matrizes de covariâncias dos dados originais, identificou-se que as quatro primeiras componentes respondiam por 87,5% da variância total dos dados espectrais (ver Figura 18). Com base nisto, decidiu-se reconstruir a matriz original utilizando 3 e 4 Componentes Principais, para em seguida aplicar a Regressão Múltipla por Mínimos Quadrados e então comparar os resultados.

Os valores ajustados das concentrações das amostras utilizadas no teste podem ser vistos na Tabela 9.

Figura 18 - Gráfico da variância total acumulada dos 30 primeiros componentes do Teste 04.

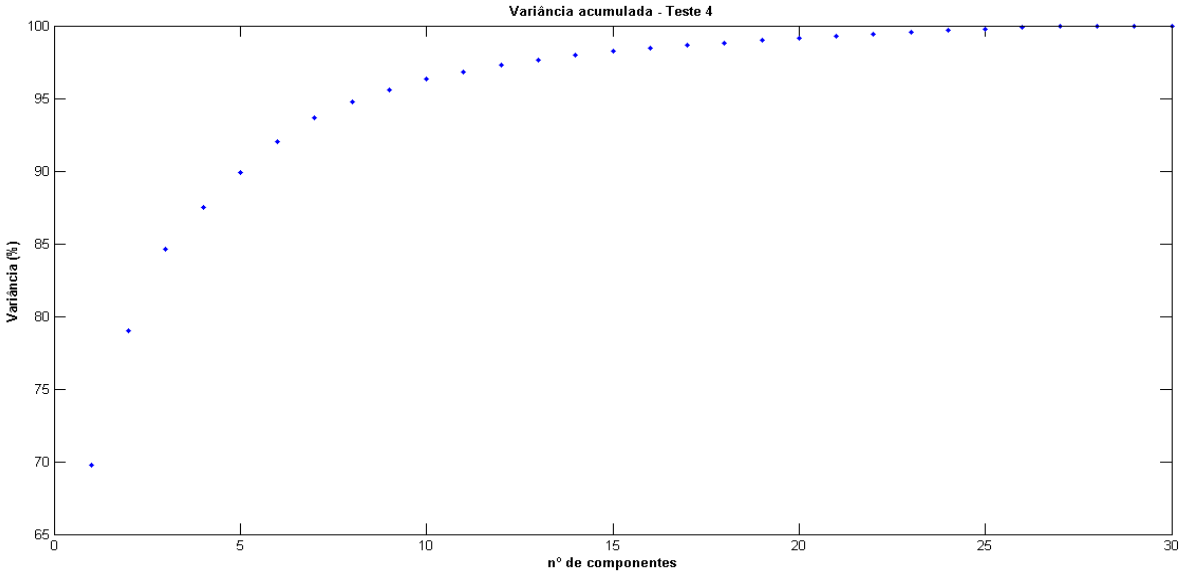


Tabela 9 - Resultados da validação cruzada da Matriz 04

Nitrato de sódio					Raminose				
Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Concentração real	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Concentração real
-2,8406E-15	0,21	0,25	0,23	0,0	-3,3203E-15	0,50	0,55	0,57	0,0
-2,4303E-15	0,14	0,17	0,17	0,0	-4,0835E-15	0,46	0,51	0,52	0,0
-9,2478E-15	0,05	0,07	0,05	0,0	-4,8902E-15	0,38	0,40	0,42	0,0
-9,2912E-15	-0,22	-0,19	-0,17	0,0	-1,5717E-15	0,23	0,27	0,23	0,0
0,50	0,33	0,34	0,34	0,5	0,50	0,33	0,35	0,35	0,5
0,50	0,25	0,28	0,28	0,5	0,50	0,32	0,36	0,36	0,5
0,50	0,29	0,31	0,31	0,5	0,50	0,31	0,34	0,34	0,5
0,50	0,06	0,08	0,09	0,5	0,50	0,20	0,24	0,22	0,5
0,50	0,29	0,03	0,03	0,5	1,00	1,42	1,06	1,04	1,0
0,50	0,94	1,00	0,98	0,5	1,00	0,83	0,92	0,96	1,0
0,50	0,13	0,16	0,18	0,5	1,00	0,29	0,34	0,32	1,0
0,50	-0,02	0,00	0,01	0,5	1,00	0,21	0,24	0,23	1,0
1,00	1,20	1,21	1,20	1,0	0,50	0,64	0,66	0,68	0,5
1,00	1,34	1,33	1,31	1,0	0,50	0,66	0,65	0,68	0,5
1,00	1,32	1,33	1,32	1,0	0,50	0,68	0,70	0,72	0,5
1,00	2,21	2,29	2,37	1,0	0,50	0,72	0,83	0,72	0,5
1,00	0,87	0,51	0,57	1,0	1,00	1,28	0,79	0,71	1,0
1,00	1,22	1,32	1,30	1,0	1,00	0,75	0,89	0,92	1,0
1,00	1,30	1,39	1,37	1,0	1,00	0,81	0,93	0,95	1,0
1,00	0,86	1,18	1,16	1,0	1,00	0,62	1,06	1,09	1,0
2,50	3,64	3,15	3,07	2,5	0,50	0,84	0,18	0,30	0,5
2,50	3,05	3,05	3,04	2,5	0,50	0,88	0,88	0,90	0,5
2,50	2,19	2,21	2,22	2,5	0,50	0,49	0,52	0,52	0,5
2,50	2,56	2,56	2,55	2,5	0,50	0,50	0,50	0,50	0,5
5,00	4,83	4,92	4,93	5,0	1,00	1,03	1,15	1,14	1,0
5,00	3,73	3,75	3,74	5,0	1,00	0,79	0,82	0,82	1,0
5,00	4,32	4,31	4,30	5,0	1,00	0,85	0,84	0,86	1,0
5,00	4,95	4,99	5,06	5,0	1,00	0,98	1,04	0,93	1,0

Comparando os resultados dos três procedimentos de regressão realizados, nota-se novamente o ajuste total das predições ao utilizar todos os dados originais na regressão, o que não ocorre com a predição da regressão por componentes principais utilizando 3 e 4 componentes. Como citado anteriormente, tal comportamento no resultado da predição é consequência da aplicação da validação cruzada, quando se utiliza exatamente os mesmos dados para a estimação dos parâmetros e para a validação do modelo. Nas etapas seguintes, utilizou-se o modelo de predição criado para validar outros conjuntos de dados.

Na tabela abaixo (Tabela 10) observa-se o resultado de um teste de validação de apenas 4 das mesmas amostras da Matriz 04, com a diferença que esta parte dos dados originais foi retirada do conjunto inicial, não participando da estimação de parâmetros e Análise de Componentes Principais, sendo reservada exclusivamente para validação do modelo de predição. Os dados reservados correspondem às amostras contendo 2,5 g/l de nitrato de sódio + 0,5 g/l de raminose. Assim, neste caso, a matriz original passou a ter 24 amostras, em vez de 28.

Tabela 10 - Comparação entre as predições das concentrações das amostras de validação da Matriz 04

Nitrato de sódio				Raminose			
Regressão com dados originais	Regressão com 3PC	Regressão com 4PC	Concentração real	Regressão com dados originais	Regressão com 3PC	Regressão com 4PC	Concentração real
3,572	3,031	3,116	2,500	0,943	0,889	0,911	0,500
3,280	3,018	3,113	2,500	0,840	0,931	0,955	0,500
2,812	2,236	2,284	2,500	0,746	0,513	0,525	0,500
3,251	2,573	2,630	2,500	0,834	0,555	0,570	0,500

Aqui, observa-se que já não houve uma predição exata, quando utilizado o modelo de predição sem tratamento por PCA. Ao contrário, a regressão com PCA apresentou melhores aproximações, com o detalhe que o modelo de regressão utilizando apenas 3 PC's apresentou resultados ligeiramente melhores que o modelo utilizando 4 PC's. A título de comparação, segue abaixo o resultado do teste de validação das mesmas amostras utilizando parte da Matriz 07:

Tabela 11 - Resultado da predição das mesmas amostras, na validação da Matriz 07 (segunda versão)

Nitrato de sódio				Raminose			
Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Concentração real	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Concentração real
2,202	1,785	1,957	2,500	0,794	0,239	0,260	0,500
1,602	1,579	1,753	2,500	-0,213	0,153	0,174	0,500
1,472	1,391	1,373	2,500	0,214	0,182	0,180	0,500
1,770	1,638	1,583	2,500	0,188	0,242	0,236	0,500

As mesmas amostras também foram utilizadas na calibração inicial da Matriz 04, fazendo parte da validação cruzada. Os resultados encontram-se em destaque, na cor vermelha, na Tabela 9 acima.

O mesmo modelo de predição inicial da Matriz 04, com 28 amostras ao total, foi utilizado para a tentativa de determinação de concentrações de outros conjuntos de dados, contendo amostras diferentes das utilizadas para a determinação dos modelos de predição por PCR e Regressão direta sem outras técnicas estatísticas.

O conjunto de dados de teste 05 consistiu de amostras de água, soluções de nitrato de sódio em diversas concentrações e soluções de raminose, também em diversas concentrações. Este conjunto de dados também foi submetido a predições para validação da calibração da Matriz 04,onde se pode ver os resultados abaixo (vale lembrar que as amostras da Matriz 04 eram misturas de nitrato de sódio com a raminose, enquanto neste conjunto de dados, os espectros são de soluções separadas).

A Tabela 12 apresenta o resultado da regressão total do conjunto de dados 05, utilizando a predição do conjunto de dados 04, com destaque às concentrações de nitrato de sódio a 2,5 g/l e raminose a 0,5 g/l. Não é possível chegar a conclusões óbvias através de simples observações desta regressão e comparando-a com a validação cruzada dos dados 04 e com a validação da Tabela 10 (onde estas concentrações foram reservadas do escopo de calibração, exclusivamente para validação). Ainda assim, nota-se nas Tabelas que não existe um padrão de comportamento ao longo de todos os valores de concentração preditos pelas regressões (o padrão de desvio nas determinações aparentemente variou em função da concentração testada). Isso nos estimula a examinar os subconjuntos dos dados testados para melhor compreensão dos resultados.

Tabela 12- Predições das concentrações das amostras da Matriz 05, utilizando a calibração da Matriz 04

Nitrato de sódio					Raminose				
Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Concentração real	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Concentração real
-2,8406E-15	0,21	0,25	0,23	0,0	-3,3203E-15	0,50	0,55	0,57	0,00
-2,4303E-15	0,14	0,17	0,17	0,0	-4,0835E-15	0,46	0,51	0,52	0,00
-9,2478E-15	0,05	0,07	0,05	0,0	-4,8902E-15	0,38	0,40	0,42	0,00
-9,2912E-15	-0,22	-0,19	-0,17	0,0	-1,5717E-15	0,23	0,27	0,23	0,00
0,62	0,73	0,76	0,75	0,5	0,34	0,51	0,55	0,56	0,00
0,47	0,57	0,59	0,57	0,5	0,31	0,46	0,48	0,49	0,00
0,43	0,54	0,55	0,53	0,5	0,32	0,43	0,44	0,46	0,00
0,88	0,53	0,58	0,59	0,5	0,60	0,33	0,40	0,39	0,00
1,19	0,82	0,81	0,81	1,0	0,46	0,21	0,19	0,19	0,00
1,30	0,85	0,85	0,86	1,0	0,59	0,28	0,28	0,27	0,00
1,53	0,95	0,95	0,96	1,0	0,66	0,29	0,29	0,27	0,00
1,40	0,87	0,85	0,87	1,0	0,65	0,30	0,28	0,26	0,00
2,93	2,67	2,68	2,67	2,0	0,67	0,68	0,69	0,70	0,00
3,13	2,29	2,26	2,27	2,0	0,88	0,40	0,37	0,36	0,00
2,94	2,29	2,28	2,28	2,0	0,78	0,47	0,44	0,44	0,00
3,27	2,61	2,62	2,63	2,0	0,81	0,52	0,53	0,51	0,00
2,80	2,50	2,50	2,49	2,5	0,62	0,53	0,53	0,55	0,00
3,37	2,88	2,86	2,86	2,5	0,73	0,59	0,56	0,56	0,00
3,65	3,46	3,52	3,52	2,5	0,95	0,88	0,96	0,96	0,00
3,54	2,69	2,67	2,68	2,5	1,02	0,51	0,48	0,47	0,00
7,78	6,95	6,84	6,82	5,0	2,13	1,56	1,42	1,45	0,00
8,57	7,32	7,25	7,24	5,0	1,55	1,07	0,97	0,97	0,00
8,06	6,84	6,74	6,73	5,0	1,36	0,90	0,76	0,78	0,00
8,27	7,21	6,90	6,89	5,0	2,08	1,57	1,15	1,16	0,00
20,11	16,62	16,56	16,63	10,0	4,35	2,76	2,68	2,59	0,00
19,50	15,72	15,59	15,65	10,0	3,98	2,11	1,93	1,84	0,00
18,83	15,89	15,73	15,77	10,0	4,15	2,53	2,31	2,26	0,00
22,62	19,44	19,40	19,55	10,0	4,50	2,75	2,70	2,48	0,00
0,10	0,64	0,70	0,73	0,0	0,18	0,59	0,67	0,63	0,25
0,52	0,92	0,86	0,86	0,0	1,90	1,44	1,36	1,35	0,25
0,22	-0,30	-0,25	-0,23	0,0	0,64	0,15	0,22	0,18	0,25
0,04	-0,41	-0,38	-0,36	0,0	0,58	0,14	0,18	0,15	0,25
-0,02	0,42	0,46	0,44	0,0	0,32	0,64	0,68	0,70	0,50
0,01	0,32	0,37	0,36	0,0	0,28	0,53	0,60	0,61	0,50
0,20	-0,02	0,04	0,05	0,0	0,39	0,32	0,39	0,38	0,50
-0,21	0,46	0,50	0,56	0,0	0,02	0,37	0,43	0,35	0,50
0,20	0,84	0,91	0,96	0,0	0,16	0,53	0,63	0,56	1,00
-0,08	0,06	0,09	0,08	0,0	0,25	0,40	0,44	0,45	1,00
-0,06	-0,43	-0,42	-0,41	0,0	0,50	0,16	0,18	0,16	1,00
0,11	-0,46	-0,43	-0,42	0,0	0,52	0,05	0,09	0,07	1,00
-0,01	0,93	0,83	0,79	0,0	0,91	1,34	1,21	1,27	2,00
0,14	0,58	0,64	0,64	0,0	0,26	0,61	0,69	0,68	2,00
-0,01	-0,19	-0,16	-0,15	0,0	0,42	0,21	0,25	0,25	2,00
0,06	-0,20	-0,16	-0,15	0,0	0,42	0,20	0,25	0,23	2,00

Tomando-se exclusivamente os dados referentes às mesmas concentrações nas Tabelas 9 a 12 (valores em destaque), é possível compará-los através de Análise Residual dos valores. Os resultados da soma dos quadrados dos resíduos e da média dos quadrados dos resíduos das respectivas concentrações destacadas podem ser vistos abaixo, na Tabela 13.

Tabela 13 - Comparativo da resumo da análise residual de todas as predições realizadas para as mesmas amostras (Nitrato de sodio 2,5 g/l e raminose 0,5 g/l)

Análise residual – Matriz 04 (destaque)								
	Nitrato de sódio				Raminose			
	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC
Soma Quad Res	0,000	1,689	0,809	0,690	0,000	0,258	0,251	0,203
Média Quad Res	0,000	0,422	0,202	0,173	0,000	0,065	0,063	0,051
Análise residual – Validação Matriz 04								
	Nitrato de sódio				Raminose			
	Regressão com dados originais		Regressão com 3PC	Regressão com 4PC	Regressão com dados originais		Regressão com 3PC	Regressão com 4PC
Soma Quad Res	2,420		0,625	0,819	0,483		0,340	0,381
Média Quad Res	0,605		0,156	0,205	0,121		0,085	0,095
Análise residual – Validação Matriz 05 (destaque)								
	Nitrato de sódio				Raminose			
	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC
Soma Quad Res	3,231	1,106	1,201	1,199	0,328	0,067	0,058	0,091
Média Quad Res	0,808	0,277	0,300	0,300	0,082	0,017	0,015	0,023
Análise residual – validação Matriz 07 (destaque)								
	Nitrato de sódio				Raminose			
	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com 4PC
Soma Quad Res	2,486	3,332	2,965	-	0,773	0,356	0,336	-
Média Quad Res	0,622	0,833	0,741	-	0,193	0,089	0,084	-

Sob esta análise, confrontando os resultados, consegue-se perceber melhor o efeito dos diferentes modelos de calibração sobre a determinação das concentrações das mesmas amostras. O melhor desempenho de predição, utilizando dados externos à calibração ficou por conta da Matriz 04 adaptada, tanto para o nitrato de sódio, quanto para a raminose.

4.2.2.3. Análise e validação do modelo de calibração da Matriz 08

A Matriz 08 foi composta basicamente por amostras de água e soluções simples de Nitrato de sódio e glicerol, num total de 30 amostras. A Tabela 14 abaixo apresenta as concentrações presentes na matriz. Na aplicação da Análise de Componentes Principais, obteve-se como resultado para melhor descrição dos

dados a utilização de 2 e 3 PC's. Para o teste de predição, utilizou-se dados externos à matriz de calibração, com o detalhe de que todas as amostras testadas na predição pertenciam a soluções de misturas, como pode ser visto na Tabela 15.

Tabela 14 - Amostras presentes na Matriz 08

Substância	Concentração
Água destilada	
Glicerol	2,5%
Glicerol	10,0%
Nitrato de sódio	2,5g/l
Nitrato de sódio	10,0g/l

Tabela 15 - Amostras de validação da Matriz 08

Substância	Concentração
Nitrato + Glicerol	10 g/l + 10%
Nitrato + Glicerol	5 g/l + 5%
Nitrato + Glicerol	2,5 g/l + 2,5%

Para este teste, também foi realizada a calibração sem utilizar as amostras de água, para avaliação do efeito da mesma sobre as predições. Os resultados das predições e da análise residual podem ser visualizados abaixo, nas Tabelas 16 e 17.

Tabela 16 - Resultados da predição utilizando a Matriz 08 (com uso da água e sem uso da água)

Utilizando o espectro da água na matriz de calibração							
Nitrato de sódio (g/l)				Glicerol (%)			
Regressão sem PCA	Regressão com 2PC	Regressão com 3PC	Concentração real	Regressão sem PCA	Regressão com 2PC	Regressão com 3PC	Concentração real
2,8414	2,95	3,01	2,5	1,8441	1,38	1,81	2,5
2,8934	3,01	3,09	2,5	1,8822	1,30	1,81	2,5
2,9589	3,09	3,16	2,5	1,8623	1,33	1,82	2,5
4,7339	4,97	5,04	5,0	3,3074	2,81	3,25	5,0
4,88	5,10	5,19	5,0	3,54	2,96	3,50	5,0
4,73	4,86	4,96	5,0	3,78	3,15	3,84	5,0
9,88	10,33	10,33	10,0	8,63	8,57	8,57	10,0
9,61	10,02	10,07	10,0	8,95	8,65	8,94	10,0
10,23	10,73	10,72	10,0	9,09	9,05	9,02	10,0
Sem o espectro da água na matriz de calibração							
Nitrato de sódio (g/l)				Glicerol (%)			
Regressão sem PCA	Regressão com 2PC	Regressão com 3PC	Concentração real	Regressão sem PCA	Regressão com 2PC	Regressão com 3PC	Concentração real
2,8738	2,81	2,91	2,5	1,7963	1,66	1,77	2,5
2,9487	2,84	2,99	2,5	1,8066	1,57	1,74	2,5
3,0001	2,93	3,05	2,5	1,8063	1,65	1,78	2,5
4,8453	4,78	4,92	5,0	3,1269	2,94	3,09	5,0
5,01	4,87	5,06	5,0	3,32	3,12	3,31	5,0
4,86	4,59	4,95	5,0	3,58	3,27	3,65	5,0
10,09	10,20	10,36	10,0	8,30	8,06	8,23	10,0
9,84	9,79	10,17	10,0	8,56	8,14	8,54	10,0
10,45	10,59	10,77	10,0	8,73	8,47	8,66	10,0

Tabela 17 - Análise residual das predições com a Matriz 08

Análise residual – Matriz 08 (com água)						
	Nitrato de sódio			Glicerol		
	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC
Soma Quad Res	8,554E-01	1,484	1,725	1,151E+01	21,206	12,217
Média Quad Res	9,504E-02	0,165	0,192	1,279E+00	2,356	1,357

Análise residual – Matriz 08 (sem água)						
	Nitrato de sódio			Glicerol		
	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC	Regressão com dados originais	Regressão com 2PC	Regressão com 3PC
Soma Quad Res	8,706E-01	1,058	1,469	1,636E+01	22,635	17,025
Média Quad Res	9,673E-02	0,118	0,163	1,818E+00	2,515	1,892

A Tabela 18 abaixo resume o desempenho de todos os testes de validação até então descritos aqui, em função dos respectivos Coeficientes de Determinação. Este coeficiente determina a qualidade de ajuste do modelo de regressão aos valores testados na predição, indicando o quanto o mesmo foi capaz de explicar os dados preditos.

Tabela 18 - Coeficientes de determinação dos testes de predição realizados

		Coeficientes de Determinação			
Dados de calibração	Dados de predição	Dados originais	2PC's	3PC's	4PC's
Matriz 04	Matriz 04	1	0,9038	0,9073	-
Matriz 04	Matriz 05	-0,2522	0,4136	0,4324	0,4204
Matriz 07	Matriz 07	1	0,9769	0,9799	-
Matriz 07 modificada	Dados externos	0,6278	0,5605	0,5993	-
Matriz 08	Dados externos	0,9293	0,8703	0,9203	-

É compreensível um coeficiente de determinação R^2 de 100% no modelo de regressão da validação cruzada com dados integrais, apresentando melhores resultados na regressão. Naturalmente, os modelos de regressão simplificados por PCA, utilizando apenas 3 e 4 PC's deixaram de fora uma boa carga de informações contidas no modelo e que, pelo visto, correspondem a *background* e ao espectro da água, uma vez que o erro relacionado às amostras com água pura (que deveriam retornar valor nulo) foi significativo. Um detalhe que reforça mais ainda esta observação é o fato de a análise residual da validação do teste com a Matriz 04 apresentar melhor desempenho para 3 PC's do que com 4 PC's, para ambas as substâncias (evidenciando o ruído contido nas PC's maiores). Outro fato interessante que reforça a hipótese sobre a fraqueza do sinal/atividade raman da

raminose em meio aquoso é o aumento do seu erro residual na predição das misturas (como se o sinal da raminose fosse mascarado pela água ou pelo nitrato, ou por ambos); nas soluções separadas, o erro é sensivelmente menor.

5 CONCLUSÕES

Diante do exposto neste trabalho, conclui-se a viabilidade de se determinar a presença e concentração de substâncias específicas em misturas, utilizando a associação entre a espectroscopia Raman e métodos de Estatística Multivariada, como a Análise de Componentes Principais e Regressão Multivariada. O fator dominante para o êxito nas predições e identificações das substâncias é a obtenção de amostras adequadas para a realização de um procedimento de calibração tal que tenha como resultado um modelo de regressão que se aproxime ao máximo das condições reais. A técnica de PCA mostrou-se eficiente como técnica de pré-processamento dos dados espectroscópicos, uma vez que permitiu a avaliação adequada da natureza das amostras, agrupando-as por similaridade, além de promover a redução do nível de ruído nos sinais. O método de Regressão multivariada também apresentou bons resultados na predição das concentrações das amostras testadas, apesar de alguns *outliers*, devido à natureza das amostras.

As amostras de grande diluição em água apresentam-se como o maior desafio para o sucesso do método, uma vez que o espectro da água pode vir a mascarar ou mesmo eliminar alguns sinais espectrais de algumas substâncias, a depender da concentração das mesmas. Como foi apresentado nos gráficos dos score plots, as amostras mais diluídas apresentam padrões muito próximos o que pode aumentar a margem de erro nos resultados. Ainda assim, a análise residual dos testes de validação demonstraram a viabilidade da técnica.

Como perspectivas futuras, torna-se imprescindível a realização de novos testes, com novas substâncias e novas misturas, inclusive com meios contendo microorganismos. O êxito em experimentos desta natureza garantiria a possibilidade de estudos quantitativos de diversas substâncias sem a necessidade de pré-tratamento das amostras.

REFERÊNCIAS

ABDEL-MAWGOUD, A. M. et al. Rhamnolipids : Detection, Analysis, Biosynthesis, Genetic Regulation and Bioengineering of Production. In: SOBERÓN-CHÁVEZ, G. (ed) Biosurfactants. **Microbiology Monographs**, vol 20. Springer, Berlin, Heidelberg, p. 13–56, 2011.

_____; LÉPINE, F.; DÉZIEL, E. Rhamnolipids : diversity of structures , microbial origins and roles. **Applied Microbiol Biotechnol**, n. 86, p. 1323–1336, 2010.

ADAM, C. D.; SHERRATT, S. L.; ZHOLOBENKO, V. L. Classification and individualisation of black ballpoint pen inks using principal component analysis of UV – vis absorption spectra. **Forensic Science International**, v. 174, p. 16–25, 2008.

ARRUDA, M. A. Z. et al. Análise exploratória em química analítica com emprego de quimiometria : pca e pca de imagens. **Revista Analytica**, v. 6, p. 38–50, 2003.

BASCHENKO, S. M.; MARCHENKO, L. S. On Raman spectra of water, its structure and dependence on temperature. **Semicond. Phys. Quantum Electron. Optoelectron.**, v. 14, n. 1, p. 77–79, 2011.

CADUSCH, P. J. et al. Improved methods for fluorescence background subtraction from Raman spectra. **Journal of Raman Spectroscopy**, v. 44, n. 11, p. 1587–1595, nov. 2013.

CENTENO DA ROSA, C. F. et al. Production of a rhamnolipid-type biosurfactant by Pseudomonas aeruginosa LBM10 grown on glycerol. **African Journal of Biotechnology**, v. 9, n. 53, p. 9012–9017, 2010.

CHRZANOWSKI, Ł.; ŁAWNICZAK, Ł.; CZACZYK, K. Why do microorganisms produce rhamnolipids ? **World J Microbiol Biotechnol**, 2011.

DA SILVA, J. G. B. **Aplicação da Análise de Componentes Principais (PCA) no Diagnóstico de Defeitos em Rolamentos através da Assinatura Elétrica de Motores de Indução**. 2008. 108 f. Dissertação (Mestrado) - Universidade Federal de Itajubá, Minas Gerais 2008.

ERIX, D. W. **Optical fingerprinting in medical microbiology ; Raman spectroscopy as a bacterial typing tool**. Rotterdam: Erasmus Universiteit Rotterdam, 2011.

ESBENSEN, K. H.; GELADI, P. Principal Component Analysis : Concept , Geometrical Interpretation , Mathematical Background , Algorithms , History , Practice. *In*: S, B.; R, T.; B, W. (Eds.). . **Comprehensive Chemometrics, Chemical and Biochemical Data Analysis**. Amsterdam: Elsevier, 2009. p. 211–226.

FERRARO, J. R.; NAKAMOTO, K.; BROWN, C. W. **Introductory Raman Spectroscopy**. 2. ed. San Diego: Elsevier, 2003. .

HIDALGO, B.; GOODMAN, M. Multivariate or multivariable regression? **American Journal of Public Health**, v. 103, n. 1, p. 39–40, 2013.

IM, H. et al. Classification of materials for explosives from prompt gamma spectra by using principal component analysis. **Applied Radiation and Isotopes**, v. 67, n. 7–8, p. 1458–1462, 2009.

JACOBSSON, P.; JOHANSSON, P. Measurement Methods | Vibrational Properties: Raman and Infra Red. **Encyclopedia of Electrochemical Power Sources**, p. 802–812, 2009.

KANG, J. et al. The effect of aqueous solution in Raman spectroscopy. **Medicine**, v. 7519, n. Pibm, p. 1–7, 2009.

KEITHLEY, R. B.; HEIEN, M. L.; WIGHTMAN, R. M. Multivariate concentration determination using principal component regression with residual analysis. **Trends in Analytical Chemistry**, v. 28, n. 9, p. 1127–1136, 1 out. 2009.

KOCH, I. **Analysis of Multivariate and High-Dimensional Data**. 1. ed. New York: Cambridge University Press, 2013.

KRAMER, R. **Chemometric Techniques for Quantitative Analysis**. New York: Marcel Dekker, 1998.

LEITERMANN, F.; SYLDATK, C.; HAUSMANN, R. Fast quantitative determination of microbial rhamnolipids from cultivation broths by ATR-FTIR Spectroscopy. **Journal of Biological Engineering**, v. 2, n. 13, p. 1–8, 2008.

MARCELO, M. C. A. et al. Profiling cocaine by ATR-FTIR. **Forensic Science International**, v. 246, p. 65–71, 2015.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada:** Uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005. 297 p.

NOGUEIRA, F. E. **Modelos de regressão multivariada**. 2007. 95 f. Dissertação (Mestrado) - Universidade de São Paulo, São Paulo, 2007.

PATEL, R. M.; DESAI, A. J. Biosurfactant production by Pseudomonas aeruginosa GS3 from molasses. **Letters in Applied Microbiology**, p. 91–94, 1997.

RAMOS, L. S. et al. Chemometrics. **Anal.Chem.**, v. 58, n. 5, p. 294–315, 1986.

RANDHAWA, H. S. Raman Spectroscopy. In: _____ **Modern Molecular Spectroscopy**. Delhi: Macmillan India Ltd., 2003. cap. 4, p. 284-326.

RIBEIRO, R. M. **Infravermelho e PCA na Análise da Natureza Química do Carbono em Diferentes Culturas**. 2012. 64 f. Tese (Doutorado) - Universidade Federal de Lavras, Lavras, 2012.

RYDER, A. G.; CONNOR, G. M. O.; GLYNN, T. J. Quantitative analysis of cocaine in solid mixtures using Raman spectroscopy and chemometric methods. **Journal of Raman Spectroscopy**, v. 31, n. 3, p. 221–227, 2000.

SALA, O. **Fundamentos da Espectroscopia Raman e no Infravermelho**. 2. ed. São Paulo, SP: Editora UNESP, 2008.

SILVA, S. N. R. L. et al. Glycerol as substrate for the production of biosurfactant by Pseudomonas aeruginosa UCP0992. **Colloids and Surfaces B: Biointerfaces**, v. 79, n. 1, p. 174–183, 2010.

SMITH, E.; DENT, G. **Modern Raman Spectroscopy - A Practical Approach**. Chichester, UK: John Wiley & Sons, 2005.

SWINEHART, D. F. The Beer-Lambert Law. **Journal of Chemical Education**, v. 39, n. 7, p. 333-335, 1962.

WANG, Q. *et al.* Engineering Bacteria for Production of Rhamnolipid as an Agent for Enhanced Oil Recovery. **Biotechnology and Bioengineering**, v. 98, n. 4, p. 842–853, 2007.

WOLD, S.; ESBENSEN, K. I. M.; GELADI, P. Principal Component Analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, p. 37–52, 1987.

XU, L. et al. The feasibility of using near-infrared spectroscopy and chemometrics for untargeted detection of protein adulteration in yogurt: Removing unwanted variations in pure yogurt. **Journal of Analytical Methods in Chemistry**, v. 2013, 2013.