Revista Virtual de Química
ISSN 1984-6835

Artigo

Uma Proposta Didática no Ensino de Análise Exploratória de Dados com Imagens de MDF (*Medium-Density Fiberboard*)

Böck, F. C.;* Assmann, D.; Helfer, G. A.; Costa, A. B.

Rev. Virtual Quim., **2015**, *7* (6), 2475-2486. Data de publicação na Web: 23 de setembro de 2015

http://www.uff.br/rvq

A Didactic Proposal in Teaching Exploratory Data Analysis with MDF Pictures

Abstract: This study presents a practical activity adopted in the classroom to aid in comprehension of multivariate tools, particularly in Image principal component analysis (Image PCA), through reproduction of the scores plotted with printed images samples of Medium-Density Fiberboard (MDF). This pedagogical approach allows a more realistic view of the groups formed in the chart, and shows how this technique can be explored in a dynamic manner, allowing a better learning and understanding of the subject. Besides that, this study discloses, among academics, ChemoStat software®, a free software for PCA.

Keywords: PCA; Images; Multivariate Analysis.

Resumo

Este estudo apresenta uma atividade prática adotada em sala de aula para auxiliar na compreensão das ferramentas multivariadas, particularmente na análise de componentes principais de imagens (PCA de imagens), através da reprodução do gráfico de escores com imagens impressas de amostras de *Medium-Density Fiberboard* (MDF). Esta abordagem pedagógica possibilita uma visualização mais realista dos agrupamentos formados no gráfico e mostra como essa técnica pode ser explorada de uma forma dinâmica, permitindo uma melhor aprendizagem e fixação do conhecimento. Além disso, este trabalho divulga, entre os acadêmicos, o software ChemoStat®, um software gratuito apto à PCA de imagens.

Palavras-chave: PCA; Imagens; Análise multivariada.

* Universidade de Santa Cruz do Sul, Programa de Pós-graduação em Sistemas e Processos Industriais, Campus Santa Cruz do Sul, CEP 96815-900, Santa Cruz do Sul-RS, Brasil.
fernanda.c.bock@gmail.com
DOI: 10.5935/1984-6835.20150147

Volume 7, Número 6 Novembro-Dezembro 2015

Revista Virtual de Química
ISSN 1984-6835

Uma Proposta Didática no Ensino de Análise Exploratória de Dados com Imagens de MDF (*Medium-Density Fiberboard*)

Fernanda Carla Böck,[a] Daniel Assmann,[a] Gilson Augusto Helfer,[a] Adilson Ben da Costa[a,b]

[a] Universidade de Santa Cruz do Sul, Programa de Pós-graduação em Sistemas e Processos Industriais, Campus Santa Cruz do Sul, CEP 96815-900, Santa Cruz do Sul-RS, Brasil.

[b] Universidade de Santa Cruz do Sul, Departamento de Biologia e Farmácia, Campus Santa Cruz do Sul, CEP 96815-900, Santa Cruz do Sul-RS, Brasil.

* fernanda.c.bock@gmail.com

Recebido em 29 de maio de 2015. Aceito para publicação em 16 de setembro de 2015

1. Introdução

A química é uma ciência com um elevado grau de abstração, o que muitas vezes leva a certa dificuldade em sua aprendizagem, principalmente quando apresentada exclusivamente de forma teórica.[1] Durante muito tempo responsabilizou-se exclusivamente os estudantes sobre o seu sucesso na aprendizagem, acreditando que ela ocorria através da repetição, de forma que os estudantes que não eram capazes de aprender eram os únicos responsáveis pelo seu insucesso. Atualmente o sucesso na busca da aprendizagem é um desafio para o professor, que precisa buscar novos métodos para a motivação dos estudantes na busca do conhecimento.[2]

Diversas são as estratégias utilizadas para motivar e transmitir o conhecimento para os estudantes de química, dentre as quais podemos citar atividades como jogos,[3,4] olimpíadas[5] e "*shows* de mágica"[6] que têm possibilitado uma aprendizagem mais prazerosa aos estudantes.

Na área de quimiometria, a utilização de muitas ferramentas matemáticas pode dificultar, ao menos inicialmente, o interesse dos estudantes no seu estudo, de forma que a exploração de atividades mais participativas que compreendam desde a aquisição de dados até a interpretação final dos resultados deve ser fomentada.

1.1. Análise exploratória de dados

A análise exploratória de dados compreende um conjunto de técnicas matemáticas que permitem o processamento de dados com diversas variáveis simultaneamente, sendo possível visualizar o agrupamento das amostras conforme sua semelhança, e também identificar a influência das variáveis sobre as amostras.[7]

Dentre as técnicas mais utilizadas pode ser citada a Análise por Componentes Principais, do inglês *Principal Components Analysis* (PCA). A PCA tem como objetivo reduzir a dimensionalidade do conjunto de dados original em um novo sistema de eixos, denominados componentes principais, do inglês *Principal Components* (PC), permitindo a visualização da natureza multivariada dos dados em poucas dimensões, preservando assim a maior quantidade de informação possível. A PCA normalmente é utilizada com objetivo de visualizar a estrutura dos dados, encontrar similaridade entre amostras, e detectar amostras anômalas (*outliers*).[8,9]

Com a redução de dimensionalidade proporcionada pela PCA, as amostras passam a ser pontos localizados em espaços de dimensões reduzidas definidos pelos PCs, por exemplo, bi- ou tridimensionais. A matriz X é decomposta em um produto de duas matrizes denominadas escores (*scores*) e pesos (*loadings*). Os escores representam as coordenadas das amostras no sistema de eixos formados pelas PCs, cada PC é constituída pela combinação linear das variáveis originais, e os coeficientes dessas combinações são denominados pesos. Os pesos são os cossenos dos ângulos entre as variáveis originais e as PCs, representando o quanto cada variável contribui para cada PC.[10,11]

A primeira componente principal (PC1) é traçada no sentido da maior variação no conjunto de dados, a segunda (PC2) é traçada ortogonalmente à primeira, tendo o objetivo de descrever a maior porcentagem da variação que não foi explicada pela PC1 e assim sucessivamente com as outras PCs. Os escores representam as relações de similaridade entre as amostras, já avaliação dos pesos permite entender quais variáveis mais contribuem para os agrupamentos observados no gráfico dos escores. Através da análise conjunta dos gráficos, é possível verificar quais são as variáveis responsáveis pelas diferenças observadas entre as amostras. O número de componentes principais a ser utilizado no modelo PCA é determinado pela porcentagem de variância explicada. Assim, seleciona-se um número de componentes de tal maneira que a maior porcentagem da variação presente no conjunto de dados originais seja capturada.[12]

Outra técnica que vem se agregando à análise exploratória é a análise de componentes principais de imagens. A PCA de imagens é muito útil quando existe um elevado número de imagens a serem processadas. Essa técnica gera uma matriz de dados a partir das informações contidas nas imagens, onde cada pixel da imagem pode ser considerado como um objeto da matriz.[13]

1.2. Softwares

Existem diversos softwares disponíveis para análise exploratória de dados. Grande parte deles necessita de uma licença comercial, e na maioria das vezes paga, o que muitas vezes dificulta o seu uso em sala de aula, porém alguns softwares possuem licenças livres. Dentre os softwares livres podemos citar o Chemoface®, criado pela Universidade Federal de Lavras (UFLA), o Past® da *Univesity of Olso*, e o ChemoStat® desenvolvido pelo Programa de Pós-

Graduação em Sistemas e Processos Industriais da UNISC.[14]

Todos os softwares citados acima compreendem a análise exploratória de dados, porém apenas o ChemoStat permite a análise de componentes principais de imagens.

2. Objetivo

Este trabalho visa contribuir no ensino de análise multivariada de imagens, mostrando como essa técnica pode ser explorada de uma forma dinâmica, permitindo uma melhor aprendizagem e fixação do conhecimento. Além disto, este trabalho divulga, entre os acadêmicos, o software ChemoStat®[14], um dos raros softwares gratuitos aptos à PCA de imagens.

3. Metodologia

3.1. Materiais utilizados

As imagens utilizadas neste trabalho foram adquiridas, em triplicata, de 57 amostras de placas de MDF (do inglês, Medium-Density Fiberboard), utilizando o escâner de uma impressora multifuncional HP (modelo PSC 1350).

As amostras de MDF foram obtidas de diferentes fornecedores, 5 amostras fornecidas pela NR Painel Canaletado, 5 da marca Madelei e 47 da marca Trupan. As amostras estão listadas no Anexo 1.

3.2. Aquisição das imagens

Para a aquisição das imagens foi instalada uma máscara de cartolina branca, com uma abertura central de 6 x 9,5 cm, na superfície de digitalização do escâner, permitindo a captura de imagens de tamanhos idênticos. A Figura 1 representa os passos para a aquisição das imagens.

3.3. Análise de componentes principais

Após a aquisição das imagens, estas foram exportadas para o software ChemoStat®[14], para gerar o gráfico dos *scores* e *loadings*, e efetuar a análise dos componentes principais das imagens. Para isso, sete etapas foram realizadas (Figura 2).

A primeira etapa foi a de *"Importar a imagem escaneada"* para o software ChemoStat®, seguido da etapa de seleção da região de interesse, do inglês *region of interest* (ROI). A seleção da ROI deve ser realizada para adquirir imagens com o mesmo número de pixels, obtendo assim um padrão de imagens para a PCA.

Para determinar os parâmetros da ROI, selecionou-se a opção "*tools*" na barra de ferramentas do software, seguida da opção "*image cutter*", onde uma nova janela é aberta para a seleção dos parâmetros. Neste trabalho os parâmetros escolhidos para a seleção da ROI foram *"pixel central"* e o tamanho de 260 x 260 pixels, após a escolha dos parâmetros seleciona-se a opção *"cut & save"* e então as imagens geradas podem ser salvas.

Com todas as imagens salvas e padronizadas, passou-se para a terceira etapa, de *"importar as imagens da ROI"*. O novo conjunto de amostras é importado selecionando-se a opção "*add image file*". Então o software habilita a próxima etapa, "*selecionar o modelo de segregar o pixel*", permitindo a escolha de dois tipos de segregação para a criação da matriz, que são "*histogram*" e "*color model*".

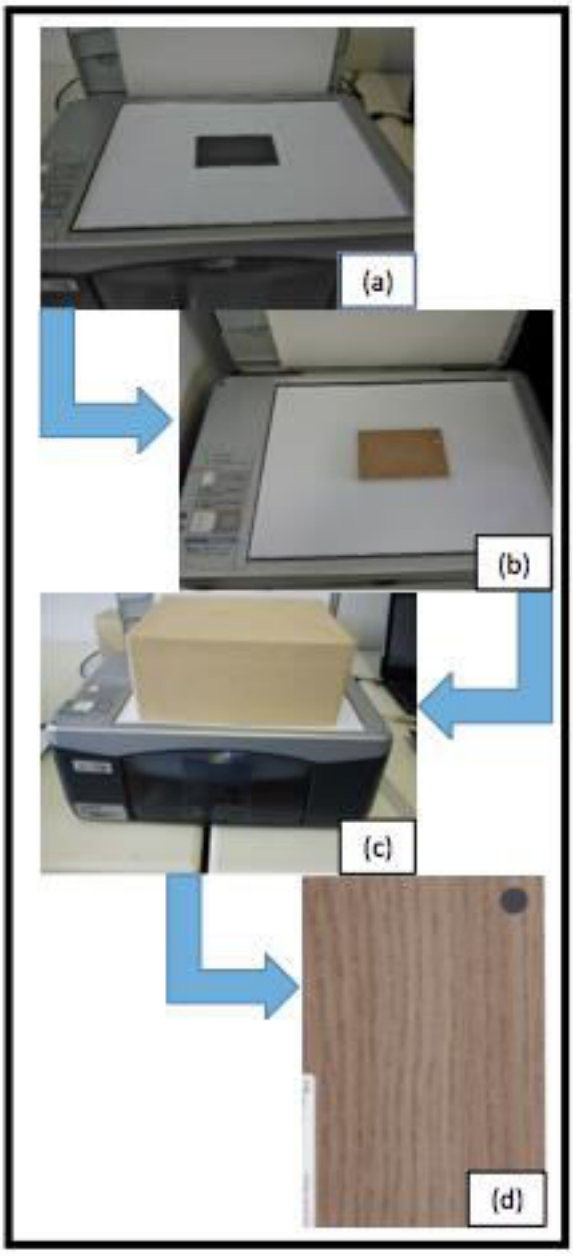

Figura 1. Esquema de aquisição de imagens com escâner para MDF. a) impressora multifuncional e papel-máscara, b) representação da posição da amostra, c) tampa para proteção de luz externa e d) imagem adquirida

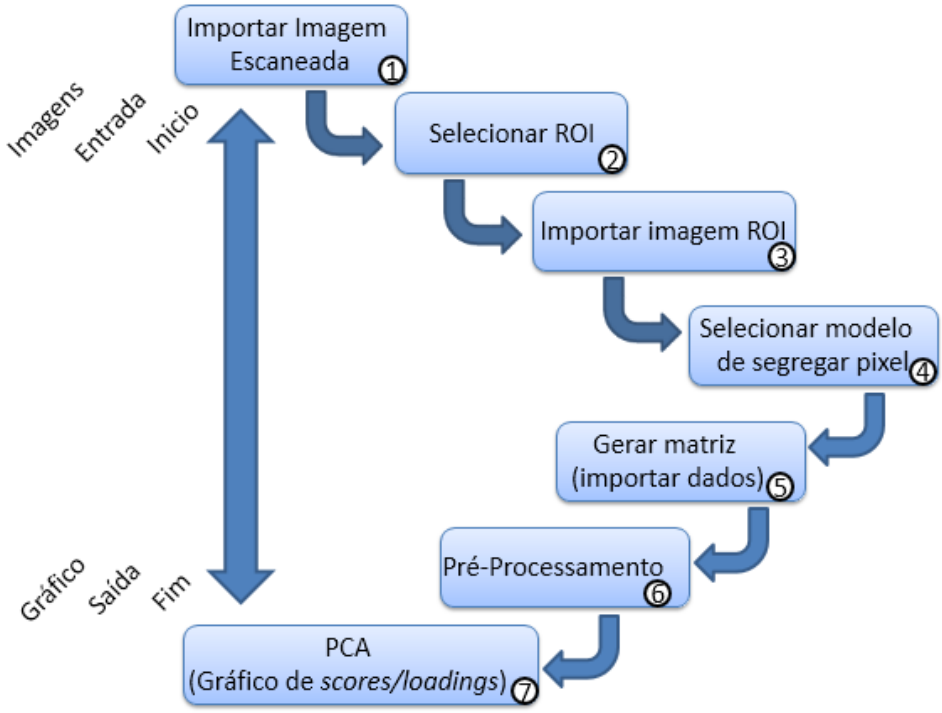

Figura 2. Esquema das etapas para realizar a análise de componentes principais

A opção "*histogram*" realiza uma distribuição de frequência baseada em 256 tons para cada componente de cor RGB (do inglês *red, green and blue*). A opção "*color model*" permite extrair informações dos modelos de cor RGB, e RGB relativo, definidos como "r%", "g%" e "b%", HSV (do inglês *hue*, *saturation* e *brightness)*, incluindo ainda as informações de intensidade ("I") e luminânica ("L"), quando selecionados. É possível optar por um ou mais componentes, separadamente.

Os dados extraídos pixel a pixel podem ser agrupados numa média, selecionando a opção "*average*", ou em uma mediana, com a opção "*median*".[15] O modelo de segregação de pixel utilizado nesse trabalho foi considerando os parâmetros HSVLI e a opção de agrupamento selecionada foi "*average*".

Selecionado o modelo de segregação de pixel, a matriz de dados é produzida na utilizando a função "*import data*". Antes de gerar os gráficos dos *scores* e *loadings*, executou-se a etapa de "*pré-processamento*". Acionando o botão direito do *mouse*, sobre a matriz, e na opção PCA, estão disponíveis dois tipos de pré-processamento, o *"meancenter"* e *"autoescale"*. Para este trabalho foi utilizado o pré-processamento "*autoescale*", no qual os dados são centrados na média e divididos pelo desvio padrão.[13]

Ao final deste procedimento são apresentados os gráficos de *scores* e *loadings* das amostras de imagens.

Para auxiliar na interpretação dos resultados exibidos, uma atividade lúdica foi desenvolvida com os estudantes. Para isso, as imagens das ROI foram impressas, e de posse as coordenadas (x, y) as imagens foram distribuídas sobre uma classe, em que os eixos de PC1 (eixo x) e PC2 (eixo y) estavam demarcados com uma fita. Os materiais necessários para essa representação estão apresentados na Figura 3.

Figura 3. Materiais necessários para a representação do gráfico de *scores*

4. Resultados e Discussão

A PCA executada a partir do modelo de segregação HSVLI mostrou que as duas primeiras componentes principais explicam 93,03% da variância dos dados, conforme apresentado na Tabela 1.

Analisando os gráficos dos *scores* da PC1 x PC2 com os dados autoescalados representados na Figura 4, é possível visualizar que ocorre uma separação das amostras pela PC1. Do lado positivo da PC1 também pode-se observar que ocorre uma separação das amostras pela PC2, formando dois grupos, um do lado negativo da PC2 e um do lado positivo.

Uma melhor interpretação dos resultados é possível através do gráfico de *scores* elaborado pelos estudantes com os recortes das imagens, Figura 5.

Tabela 1. Valores de variância e variância acumulada das componentes principais

PC	Variância (%)	Variância acumulada (%)
1	66,80	66,80
2	26,23	93,03
3	6,90	99,93
4	0,08	100

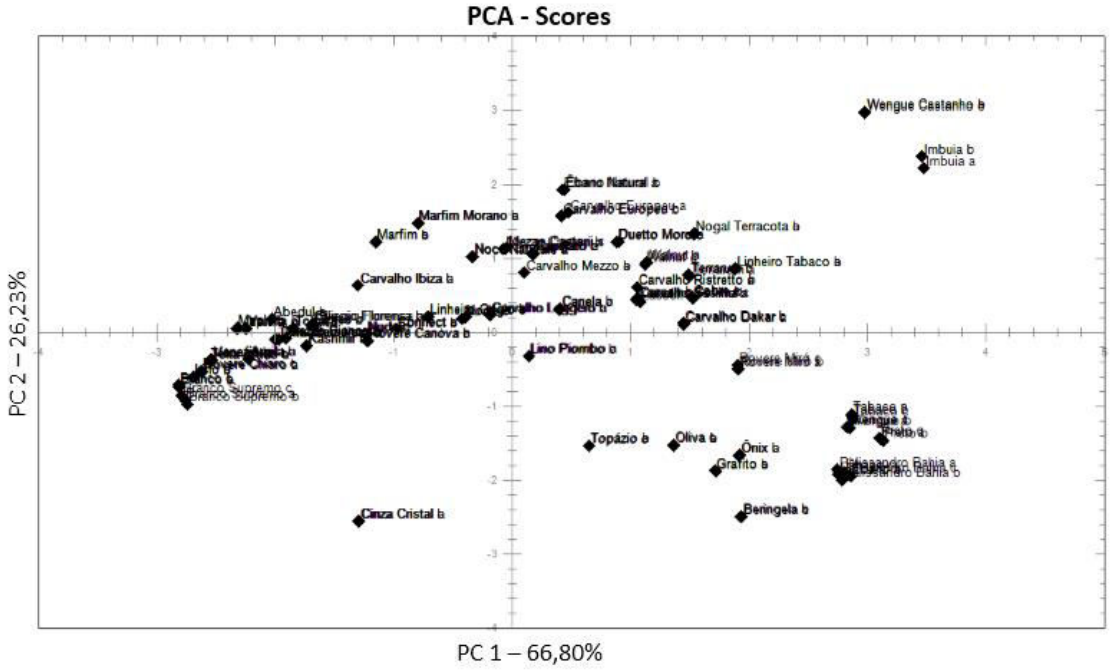

Figura 4. Gráfico de *scores* PC1 x PC2

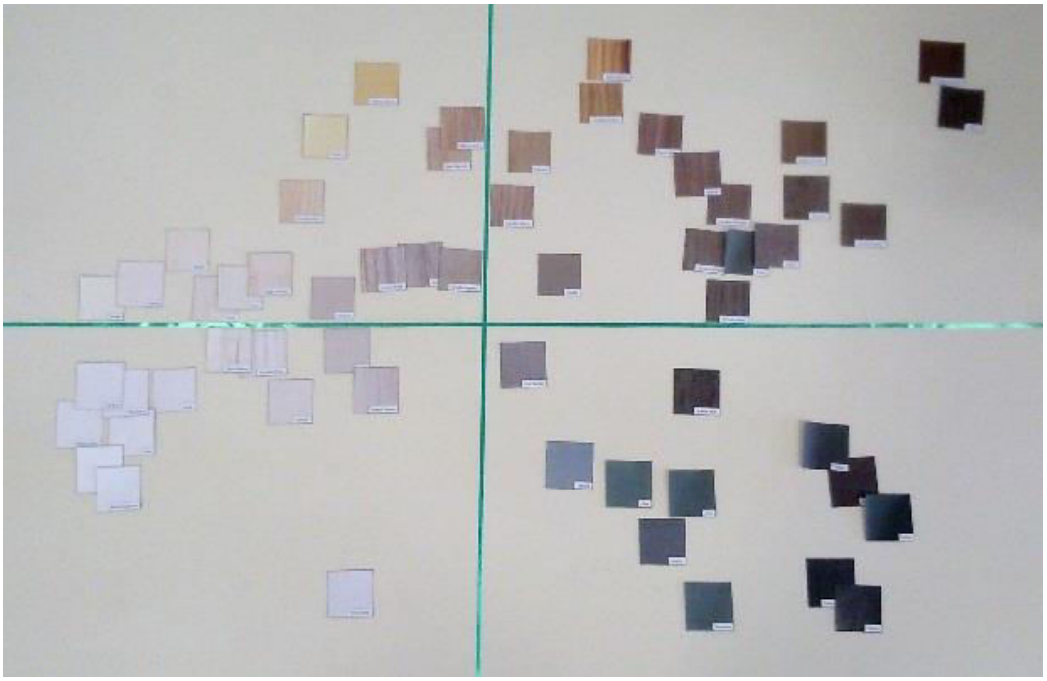

Figura 5. Representação com imagens do gráfico de *scores* da PC1 x PC2

Através dessa representação é possível visualizar que a componente principal 1 separa as amostras com tons mais claros (lado negativo) daquelas com tons mais escuros (lado positivo), e a componente principal 2 aparentemente separa as amostras que possuem uma tonalidade mais sólida (lado negativo), das amostras que possuem detalhes em outras cores (lado positivo).

Nos gráficos de pesos (*loadings)* podem

ser observadas as variáveis que possuem maior influência na separação das amostras.

Na Figura 6 é possível verificar que as variáveis com maior influência na separação das amostras na PC1 são V, L e I, que influenciam a separação para o lado negativo da PC. Do lado positivo com menor influência estão as variáveis H e S.

Na Figura 7 pode-se observar a influência das variáveis na separação das amostras pela PC2.

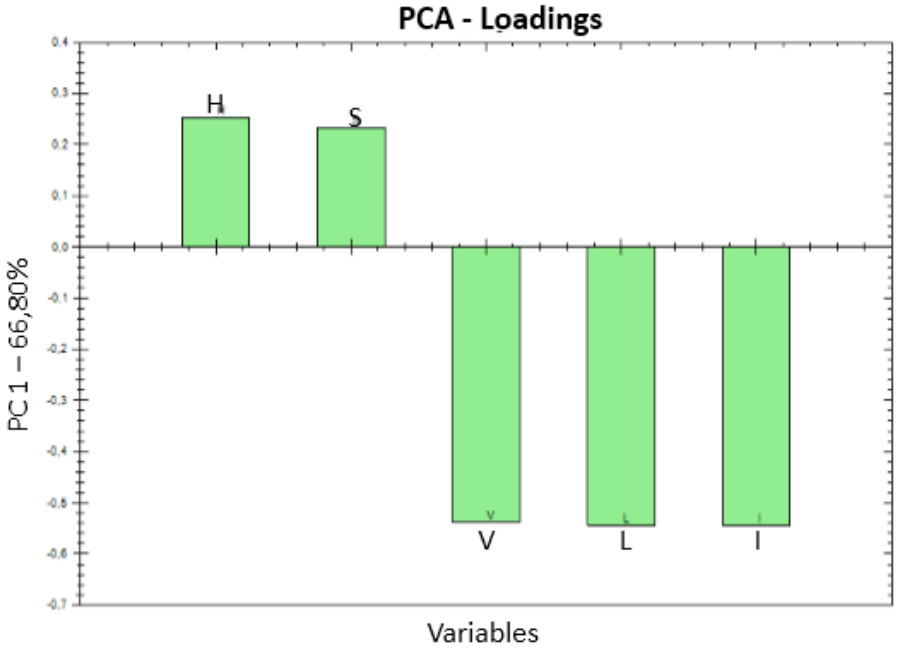

Figura 6. Gráfico de pesos (*loadings)* da PC1

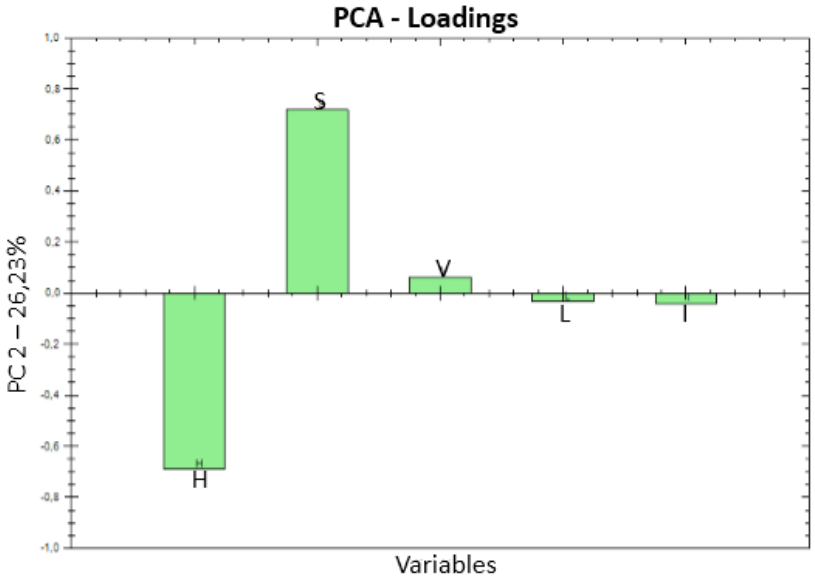

Figura 7. Gráfico de pesos (*loadings*) da PC2

Na Figura 7 observa-se que duas variáveis têm maior influência na separação, a variável H para o lado negativo, e a S para o lado positivo. As variáveis V (lado positivo) e L, I (lado negativo) não apresentam muita influência na separação das amostras pela PC2.

Por fim, cabe destacar que a atividade de montagem manual dos gráficos de scores permitiu aos estudantes a discussão das propriedades de cada imagem de MDF. Auxiliando a compreensão das propriedades de cada material responsável pelo agrupamento ou separação destes por similaridade.

5. Considerações finais

Este estudo possibilitou uma melhor compreensão da técnica de análise multivariada de imagens, através da reprodução do gráfico de escores com imagens impressas das amostras. Essa atividade permitiu uma visualização mais realista dos agrupamentos formados no gráfico, identificando semelhanças das amostras em relação aos agrupamentos formados, contribuindo assim para uma melhor aprendizagem e fixação do conhecimento sobre técnicas de análise multivariada de imagens bem como da utilização do software Chemostat®.

Referências Bibliográficas

[1] Mota, T. C.; Cleophas, M. G. Proposta para o ensino de química utilizando a planta Pterodon abruptus (Moric.) Benth. como indicador natural de pH. *Revista Virtual de Química* **2014**, *6*, 1353. [CrossRef]

[2] Cunha, M. B. Jogos no Ensino de Química: Considerações Teóricas para sua Utilização em Sala de Aula. *Química Nova na Escola* **2012**, *34*, 92. [Link]

[3] Massena, E. P.; Guzzi Filho, N. J.; Sá, L. P. Produção de casos para o ensino de química: uma experiência na formação inicial dos professores. *Química Nova* **2013**, *36*, 1066. [CrossRef]

[4] Osorio, V. K. L.; Kuya, M. K.; Maia, A. S. Oliveira, W. Um experimento-charada usando data-show e resinas de troca iônica. *Química Nova* **2003**, *26*, 960. [CrossRef]

[5] Mangrich, A. página SBQ. Disponível em: <http://www.sbq.org.br/portal2/olimpiadas/olimpiadas.htm>. Acesso em: 5 maio 2015.

[6] Arroio, A.; Honório, K. M.; Weber, K. C.; Homem-de-Mello, P.; Gambardella, M. T. P.; Silva, A. B. F. O show da Química: motivando o interesse científico. *Química Nova* **2006**, *29*, 173. [CrossRef].

[7] Hou, S.; Wentzell, P. D. Re-centered kurtosis as a projection pursuit index for multivariate data analysis. *Journal of Chemometrics* **2014**, *28*, 370. [CrossRef]

[8] Chandrasekaran, A.; Ravisankar, R.; Harikrishnan, N.; Satapathy, K. K.; Prasad, M. V. R.; Kanagasabapathy, K. V. Multivariate statistical analysis of heavy metal concentration in soils of Yelagiri Hills, Tamilnadu, India – Spectroscopical approach. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2015**, *137*, 589. [CrossRef] [PubMed]

[9] Parra, S.; Bravo, M. A.; Quiroz, W.; Moreno, T.; Karahasiou, A.; Font, O.; Vidal, V.; Cereceda-Balic, F. Source apportionment for contaminated soils using multivariate statistical methods. *Chemometrics and Intelligent Laboratory Systems* **2014**, *138*, 127. [CrossRef]

[10] Corvucci, F.; Nobili, L.; Melucci, D.; Grillenzoni, F. V. The discrimination of honey origin using melissopalynology and Raman spectroscopy techniques coupled with multivariate analysis. *Food Chemistry* **2015**, *169*, 297. [CrossRef] [PubMed]

[11] Karpushkin, E.; Bogomolov, A. Morphology assessment of poly (2-hydroxyethyl methacrylate) hydrogels using multivariate analysis of viscoelastic and swelling properties. *Polymer* **2015**, *58*, 222. [CrossRef]

[12] de Souza, A. M.; Poppi, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e

análise de componentes principais: um tutorial, parte 1. *Química Nova* **2012**, *35*, 223. [CrossRef]

[13] Matos, G. D.; Pereira-Filho, E. R.; Poppi, R. J.; Arruda, M. A. Z. Análise exploratória em química analítica com emprego de quimiometria: PCA e PCA de imagens. *Revista Analytica* **2003**, *6*, 38.

[14] Helfer, G. A.; Bock, F.; Marder, L.; Furtado, J. C.; Costa, A. B.; Ferrão, M. F. Chemostat, um software gratuito para análise exploratória de dados multivariados. *Química Nova* **2015**, *38*, 575. [CrossRef]

[15] Helfer, G. A.; *Dissertação de Mestrado*, Universidade de Santa Cruz do Sul, 2014. [Link]

Anexo 1. Relação de amostras utilizadas no estudo

Amostra	Nome	Amostra	Nome	Amostra	Nome	Amostra	Nome
	#Beringela		*Canelato		*Grafito		*Nude
	#Cipres		*Carvalho Dakar		*Imbuia		*Palissandro Bahia
	#Cobre		*Carvalho Europeu		*Kashmir		*Rovere Canova
	#Oliva		*Carvalho Ibiza		*Lassen		*Rovere Chiaro
	#Ônix		*Carvalho Leggero		*Linheiro Grigio		*Rovere Miró
	+Branco		*Carvalho Mezzo		*Linheiro Tabaco		*Teka Ártico
	+Maple		*Carvalho Ristretto		*Lino		*Teka Barcelona
	+Marfim		*Carvalho Sevilha		*Lino Piombo		*Terrarum
	+Preto		*Ciliegio Florença		*Marfim Morano		*Tokai
	+Tabaco		*Cinza Cristal		*Mezzo Bianco		*Topázio
	*Abedul		*Connect		*Mezzo Castani		*Vanilla
	*Argila		*Duetto Moro		*Noce Naturale		*Venezia
	*Branco Supremo		*Ébano		*Nodo		*Walnut
	*Canela		*Ébano Natural		*Nogal Terracota		*Wengue
							*Wengue Castanho

Fornecedores: # Madelei; +NR Painel Canaletado; *Trupan.