Desenvolvimento e validação de modelos preditivos baseados em espectrômetro NIR portátil e machine learning para previsão da composição de larvas de mosca soldado negro (*Hermetia illucens*)

Palavras-Chave: quimiometria, inseto comestível, previsão, não destrutivo

Autores:
LUIS JAM PIER CRUZ TIRADO, FEA – UNICAMP
MATHEUS SILVA DOS SANTOS VIEIRA, FEA – UNICAMP
JOSÉ MANUEL AMIGO, Universidade do País Basco (UPV/EHU)
RAÚL SICHE, Universidade Nacional de Trujillo (UNT)
Prof(ª). Dr(ª). DOUGLAS FERNANDES BARBIN (orientador), FEA – UNICAMP

1. INTRODUÇÃO

Com o aumento da população mundial, novos tipos de alimentos terão um papel importante na garantia da segurança alimentar dessas pessoas. É estimado que, em 30 anos, haja cerca de 9,7 bilhões de pessoas vivendo em nosso planeta, o que levará a um aumento de 25 a 70% da demanda por alimentos. Tendo isso em mente, os insetos se tornarão uma fonte muito importante de proteínas, tanto para animais, quanto para os seres humanos (VARUNJIKAR et al., 2022). Além dessa importância, o consumo de insetos traz três principais benefícios para as pessoas e para a sociedade: (1) eles possuem um alto valor nutricional por serem fonte de proteínas, ácidos graxos essenciais e alguns minerais; (2) são benéficos para o meio ambiente por emitirem menos gases do efeito estufa e amônia; e (3) melhoram fatores sociais por contribuírem com a segurança alimentar (MUREFU et al., 2019).

Dentre estes insetos, as larvas da mosca soldado negro (black soldier fly – BSF) chamam atenção por serem ricas em nutrientes e por colaborarem com a redução de lixo orgânico. As larvas da BSF são ricas em proteínas, lipídios, minerais e quitina, podendo ter diversas aplicações, como na produção de alimentos proteicos para peixes, aves, suínos e animais de estimação; na produção de biodiesel; e, devido à quitina, podem ser usadas nas indústrias alimentícias, de cosméticos, farmacêuticas, entre outras (CALIGIANI et al., 2018)

O tempo de desenvolvimento da larva até alcançar o estágio de abate (pré-pupa) é curto (cerca de 14 dias) e o acompanhamento da qualidade nutricional da larva pode ser um desafio para a nascente e pouco desenvolvida indústria de insetos no Brasil. Isso acontece, porque os métodos tradicionais de medição de proteína e gordura, ainda que bastante precisos e bem estabelecidos na legislação, consomem reagentes químicos, são demorados, precisam de alto conhecimento técnico e destroem as amostras. Neste contexto, a espectroscopia do infravermelho próximo (NIRS – *Near Infrared Spectroscopy*), em combinação com ferramentas de quimiometria (*machine learning, data mining e deep learning*), pode ser implementada como uma metodologia de análise para mensurar os componentes químicos e físicos de uma amostra. O NIRS permite estimar a concentração dos analitos de interesse ou classificar/autenticar as amostras de forma não destrutiva e não invasiva.

Os avanços recentes em instrumentação e redução de componentes que utilizam os sistemas NIRS tornaram muitos deles portáteis, com formato ergonômico e de fácil manipulação (PEREIRA et al., 2020). Uma vez desenvolvido os modelos preditivos, a estimação do analito pode ser feita *in situ*, o que facilita a estimação da composição das larvas dos insetos previamente ao abate e durante a preparação da farinha. Logo, isso permitirá um melhor controle de qualidade num menor tempo por parte do produtor.

2. METODOLOGIA

2.1. OBTENÇÃO DAS AMOSTRAS

As larvas da mosca soldado negro (*Hermetia illucens*) foram coletadas diretamente de um produtor agrícola localizado em Piracicaba (São Paulo, Brasil). Segredo industrial protege a divulgação do tipo de alimentação das larvas e seu sistema de abate, mas foi possível confirmar com o produtor que a alimentação dessas larvas não inclui proteína de origem animal. Cerca de 10 g de cada amostra foram moídas usando um moinho de lâminas (modelo A 11 B S32, IKA, Alemanha).

2.2. INSTRUMENTAÇÃO E AQUISIÇÃO DOS DADOS

As farinhas obtidas foram colocadas em placas de Petri e escaneadas em quatro regiões aleatórias para garantir uma melhor representação da amostra utilizando-se dois modelos espectrômetro NIR portátil: NIR-S-G1 InnoSpectra (900 - 1700 nm – espectrômetro 1) e NeoSpectra SiWare (1350 – 2562 nm – espectrômetro 2).

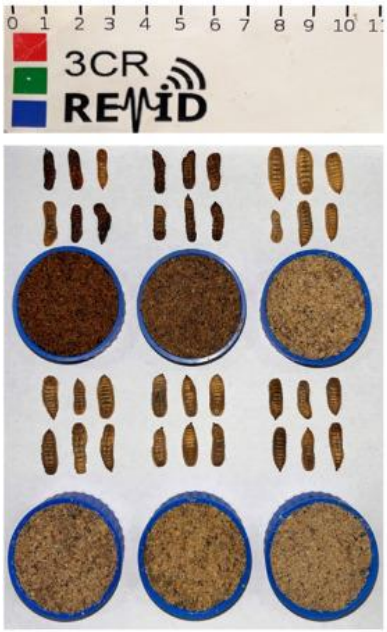

Figura 1 - Exemplo de amostras das larvas de BSF inteiras e moídas - fonte: autoria própria.

2.3. ANÁLISES DE REFERÊNCIA

O conteúdo de proteínas das farinhas das larvas de BSF foi obtido pelo método de Kjeldahl (AOAC, 2000). As análises foram feitas em triplicata. Devido à presença de quitina, foi usado um fator de conversão de nitrogênio para proteína do valor de kp=4,76.

Os lipídios da farinha das larvas de BSF foram extraídos pelo método de Bligh-Dyer (HARTMAN e LAGO, 1973). Este é um método de extração a frio, onde os lipídios são extraídos e dissolvidos em clorofórmio e, em seguida, a fase líquida foi afastada e evaporada a 35 ºC usando um evaporador rotativo. Em seguida, os lipídios residuais foram pesados e o teor de lipídios foi calculado com base no peso da farinha.

2.4 ANÁLISE DE DADOS

Os espectros obtidos foram pré-processados para remover efeitos indesejados. SNV foi implementado para remover efeitos aditivos de dispersão de luz. Primeira e segunda derivadas de Savitzky-Golay foram testados para corrigir dispersão de luz, correção de linha de base e para destacar as principais bandas. Os métodos de pré-processamento foram implementados individualmente ou combinados a SNV mais derivadas. Além disso, os dados foram centrados na média antes da análise quimiométrica.

A análise de componentes principais (PCA) foi usada nesse trabalho para observar a distribuição de variância de amostras de farinha de larvas de BSF.

Os modelos de regressão utilizados foram Regressão Parcial de Mínimos Quadrados (PLSR) e Regressão de Máquina de Vetor de Suporte (SVMR). Antes da análise, os dados foram aleatoriamente divididos em um conjunto de calibração (70%, 177 amostras) e um conjunto de teste externo (30%, 75 amostras). Para construir o modelo PLSR, os modelos foram selecionados usando variáveis latentes (latent variables – LVs) com o menor erro quadrado médio da validação cruzada (mean square error of the cross-validation – RMSECV). Já para construir o modelo SVMR, os dados primeiramente foram comprimidos com PLS usando 10-12 variáveis latentes. Após isso, os modelos SVMR foram desenvolvidos usando uma função radial kernel-based (RBF), e sua otimização foi baseada nos parâmetros c (custo), g (gamma) e ε (epsilon) com uma pesquisa de grade baseada no menor RMSECV de uma validação cruzada de 10 vezes.

Foram testados quatro métodos de seleção de variáveis: rPLS, iPLS, CovSel e GA. A seleção de variáveis foi realizada após a aplicação da melhor abordagem de pré-processamento para cada espectrômetro e propriedades estimadas (teor de proteínas e lipídios).

Para analisar os dados espectrais obtidos nos sensores portáteis NIR, o algoritmo Kennard-Stone, a análise exploratória por PCA, os modelos PLSR e SVMR de treinamento e teste assim como a etapa de seleção de variáveis foram inteiramente realizados usando caixa de ferramenta PLSToolbox (Eigenvector Research, Inc., Manson, WA, EUA) para MATLAB R2020a (Mathworks, Natick, EUA).

3. RESULTADOS E DISCUSSÃO

3.1 PERFIL DE ESPECTROS NIR

A figura 2A e 2B mostra os espectros brutos e pré-processados (SNV + 1ª derivada S-G (polinômio de 2ª ordem, 7 pontos de janela)), respectivamente, obtidos pelo espectrômetro 1. Já a figura 2C e 2D mostra os espectros brutos e pré-processados (SNV + 1ª derivada S-G (polinômio de 2ª ordem, 5 pontos de janela)), respectivamente, obtidos pelo espectrômetro 2. Os espectros NIR obtidos usando o espectrômetro 1 apresentaram picos relevantes em 928, 1170, 1225, 1410, 1450, 1490, 1510, 1530, 1570 e 1670 nm (Fig. 2B). O perfil espectral da farinha BSF obtido com o espectrômetro 2 apresentou picos relevantes em 1410, 1490, 1510, 1530, 1570, 1720, 1760, 1900, 2050, 2150, 2250 e 2450 nm (Figura 2D).

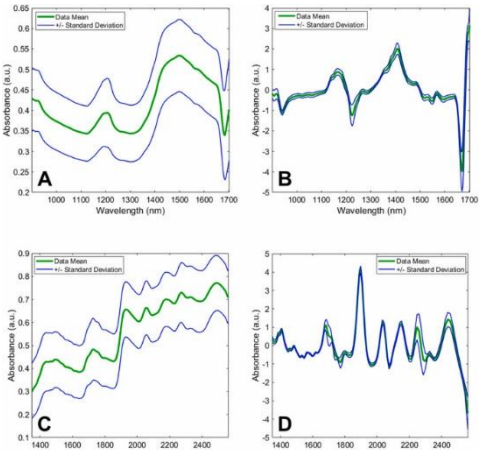

Figura 2 - Espectros brutos e pré-processados obtidos pelos espectrômetros 1 e 2 - fonte: autoria própria.

Para o espectrômetro 1, os picos em 928 nm, 1170 nm e 1225 nm são relacionados ao teor de óleo, hidrocarbonetos alifáticos (CH=CH) e estruturas CH_3 (RIU et al., 2022). A região espectral em torno de 1410 e 1450 nm está relacionada ao teor de água. A região espectral em 1490-1570 nm, exibida em ambos os perfis de espectros, está relacionada a aminas (1490 nm), conteúdo de proteína (1510 nm) e estruturas de NH_2 (1530 nm) (OSBORNE, 2006).

Para o espectrômetro 2, os picos em 1720 nm e 1760 nm esão relacionados com lipídios e ácidos graxos. O pico em 1920 nm está relacionado com a estrutura primária de proteínas (OSBORNE, 2006). O pico em 2050 nm está relacionado com o conteúdo de proteínas (RIU et al., 2022).

Ambos os espectrômetros apresentaram picos relevantes associados ao conteúdo de proteínas e lipídios da farinha da larva de BSF.

3.2 RESULTADOS PCA

O PCA foi realizado com base em espectros pré-processados (Figura 2). Juntos, PC1 e PC2 capturaram 91,6% e 90,8% da variância do conjunto de dados obtidos do espectrômetro 1 e do espectrômetro 2, respectivamente. Na figura 3A e 3D, é possível ver um gráfico de dispersão de pontuação de acordo com o conteúdo de proteína (%) e a figura 3B e 3E mostra um gráfico de dispersão de pontuação de acordo com o conteúdo de lipídios (%) para ambos os espectrômetros. De maneira geral, é possível observar uma tendência na distribuição (variância) das amostras em função do teor de proteínas e lipídeos.

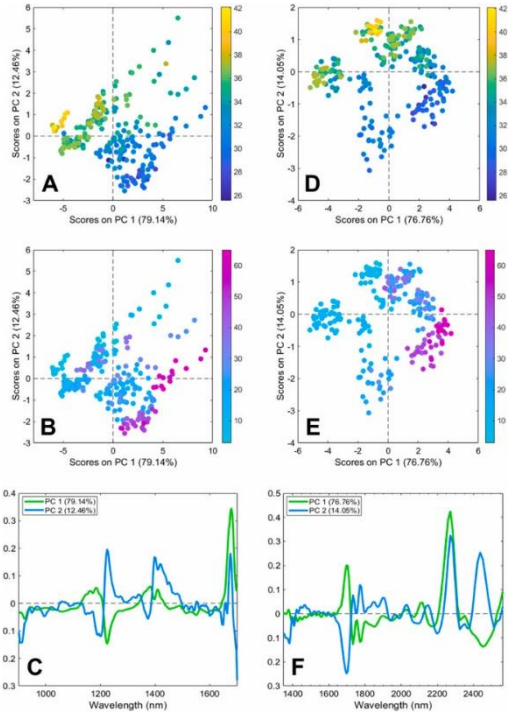

Figura 3 - Resultados PCA dos espectrômetros 1 e 2 - fonte: autoria própria.

3.3. RESULTADOS DA REGRESSÃO

Os melhores modelos de regressão baseados em variáveis selecionadas rastreadas por rPLS, GA, iPLS e CovSel são demonstrados na Tabela 1 e na Tabela 2, onde "Param" se refere ao número de LVs ótimos para modelos PLSR ou custo e valor gama obtidos da otimização SVMR, além da formatação "rPLS (modo seletivo; número de variáveis)", "iPLS (tamanho das janelas; número de variáveis)" e "GA (tamanho das janelas; número de variáveis)".

Dispositivo NIR	Modelo	Pre-proc	Param*	R^2_C	RMSEC	R^2_{CV}	RMSECV	R^2_P	RMSEP	RPD
1	PLSR	SNV+1ª	3	0,81	1,34	0,80	1,39	0,81	1,31	2,42
	SVMR	SNV+1ª	10; 0,01	0,91	0,93	0,84	1,26	0,85	1,19	2,66
		rPLS(sel;17)	3,2;0,032	0,91	0,94	0,87	1,13	0,86	1,20	2,65
2	PLSR	SNV+2ª	2	0,82	1,27	0,81	1,30	0,83	1,27	2,50
		iPLS(1;16)	3	0,84	1,18	0,83	1,24	0,84	1,21	2,62
	SVMR	1ª	100;0,003	0,91	0,89	0,82	1,28	0,85	1,19	2,66
		iPLS(1;16)	3,2;0,032	0,91	0,88	0,82	1,28	0,86	1,18	2,71

Tabela 1 - Melhores resultados obtidos pelos modelos de regressão para conteúdo de proteína (%).

Dispositivo NIR	Modelo	Pre-proc	Param*	R^2_C	RMSEC	R^2_{CV}	RMSECV	R^2_P	RMSEP	RPD
1	PLSR	SNV+1ª	4	0,80	6,75	0,78	7,04	0,76	6,44	2,35
		GA(1;33)	3	0,85	5,86	0,84	6,05	0,82	5,52	2,75
	SVMR	1ª	10;0,032	0,97	2,58	0,91	4,48	0,91	3,76	4,04
		iPLS(10;100)	3,2;0,1	0,98	2,32	0,92	4,35	0,91	3,78	4,01
2	PLSR	SNV+1ª	4	0,90	4,80	0,89	5,11	0,85	5,73	2,65
		iPLS(1;20)	3	0,92	4,30	0,92	4,42	0,88	5,13	2,95
	SVMR	SNV+1ª	3,2;0,031	0,98	1,94	0,95	3,36	0,94	3,51	4,32
		GA(15;82)	31,6;0,01	0,98	1,93	0,94	3,79	0,96	3,05	4,97

Tabela 2 - Melhores resultados obtidos pelos modelos de regressão para conteúdo de lipídios (%).

Para a previsão do conteúdo de proteína (%), tanto o espectrômetro 1 (900–1700 nm) quanto o espectrômetro 2 (1350–2562 nm) mostraram desempenho semelhante (Tabela 1). Comparando os modelos PLSR e SVMR, ambos apresentaram desempenho semelhante, com RMSEP entre 1,19 e 1,59 (%) e valores de RPD entre 2,19 e 2,66, o que indica um bom modelo de regressão para previsão (SAEYS et al., 2005). Os valores gama baixos (0,003–0,035) para SVMR indicam nenhum sobreajuste e, mais importante, um comportamento de kernel linear (CRUZ TIRADO et al., 2023). Portanto, isso explicaria a semelhança com o desempenho do modelo PLSR para prever o teor de proteína (%).

Houve diferença entre a faixa NIR avaliada e os algoritmos de regressão para prever o conteúdo de lipídios (%). Em geral, o espectrômetro 2 (1350–2562 nm) foi superior ao espectrômetro 1 (900–1700 nm), e o SVMR teve um desempenho melhor que o PLSR (Tabela 2). Provavelmente a superioridade do espectrômetro 2 em relação ao espectrômetro 1 foi devido à faixa de comprimento de onda, que incluía comprimentos de onda relacionados a estruturas químicas presentes em ácidos graxos (por exemplo, 1685–1800 nm) (CAPORASO, WHITWORTH e FISK, 2021; OSBORNE, 2006; RIU et al., 2022).

Conforme observado, ambos os espectrômetros apresentaram diferenças de desempenho, o que é esperado, considerando a diferença das peças que os compõem, o que afeta o preço, e a diferença de faixa espectral. Considerando aplicações práticas em larvas BSF, ambos os espectrômetros NIR portáteis acoplados com PLSR ou SVMR podem estimar adequadamente o teor de proteínas e lipídios, respectivamente, de farinhas de larvas BSF. Além do desempenho relatado, deve-se considerar o custo reduzido de equipamentos e análises, a isenção de produtos químicos e a velocidade de aquisição de dados em uma quantidade amostral maior do que utilizando métodos tradicionais.

4. CONCLUSÃO

Os modelos PLSR e SVMR mostraram desempenho semelhante para prever o teor de proteína (%), com RMSEP tão baixo quanto 1,18%, independentemente do espectrômetro. Diferentemente, a predição do conteúdo de lipídios foi mais eficiente usando SVMR, com valores de RMSEP inferiores a 40% em relação ao PLSR, o que pode indicar uma tendência não linear dos dados. Por fim, os modelos apresentados neste trabalho comprovam a robustez de ambos os espectrômetros NIR portáteis para predizer o teor de proteínas e lipídeos na farinha de larvas de BSF.

BIBLIOGRAFIA

CALIGIANI, A. et al. Composition of black soldier fly prepupae and systematic approaches for extraction and fractionation of proteins, lipids and chitin. **Food Research International**, v. 105, p. 812–820, 1 mar. 2018.

CAPORASO, Nicola; WHITWORTH, Martin B.; FISK, Ian D. Total lipid prediction in single intact cocoa beans by hyperspectral chemical imaging. **Food Chemistry**, v. 344, p. 128663, 2021.

CRUZ-TIRADO, J. P. et al. Rapid and non-destructive cinnamon authentication by NIR-hyperspectral imaging and classification chemometrics tools. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 289, p. 122226, 2023.

HARTMAN, L.; LAGO, R. C. Rapid preparation of fatty acid methyl esters from lipids. **Laboratory Practice**, v. 22, pp. 475-476, 1973.

MUREFU, T. R. et al. Safety of wild harvested and reared edible insects: A review. **Food Control**, v. 101, p. 209–224, 1 jul. 2019.

OSBORNE, Brian G. Near-infrared spectroscopy in food analysis. **Encyclopedia of analytical chemistry: applications, theory and instrumentation**, 2006.

PEREIRA, E. V. DOS S. et al. Simultaneous determination of goat milk adulteration with cow milk and their fat and protein contents using NIR spectroscopy and PLS algorithms. **LWT**, v. 127, 1 jun. 2020.

RIU, Jordi et al. Exploring the Analytical Complexities in Insect Powder Analysis Using Miniaturized NIR Spectroscopy. **Foods**, v. 11, n. 21, p. 3524, 2022.

SAEYS, Wouter; MOUAZEN, Abdul Mounem; RAMON, Herman. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. **Biosystems engineering**, v. 91, n. 4, p. 393-402, 2005.

VARUNJIKAR, M. S. et al. Shotgun proteomics approaches for authentication, biological analyses, and allergen detection in feed and food-grade insect species. **Food Control**, v. 137, p. 108888, jul. 2022.