XÁC ĐỊNH NHANH HÀM LƯỢNG CHẤT BÉO TRONG CÁ BẰNG ĐO QUANG PHỔ NIR KẾT HỢP PHÂN TÍCH HỒI QUY PLS

Phạm Ngọc Hưng[1*], Lê Tuấn Phúc[1], Trần Thị Thanh Hoa[1], Cung Thị Tố Quỳnh[1], Lại Quốc Đạt[2], Nguyễn Hoàng Dũng[2], Đặng Minh Nhật[3], Lê Thành Nhân[4], Hoàng Quốc Tuấn[1]

[1]*Viện Công nghệ Sinh học và Công nghệ Thực phẩm - Trường Đại học Bách Khoa Hà Nội*

[2]*Trường Đại học Bách Khoa - ĐHQGHCM*

[3]*Trường Đại học Bách Khoa - Đại học Đà Nẵng*

[4]*Viện Công nghệ Quốc tế DNIIT - Đại học Đà Nẵng*

*Email: *hung.phamngoc@hust.edu.vn*

Ngày nhận bài: 10/6/2022; Ngày chấp nhận đăng: 10/8/2022

TÓM TẮT

Một phương pháp dự đoán hàm lượng chất béo của cá được xây dựng bằng cách sử dụng quang phổ cận hồng ngoại (NIR) kết hợp với mô hình hồi quy bình phương tối thiểu một phần (PLS). Để xây dựng và tối ưu hóa mô hình dự đoán, 25 mẫu cá Nục đã được thu thập ngẫu nhiên để tiến hành đo NIR tại 4 vùng trên thân cá, đồng thời hàm lượng chất béo cũng được xác định bằng phương pháp hoá học. Mô hình dự đoán được tối ưu hoá dựa trên phương pháp lựa chọn các bước sóng có ý nghĩa và loại dần các bước sóng còn lại để đạt được giá trị của hệ số tương quan lớn và sai số trung bình bình phương nhỏ nhất. Mô hình dự đoán tốt nhất được xây dựng dựa trên dữ liệu đo NIR tại vùng bụng dưới của cá với hệ số tương quan 0,96 và sai số trung bình bình phương 0,001 cho tập xác thực chéo.

Từ khóa: NIR, cá, chất béo, mô hình đa biến, PLS.

1. ĐẶT VẤN ĐỀ

Việt Nam là nước có đường bờ biển dài 3.650 km, đa dạng về các loài thủy sản. Sản lượng khai thác thủy sản lên tới 8,4 triệu tấn vào năm 2020, tăng trung bình khoảng 8% mỗi năm trong những năm gần đây. Bên cạnh việc đánh bắt thì sản lượng nuôi trồng cũng được mở rộng với tổng sản lượng là 4,6 triệu tấn (năm 2020) [1].

Thị trường cá và các sản phẩm chế biến từ cá trên thế giới đang tăng trưởng liên tục. Chất lượng và độ an toàn của cá chủ yếu bị ảnh hưởng bởi quá trình bảo quản, thời gian và nhiệt độ bảo quản. Các sản phẩm từ cá có chứa hàm lượng chất béo cao nên rất dễ bị oxy hóa gây mùi ôi thiu ở ngay nhiệt độ môi trường [2]. Sự thay đổi về màu sắc, cấu trúc, độ đàn hồi và các đặc tính sinh hóa của cá cũng là yếu tố quan trọng ảnh hưởng đến tâm lý người tiêu dùng và quyết định hành vi mua hàng tiếp theo của họ.

Quang phổ NIR (Near Infrared Reflectance) là một kỹ thuật phân tích với khả năng phân tích nhanh, dễ sử dụng và đặc biệt là không cần phá hủy mẫu cũng như công đoạn chuẩn bị mẫu không phức tạp như các phương pháp hóa học thông thường [3], được sử dụng phổ biến trong những năm gần đây đối với ngành công nghệ thực phẩm. Tín hiệu quang phổ thu được từ máy đo quang phổ NIR đã được áp dụng để đánh giá cấu trúc hóa học trên một số loại hoa quả tươi [4]. NIR cũng đã được nghiên cứu để áp dụng xác định độ tươi của cá [5], thành phần

chất béo trên cá đông lạnh [6], độ ẩm và hàm lượng protein của cá [7]. Các nghiên cứu đã cho thấy nhiều ưu điểm khi sử dụng quang phổ NIR để phân tích và xây dựng các phương pháp đánh giá thuộc tính của các sản phẩm nông sản, thuỷ hải sản.

PCR (Principal Component Regression) và PLS (Partial Least Square) là những phương pháp được sử dụng khi xử lý các dữ liệu đa chiều. Hai phương pháp này đều thực hiện giảm chiều dữ liệu và diễn tả được các quan sát trong một không gian mới được gọi là không gian biến tiềm ẩn. Trong PCR, việc chuyển dữ liệu sang không gian mới được thực hiện bằng cách chỉ sử dụng thông tin của các đặc trưng như thông tin về độ hấp thụ của các bước sóng quang phổ. Bên cạnh đó, hồi quy PLS có thể kết hợp thông tin về đặc trưng và quan sát, nghĩa là cả độ hấp thụ của các bước sóng và nồng độ đo được của quan sát đó. Một số nghiên cứu đã chỉ ra rằng phân tích theo PLS có thể cho mô hình hồi quy với kết quả tốt hơn phương pháp PCR ở dữ liệu phổ [8, 9].

Trong hồi quy OLS (Original Least Square), các giá trị ước lượng được tính theo phương pháp tối thiểu hoá tổng bình phương khoảng cách giữa giá trị ước lượng và giá trị thực tế của điểm dữ liệu. Do vậy, để mô hình hồi quy OLS đạt được kết quả tốt nhất thì cần đảm bảo được sự tuyến tính của các hệ số hồi quy, các yếu tố dự báo phải không liên quan đến phần dư, các phần dư phải có phương sai không thay đổi [10, 11]. Một số nghiên cứu cũng đã chỉ ra rằng phương pháp OLS có hiệu quả thấp hơn so với phương pháp PLS khi áp dụng với bộ dữ liệu có kích thước mẫu nhỏ và các biến đặc trưng có xảy ra sự đa cộng tuyến [12].

Mục tiêu của nghiên cứu này là áp dụng mô hình hồi quy đa biến để phân tích dữ liệu quang phổ NIR nhằm xác định nhanh hàm lượng chất béo trong cá. Các phương pháp tiền xử lý dữ liệu bao gồm: SNV (Standard Normal Variate), MSC (Multipicative Scatter Correction) và đạo hàm đã được dùng để hiệu chỉnh dữ liệu NIR đo trên các mẫu cá. Mô hình hồi quy PLS được áp dụng để xác định mối tương quan giữa hàm lượng chất béo và dữ liệu quang phổ NIR. Các bước sóng quang phổ được khảo sát và lựa chọn nhằm tối ưu và nâng cao hiệu suất của mô hình. Hiệu suất của mô hình hồi quy được đánh giá theo hệ số tương quan (R^2) và giá trị sai số trung bình bình phương (MSE - Mean Square Error) và của bộ dữ liệu dự đoán và tập dữ liệu xác thực chéo.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Vật liệu nghiên cứu

2.1.1. Mẫu cá

Hai mươi lăm mẫu cá Nục (n = 25) được thu thập ngẫu nhiên trong tháng 12/2021, tại một số cảng cá thuộc các tỉnh phía Bắc. Các mẫu đã được thu thập theo TCVN 5276:1990 [13], được làm lạnh đông trước khi vận chuyển về Trung tâm Đào tạo và Phát triển sản phẩm thực Phẩm - Viện Công nghệ Sinh học và Công nghệ Thực phẩm - Trường Đại học Bách Khoa Hà Nội. Các mẫu được bảo quản lạnh đông ở nhiệt độ -18^{o}C để không làm ảnh hưởng đến chất lượng mẫu cho đến khi phân tích.

2.1.2. Hóa chất và thiết bị xác định hàm lượng chất béo

Các hoá chất được sử dụng để xác định hàm lượng chất béo của cá như natri sulfat (Na_2SO_4) khan (Đức), ete etylic ($C_2H_5OC_2H_5$) (Đức) và một số hoá chất khác đều có độ sạch PA và sử dụng một lần. Thiết bị được sử dụng bao gồm: thiết bị chưng cất Soxhlet (Đức); cân phân tích độ chính xác 0,001 g (Đức); nồi cách thủy có thể điều chỉnh nhiệt độ và bình hút ẩm (Trung Quốc).

2.1.3. Thiết bị đo quang phổ cận hồng ngoại

Nghiên cứu sử dụng thiết bị quang phổ DLP® NIRscan Nano EVM, được cung cấp bởi Texas Instruments, Dallas, Texas, Hoa Kỳ. Thiết bị đo được thiết kế để thu thập tín hiệu phản xạ của 228 bước sóng quang phổ phân bố đều từ 900 nm đến 1700 nm.

Hình 1. Thiết bị đo quang phổ cận hồng ngoại DLP® NIRscan Nano EVM

2.2. Phương pháp nghiên cứu

2.2.1. Thu thập dữ liệu quang phổ NIR

Quang phổ NIR của các mẫu cá được thu thập theo nguyên lý phản xạ và biến đổi Fourier. Mỗi con cá tiến hành đo tiếp xúc trực tiếp tại 4 vùng vị trí trên thân cá bao gồm: vị trí thân trên (1), thân giữa (2), thân dưới (3) và thân dưới bụng (4) được mô tả như trong Hình 2. Tại mỗi vùng vị trí, thực hiện đo quang phổ tại 20 điểm ngẫu nhiên trong vùng đó. Dữ liệu quang phổ của mỗi lần đo là giá trị trung bình của 6 lần quét liên tiếp được tự động thực hiện theo phần mềm của thiết bị đo quang phổ. Sau khi thực hiện thu dữ liệu quang phổ, các mẫu cá sẽ được xác định hàm lượng chất béo theo phương pháp hóa học.

Hình 2. Các vùng vị trí đo quang phổ NIR trên mẫu cá Nục

2.2.2. Phương pháp xác định hàm lượng chất béo

Hàm lượng chất béo được xác định theo TCVN 3703:2009 [14] về thủy sản và sản phẩm thủy sản – xác định hàm lượng chất béo, sử dụng phương pháp chưng cất Soxhlet với dung môi chiết là dung môi hữu cơ. Mẫu cá được bỏ đầu, vảy, đuôi, vây, ruột và phần xương không ăn được, lọc phi lê toàn bộ phần thịt và da từ đầu đến đuôi. Sử dụng máy xay nhanh mẫu thử 3 lần. Sau mỗi lần, loại bỏ phần mẫu không nghiền được nhỏ từ máy xay và trộn kỹ mẫu. Mỗi mẫu tiến hành đo lặp lại 3 lần, hàm lượng chất béo được biểu diễn bằng giá trị trung bình ± độ lệch chuẩn.

Hàm lượng chất béo X, được biểu thị bằng phần trăm khối lượng (%), theo công thức:

$$X = \frac{m_1.100}{m}$$

Trong đó:

- m_1 là khối lượng chất béo thu được tính bằng gam (g);
- m là khối lượng mẫu thử tính bằng gam (g).

Biểu thị kết quả đến hai chữ số thập phân.

2.2.3. Phương pháp tiền xử lý dữ liệu

Nghiên cứu sử dụng phương pháp SNV [15], MSC [16], Savitzky – Golay [17] để tiền xử lý các dữ liệu quang phổ nhằm giảm các tác động của nhiễu, xử lý các dữ liệu khuyết thiếu, giảm ảnh hưởng của hiệu ứng cộng và nhân khi đo quang phổ.

2.2.4. Xác thực chéo

Phương pháp xác thực chéo K-Fold CV (Cross Validation) được sử dụng để đánh giá hiệu quả của mô hình hồi quy, thể hiện trên các giá trị R^2_CV và MSE_CV trên tập xác thực chéo.

2.2.5. Mô hình dự đoán

Mô hình dự đoán giá trị hàm lượng chất béo trong cá, theo phương pháp hồi quy PLS, được dựa trên dữ liệu quang phổ NIR thu được từ 228 bước sóng trong dải từ 900 đến 1700 nm. Trong xây dựng mô hình dự đoán này, mỗi bước sóng sẽ được coi là một biến. Dữ liệu ban đầu được chiếu trong không gian biến tiềm ẩn nhằm tối đa hóa hiệp phương sai giữa các biến trong không gian mới với giá trị cần được ước lượng. Phương pháp lọc được sử dụng để xác định và khảo sát sự ảnh hưởng và lựa chọn bước sóng đến kết quả dự đoán của mô hình hồi quy [21]. Theo đó, trong mỗi vòng lặp, bước sóng có giá trị tuyệt đối thấp nhất của hệ số hồi quy sẽ được loại bỏ và mô hình hiệu chuẩn được xây dựng lại dựa trên việc sử dụng MSE của bộ xác thực chéo làm số liệu tham chiếu. Quy trình được lặp lại cho đến khi còn lại 1 bước sóng cuối cùng. Hệ số MSE được lưu lại mỗi khi loại bỏ một bước sóng. Sau khi kết thúc vòng lặp, tập hợp các bước sóng cho kết quả MSE thấp nhất được chọn là mô hình hồi quy tốt nhất.

Đánh giá hiệu quả của mô hình căn cứ theo các hệ số tương quan của tập huấn luyện R^2_Calib, hệ số tương quan của tập xác thực chéo R^2_CV, sai số trung bình bình phương của tập huấn luyện MSE_Calib và sai số trung bình bình phương của tập xác thực chéo MSE_CV. Giá trị R^2 càng lớn và MSE càng nhỏ thì năng lực và hiệu quả dự đoán của mô hình càng tốt. Để tối ưu hoá mô hình, các biến thích hợp sẽ được lựa chọn, biến không thích hợp sẽ được loại bỏ.

Công thức tính MSE và R^2 như sau:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}$$

Trong đó:

- y_i: giá trị thực tế
- $\hat{y}_i$: giá trị ước lượng từ mô hình
- $\bar{y}_i$: giá trị trung bình của tập giá trị thực tế

Dữ liệu tính toán và xử lý cũng như công cụ tiền xử lý được áp dụng thực hiện trong môi trường Python 3.9 với các thư viện đi kèm là Numpy, Pandas, Sklearn, Scipy. Kiểm định Paired T-Test được sử dụng để đánh giá ý nghĩa về mặt thống kê kết quả dự đoán của mô hình với giá trị tham chiếu được đo bằng phương pháp hóa học

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Hàm lượng chất béo trong các mẫu cá

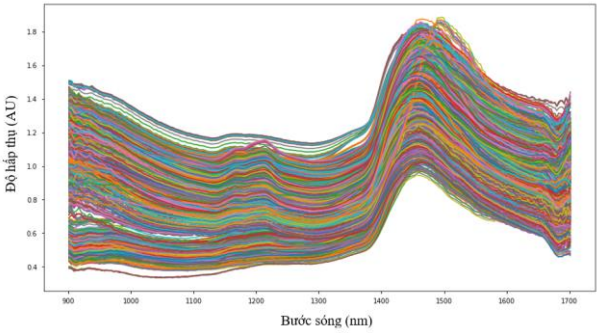
Hàm lượng chất béo trong 25 mẫu cá nục trong nghiên cứu được định lượng theo TCVN 3703:2009 [14] bằng phương pháp chưng cất lôi cuốn trong dung môi hữu cơ. Kết quả được thể hiện trong Bảng 1 với hàm lượng chất béo tính bằng g/100g mẫu.

Bảng 1. Hàm lượng chất béo trong các mẫu cá nghiên cứu

STT	Mã hóa mẫu	Hàm lượng chất béo (g/100g)	STT	Mã hóa mẫu	Hàm lượng chất béo (g/100g)
1	TN1	1,12 ± 0,05	14	TN14	1,20 ± 0,13
2	TN2	1,28 ± 0,06	15	TN15	1,53 ± 0,13
3	TN3	1,06 ± 0,08	16	TN16	1,42 ± 0,04
4	TN4	1,48 ± 0,03	17	TN17	1,27 ± 0,13
5	TN5	1,28 ± 0,03	18	TN18	1,17 ± 0,05
6	TN6	1,10 ± 0,04	19	TN19	1,53 ± 0,13
7	TN7	1,25 ± 0,04	20	TN20	1,10 ± 0,11
8	TN8	1,54 ± 0,13	21	TN21	1,42 ± 0,13
9	TN9	1,07 ± 0,04	22	TN22	1,40 ± 0,09
10	TN10	1,08 ± 0,17	23	TN23	1,51 ± 0,09
11	TN11	1,07 ± 0,08	24	TN24	1,07 ± 0,05
12	TN12	1,48 ± 0,12	25	TN25	1,51 ± 0,07
13	TN13	1,51 ± 0,07			

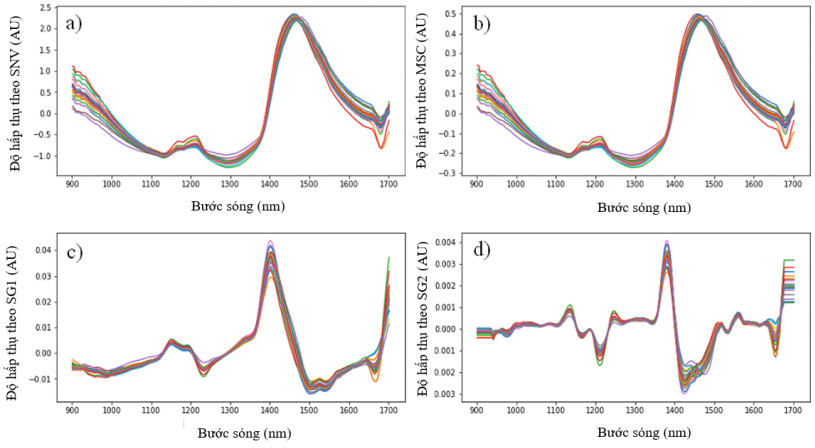
3.2. Tiền xử lý dữ liệu quang phổ

Hình 3 mô tả dữ liệu từ 2000 phép đo quang phổ NIR của 25 mẫu cá trong dải bước sóng từ 900 đến 1700 nm. Kết quả cho thấy, hình dáng của các quang phổ có nét tương đồng do sự có mặt của nước và chất béo trong các mẫu cá được hấp thụ. Các quang phổ có các đỉnh tại các bước sóng liên quan tới đặc trưng của các liên kết O-H và C-H. Cụ thể, bước sóng 1600-1700 nm liên quan tới cộng hưởng O-H của nước; bước sóng 1400-1500 nm liên quan tới âm bội của O-H và C-H; bước sóng 1150-1250 nm liên quan tới âm bội 2 của C-H và bước sóng 900-1000 nm liên quan tới dải âm bội thứ 3 của O-H và của C-H.



Hình 3. Dữ liệu quang phổ NIR đo được trên 25 mẫu cá

Dữ liệu quang phổ NIR thu được được tiền xử lý bằng cách áp dụng 4 phương pháp: SNV, MSC; SG1 (Savitzky-Golay Derivative 1) và SG2 (Savitzky-Golay Derivative 2). Sử dụng phương pháp SNV và MSC, các hiệu ứng cộng và nhân trong dữ liệu quang phổ thô đã bị loại bỏ, đồ thị của quang phổ đã được tập trung hơn và được thể hiện trên Hình 4.a và Hình 4.b. Trong khi đó, phương pháp tiền xử lý Savitzky-Golay, đồ thị của quang phổ cũng được tập trung, đồng thời đỉnh và đáy của đồ thị được thể hiện rõ ràng hơn do được khuếch đại nhờ đạo hàm (Hình 4.c và Hình 4.d).



Hình 4. Các phép tiền xử lý a) SNV; b) MSC; c) Đạo hàm cấp 1 SG1; d) Đạo hàm cấp 2 SG2;

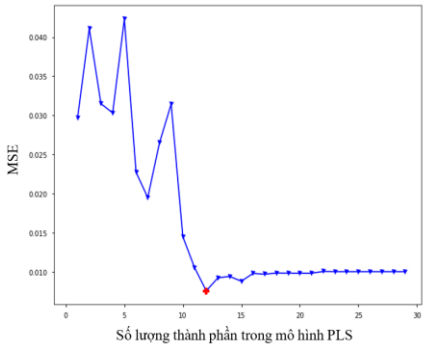
3.3. Ước lượng hàm lượng chất béo theo mô hình PLS

Dữ liệu quang phổ NIR được đo trên 4 vùng vị trí: thân trên, thân giữa, thân dưới, và thân bụng dưới. Các vị trí này có hàm lượng chất béo khác nhau. Do vậy, việc gộp chung dữ liệu đo NIR từ tất cả các vùng để xây dựng mô hình ước lượng chất béo sẽ không đạt độ chính xác cao và có sai số lớn. Để tăng độ chính xác cho mô hình dự đoán, bốn mô hình hồi quy PLS sẽ được xây dựng cho cả 4 vùng đo một cách riêng biệt. Mô hình có hiệu quả tốt nhất sẽ được sử dụng để ước lượng hàm lượng chất béo trong cá.

Theo quan sát, phần bụng dưới của cá có thể coi là vùng đặc trưng để đánh giá hàm lượng chất béo. Do vậy, phần trình bày sau đây là kết quả của việc xây dựng mô hình dự đoán chất béo trong cá dựa trên dữ liệu đo quang phổ từ vùng bụng dưới. Với 3 vùng còn lại, việc xây dựng mô hình và tính toán các giá trị tương quan, sai số được tiến hành tương tự.

Xác định số lượng thành phần tối ưu là một vấn đề quan trọng đối với mô hình hồi quy PLS. Việc lựa chọn số lượng thành phần quá ít dẫn đến mất thông tin, trong khi lựa chọn số

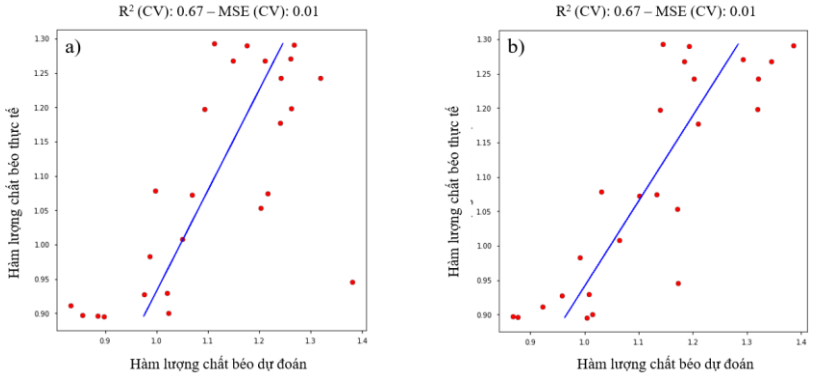
lượng thành phần quá nhiều dễ dẫn đến mô hình có khả năng dự đoán kém [18]. Một số nghiên cứu đã xác định số lượng thành phần cần giữ lại trong hồi quy PLS của tập dữ liệu Latex và Oxy đã chỉ ra khi số lượng thành phần của mô hình hồi quy PLS vượt quá một giá trị chặn trên thì mô hình hồi quy không thay đổi và cho kết quả kém [19].



Hình 5. Giá trị MSE_CV khi thay đổi số thành phần trong mô hình PLS

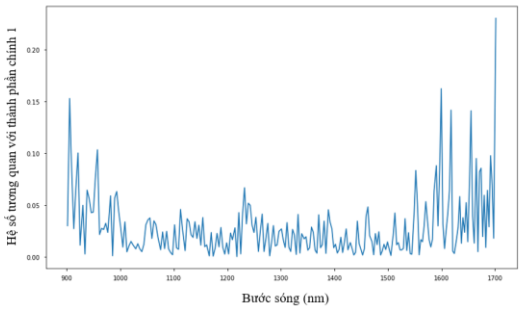
Một vòng lặp được thực hiện để khảo sát ảnh hưởng của số lượng thành phần của mô hình hồi quy tới chỉ số MSE_CV khi hồi quy dữ liệu quang phổ NIR với hàm lượng chất béo trong cá. Hình 5 cho thấy khi số lượng thành phần của mô hình là 12 thì sai số MSE_CV trong mô hình hồi quy PLS là thấp nhất.

Độ tương quan của mô hình được cải thiện hơn khi sử dụng số thành phần tối ưu đã khảo sát cho mô hình hồi quy. Trong Hình 6, hệ số tương quan R^2_CV tăng từ 0,36 lên 0,67 đồng thời sai số bình phương trung bình MSE_CV giảm từ 0,0144 xuống 0,0075 tương ứng với số thành phần của mô hình ban đầu để mặc định là 10 và số lượng thành phần sau khi khảo sát là 12.



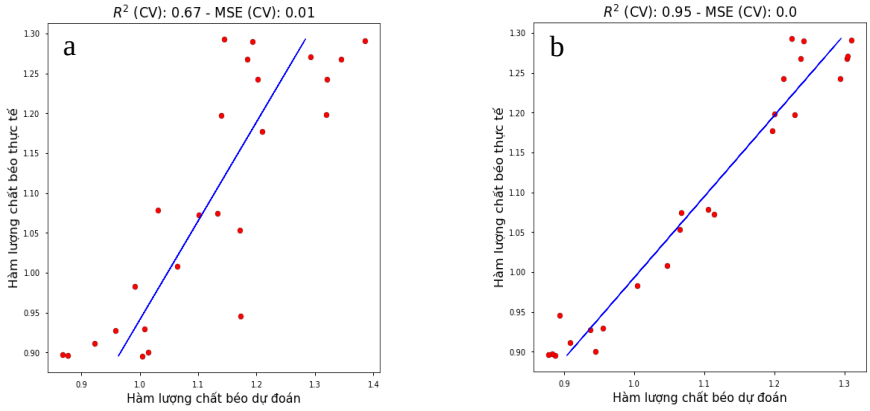
Hình 6. Đồ thị biểu thị mối tương quan kết quả ước lượng Chất béo của mô hình PLS.
a) Số thành phần của mô hình PLS bằng 10; b) Số thành phần của mô hình PLS bằng 12

Lựa chọn bước sóng của quang phổ cho mô hình hồi quy PLS đã được khảo sát trong nghiên cứu xác định hàm lượng vitamin B_{12} của tá dược bằng phương pháp quang phổ và cho kết luận rằng: Lựa chọn 33 bước sóng trong tổng số 246 bước sóng của quang phổ UV-Vis đã giúp mô hình hồi quy PLS cải thiện 86% giá trị ước lượng vitamin B_{12} [20]. Đầu tiên, mô hình PLS được áp dụng trên toàn bộ bước sóng của tập dữ liệu. Tiếp theo, sắp xếp các bước sóng đó theo thứ tự tăng dần về độ mạnh của các hệ số hồi quy tương quan với thành phần đầu tiên của mô hình PLS. Cuối cùng, loại bỏ dần các bước sóng đã sắp xếp. Các chỉ số R^2 và MSE được lưu lại mỗi khi loại bỏ một bước sóng để so sánh và tìm ra những bước sóng cho kết quả hồi quy tốt nhất.



Hình 7. Hệ số tương quan của bước sóng với thành phần thứ 1 của mô hình PLS

Từ bảng giá trị hệ số hồi quy tương quan của từng bước sóng, được thể hiện trong Hình 7. Tiến hành sắp xếp các bước sóng theo thứ tự mức độ tương quan tăng dần. Sau đó, loại bỏ dần các bước sóng theo thứ tự có độ tương quan từ thấp đến cao. Sau khi thực hiện thuật toán trên, mô hình PLS đã loại bỏ được 185 bước sóng có mức độ tương quan thấp, số lượng thành phần trong mô hình được chạy lặp lại một lần nữa để tìm ra số thành phần tối ưu cho các bước sóng còn lại. Số thành phần cho các bước sóng còn lại là 20. Mô hình PLS sau khi lựa chọn được biến đã tốt hơn hơn so với mô hình áp dụng toàn bộ các bước sóng và được thể hiện trên Hình 8. Cụ thể, hệ số tương quan R^2_CV tăng từ 0,67 lên 0,95 và giá trị MSE_CV giảm từ 0,007 xuống còn 0,001.



Hình 8. Mối tương quan kết quả dự đoán của mô hình PLS:
a) Trước khi loại bỏ bước sóng. b) Sau khi loại bỏ bước sóng

Mô hình đã ước lượng được hàm lượng chất béo trong phạm vi: 0,823 – 1,24 (g/100g). Kiểm định Paired T-Test (α = 0,05; n = 25) cho kết quả p_value = 0,17 > 0,05 chứng minh rằng kết quả dự đoán NIR kết hợp mô hình hồi quy PLS và chọn lọc bước sóng cho kết quả dự đoán không có sự khác biệt đáng kể có ý nghĩa về mặt thống kê so với giá trị tham chiếu được đo bằng phương pháp hóa học [22].

Tiến hành tương tự để xây dựng mô hình hồi quy ước lượng chất béo cho cá dựa trên dữ liệu quang phổ NIR, kết hợp các phương pháp tiền xử lý dữ liệu nêu trên, cho ba vùng còn lại trên thân cá: thân trên, thân giữa và thân dưới. Kết quả cuối cùng được thể hiện trong Bảng 2.

Kết quả thể hiện trong Bảng 2 cho thấy, mô hình hồi quy PLS cho vùng thân dưới bụng của cá có kết quả hồi quy cao nhất so với các vùng còn lại, khi phân tích quang phổ NIR với hàm lượng chất béo. Mô hình hồi quy PLS tại vị trí thân dưới bụng cho kết quả giá trị tương quan R^2_Calib ≥ 0,99 và MSE_Calib < 0,001. Đồng thời, khi áp dụng phương pháp tiền xử lý SNV với tập xác thực chéo thì các giá trị tương quan và sai số đạt tốt nhất, ở giá trị 0,96 và

0,001 tương ứng. Hệ số tương quan trong tập xác thực chéo R^2_CV nhỏ hơn giá trị R^2_calib tương ứng trong tập hiệu chuẩn cho thấy phương pháp xác thực chéo đã giúp mô hình tránh bị hiện tượng mô hình đào tạo dự đoán tốt trên tập xác thực và dự đoán kém trong tập kiểm tra (overfitting).

Bảng 2. Kết quả mô hình hồi quy PLS cho từng vị trí đo với các phép tiền xử lý khác nhau

	Tiền xử lý	Số bước sóng	Số lượng thành phần chính PLS	R^2_Calib	R^2_CV	MSE_Calib	MSE_CV
Thân trên	MSC	17	14	0,97	0,70	0,001	0,007
	SNV	15	14	0,90	0,39	0,002	0,014
	SG1	45	18	1,00	0,04	0,000	0,022
	SG2	19	10	0,91	0,29	0,002	0,016
Thân giữa	MSC	24	15	0,99	0,26	0,000	0,018
	SNV	26	15	0,99	0,75	0,001	0,006
	SG1	39	13	0,97	0,53	0,001	0,011
	SG2	34	20	1,00	0,47	0,000	0,012
Thân dưới	MSC	25	15	0,99	0,21	0,000	0,018
	SNV	25	13	0,97	0,70	0,001	0,007
	SG1	39	13	0,97	0,53	0,001	0,011
	SG2	41	17	1,00	0,39	0,000	0,014
Thân dưới bụng	MSC	43	20	0,99	0,95	0,000	0,001
	SNV	39	20	1,00	0,96	0,000	0,001
	SG1	18	13	1,00	0,90	0,000	0,002
	SG2	23	15	1.00	0,67	0,000	0,008

4. KẾT LUẬN

Kết quả của nghiên cứu này đã cung cấp một phương pháp tiếp cận, xử lý và tối ưu hóa thuật toán hồi quy PLS để dự đoán nhanh hàm lượng chất béo trong cá thông qua phép đo quang phổ NIR. Mô hình hồi quy PLS kết hợp lựa chọn biến và phương pháp tiền xử lý dữ liệu SNV, áp dụng cho tập dữ liệu đo NIR ở vùng bụng dưới của cá, cho hệ số tương quan là 0,96 và sai số bình phương trung bình là 0,001 cho tập xác thực chéo. Điều này cho thấy phương pháp phân tích quang phổ NIR kết hợp hồi quy PLS để ước lượng hàm lượng chất béo có thể là một phương pháp nhanh chóng, dễ sử dụng và có độ chính xác tương đối lớn.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi đề tài "Nghiên cứu ứng dụng các phương pháp phân tích nhanh kết hợp xử lý dữ liệu đa chiều và học máy trong kiểm soát chất lượng một số loại hải sản" Mã số: ĐTĐL.CN-33/20, Bộ Khoa học & Công nghệ.

TÀI LIỆU THAM KHẢO

1. Hiệp hội Chế biến và Xuất khẩu Thuỷ sản Việt Nam. - Tổng quan ngành thủy sản Việt Nam, https://vasep.com.vn/gioi-thieu/tong-quan-nganh.
2. Hamilton, R. J., Kalu, C., Prisk, E., Padley, F. B. & Pierce, H. -Chemistry of free radicals in lipids, Food Chemistry **60** (2) (1997) 193–199. https://doi.org/10.1016/S0308-8146(96)00351-2.

3. ROWLAND, S. J. & Rook, J. A. F. -Analytical Methods, International Journal of Dairy Technology **14** (3) (1961) 112–114. https://doi.org/10.1111/j.1471-0307.1961.tb00962.x.
4. Chen, H., Lin, B., Cai, K., Chen, A. & Hong, S. -Quantitative analysis of organic acids in pomelo fruit using FT-NIR spectroscopy coupled with network kernel PLS regression, Infrared Physics and Technology **112** (2021) https://doi.org/10.1016/j.infrared.2020.103582.
5. Kaavya, R. *et al.* -Application of infrared spectroscopy techniques for the assessment of quality and safety in spices: a review, Applied Spectroscopy Reviews **55** (7) (2020) https://doi.org/10.1080/05704928.2020.1713801.
6. Karlsdottir, M. G., Arason, S., Kristinsson, H. G. & Sveinsdottir, K. -The application of near infrared spectroscopy to study lipid characteristics and deterioration of frozen lean fish muscles, Food Chemistry **159** (2014) 420–427. https://doi.org/10.1016/j.foodchem.2014.03.050.
7. Masoum, S., Alishahi, A. R., Farahmand, H., Shekarchi, M. & Prieto, N. -Determination of protein and moisture in fishmeal by near-infrared reflectance spectroscopy and multivariate regression based on partial least squares, Iranian Journal of Chemistry and Chemical Engineering **31** (3) (2012) 51–59.
8. Zarzo, M. & Ferrer, A. -Batch process diagnosis: PLS with variable selection versus block-wise PCR, Chemometrics and Intelligent Laboratory Systems **73** (1) (2004) 15–27. https://doi.org/10.1016/J.CHEMOLAB.2003.11.009.
9. Ghasemi, J. & Niazi, A. -Simultaneous determination of cobalt and nickel. Comparison of prediction ability of PCR and PLS using original, first and second derivative spectra, Microchemical Journal **68** (1) (2001) 1–11. https://doi.org/10.1016/S0026-265X(00)00159-4.
10. Abdi, H. -Partial Least Square Regression PLS-Regression, (2003) 1–7.
11. Campbell, A. & Ntobedzi, A. -Emotional Intelligence, Coping and Psychological Distress: A Partial Least Squares Approach to Developing a Predictive Model, E-Journal of Applied Psychology **3** (2) (2007) 39–54. https://doi.org/10.7790/ejap.v3i2.91.
12. Farahani, H. A., Rahiminezhad, A., Same, L. & Immannezhad, K. -A Comparison of Partial Least Squares (PLS) and Ordinary Least Squares (OLS) regressions in predicting of couples mental health based on their communicational patterns, Procedia - Social and Behavioral Sciences **5** (2010) 1459–1463. https://doi.org/10.1016/J.SBSPRO.2010.07.308.
13. TCVN 5276:1990: Tiêu chuẩn Việt Nam về Thủy sản - Lấy mẫu và chuẩn bị mẫu.
14. TCVN 3703-2009: Tiêu chuẩn Việt Nam về Thủy sản và các sản phẩn thủy sản - Xác định hàm lượng chất béo.
15. Barnes, R. J., Dhanoa, M. S. & Lister, S. J. -Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Applied Spectroscopy **43** (5) (1989) 772–777. https://doi.org/10.1366/0003702894202201.
16. Isaksson, T. & Naes, T. -Effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Applied Spectroscopy **42** (7) (1988) 1273–1284. https://doi.org/10.1366/0003702884429869.
17. Roger, J. M., Biancolillo, A. & Marini, F. -Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems **199** (February) (2020) 103975.

https://doi.org/10.1016/j.chemolab.2020.103975.

18. Wiklund, S. *et al*. -A randomization test for PLS component selection, Journal of Chemometrics **21** (10–11) (2007) 427-439. https://doi.org/10.1002/cem.1086.
19. Lazraq, A., Cléroux, R. & Gauchi, J. P. -Selecting both latent and explanatory variables in the PLS1 regression model, Chemometrics and Intelligent Laboratory Systems **66** (2) (2003) 117-126. https://doi.org/10.1016/S0169-7439(03)00027-3.
20. Sratthaphut, L. & Ruangwises, N. -Genetic Algorithms-Based Approach for Wavelength Selection in Spectrophotometric Determination of Vitamin B12 in Pharmaceutical Tablets by Partial Least-Squares, Procedia Engineering **32** (2012) 225–231. https://doi.org/10.1016/J.PROENG.2012.01.1261.
21. Mehmood, T., Liland, K. H., Snipen, L. & Sæbø, S. -A review of variable selection methods in Partial Least Squares Regression, Chemometrics and Intelligent Laboratory Systems **118** (2012) 62-69. https://doi.org/10.1016/J.CHEMOLAB.2012.07.010.
22. Clua-Palau, G., Jo, E., Nikolic, S., Coello, J. & Maspoch, S. -Finding a reliable limit of detection in the NIR determination of residual moisture in a freeze-dried drug product, Journal of Pharmaceutical and Biomedical Analysis **183** (2020) 113163. https://doi.org/10.1016/j.jpba.2020.113163.

ABSTRACT

RAPID DETERMINATION OF FAT CONTENT IN FISH BY NEAR-INFRARED SPECTROSCOPY COMBINED WITH PARTIAL LEAST SQUARES REGRESSION

Pham Ngoc Hung[1]*, Le Tuan Phuc[1], Tran Thi Thanh Hoa[1],
Cung Thi To Quynh[1], Lai Quoc Dat[2], Nguyen Hoang Dung[2],
Dang Minh Nhat[3], Le Thanh Nhan[4], Hoang Quoc Tuan[1]
[1]*Hanoi University of Science and Technology*
[2]*Ho Chi Minh University of Technology - VNUHCM*
[3]*The University of Danang - University of Science and Technology*
[4]*Danang International Institute of Technology*
*Email: *hung.phamngoc@hust.edu.vn*

A fat content prediction model was built using near-infrared spectroscopy (NIR) combined with partial least squares regression (PLS). Twenty five samples of Nuc fishes were randomly collected for NIR measurements in 4 individual regions, and the fat content was also determined by chemical method. All data measured was used to develop and optimize the model. The model prediction was optimized based on the method of choosing significant wavelengths to obtain the highest R-square value and the smallest mean square error (MSE). The best prediction model was built based on NIR data in the lower abdomen of fish with the 0.96 of R-square and 0.001 of MSE for the cross-validation set.

Keywords: NIR, fish, fat, multivariable model, PLS.