

NGHIÊN CỨU PHÁT TRIỂN PHƯƠNG PHÁP ĐO NHANH HÀM LƯỢNG PROTEIN TRONG CÁ BIỂN BẰNG THIẾT BỊ ĐO CẬN HỒNG NGOẠI (NIRS) CẦM TAY

RESEARCH ON DEVELOPING A QUICK METHOD OF MEASURING THE PROTEIN CONTENT IN THE SEA FISH BY PORTABLE NEAR-IR REFLECTANCE SPECTROSCOPY DEVICE (NIRS)

Trần Thị Ánh Tuyết*, Nguyễn Thanh Trọng, Mạc Thị Hà Thanh, Đặng Minh Nhật

Trường Đại học Bách khoa - Đại học Đà Nẵng, Việt Nam¹

*Tác giả liên hệ / Corresponding author: ttatuyet@dut.udn.vn

(Nhận bài / Received: 24/5/2023; Sửa bài / Revised: 27/6/2023; Chấp nhận đăng / Accepted: 21/8/2023)

Tóm tắt - Nghiên cứu này nhằm phát triển phương pháp xác định nhanh hàm lượng protein của cá biển bằng cách xây dựng mô hình hồi quy tuyến tính của độ hấp thụ quang trên dải bước sóng cận hồng ngoại. Để làm được điều này, nhóm nghiên cứu đã đo độ hấp thụ quang của 37 mẫu cá, với mỗi giá trị bước sóng thu được giá trị hấp thụ X và đáp ứng Y là hàm lượng protein được phân tích bằng phương pháp chuẩn. Kết quả cho thấy mô hình hồi quy PLS tại vị trí gáy sau khi tiền xử lý dữ liệu cho kết quả tốt nhất với số yếu tố chính là 16, R-Square Cal là 0,968, R-Square Val là 0,8592. Sử dụng mô hình đó dự đoán hàm lượng protein cho 10 mẫu kiểm chứng cho sai số tương đối trong khoảng 0,06 - 8,75%. Sau khi kiểm định lại hệ số β của mô hình hồi quy bằng T-Test hai phía cho kết quả tất cả các hệ số β đều có ý nghĩa thống kê, chứng minh rằng tất cả các biến X đều ảnh hưởng đến đáp ứng Y.

Từ khóa - Quang phổ cận hồng ngoại; PLS; PCR; Protein; Thủy sản

1. Đặt vấn đề

Việc phân tích hàm lượng các chất dinh dưỡng trong mẫu thực phẩm bằng phương pháp đo phổ cận hồng ngoại dựa trên nguyên lý: Sự hấp phụ và phản xạ ánh sáng ở bước sóng 900 – 1700 nm của các chất với cấu trúc hoá học khác nhau là khác nhau [1]. Khi một mẫu của hợp chất hữu cơ được chiếu sáng thì các liên kết liên tục dao động, chúng sẽ bị bẻ cong, dẫn ra hoặc co lại. Sự dao động này tạo ra một kiểu vận động dạng sóng trong các cầu nối ở một tần số đặc trưng cho từng nhóm chức năng. Một số tia tới có tần số phù hợp với tần số của sóng dao động sẽ bị hấp thụ, trong khi các sóng có tần số khác sẽ bị phản xạ lại hoặc bị truyền qua [2]. Từ phổ thu được về các cầu nối đặc trưng cho một chất nào đó của mẫu đo nhờ phần mềm quét phổ (DLP NIRscan nano) sẽ có được một ma trận các giá trị số của phổ cho một chất hữu cơ tương ứng. Mẫu đó đồng thời cũng được phân tích thành phần hoá học. Sử dụng mô hình thống kê nhiều biến cho phép mô tả quan hệ giữa phổ hấp thụ và thành phần hoá học, quan hệ này sau đó chính là mô hình toán học để chẩn đoán thành phần hoá học của các mẫu chưa phân tích tại phòng thí nghiệm.

Một số ưu điểm của phương pháp này có thể kể đến như

Abstract - This research aims to develop the method of quickly determining the protein content of sea fish by building a linear regression model of the absorption on the near-infrared wave range. In order to do this purpose, this study measured the optical absorption on 37 fish samples, with each wavelength value obtaining the absorption value X and a response Y is the protein content of protein analyzed by the standard laboratory method. The results showed that the regression model PLS at the nape position after the data pre-processing gives the best results with the main factor: 16, R-Square Cal is 0.968, and R-Square Val is 0.8592. Using that model predicts the protein content for 10 verification samples with a relative error from 0.06 to 8.75%. After re-testing the coefficient β of the regression model by 2-sided T-Test, all values β have statistical significance, proving that all X variables affect the Y response.

Key words - Near-Infrared Analysis; PLS; PCR; Protein; Fish Science

là một kỹ thuật xanh thân thiện với môi trường do không yêu cầu dung môi và thuốc thử, không phá hủy cấu trúc mẫu và có thể tái sử dụng, phương pháp này phân tích nhanh nhiều thông số khác nhau và cung cấp kết quả chính xác trong vòng chưa đầy một phút chỉ bằng một phép đo duy nhất và có thể vận hành bởi những người không có kiến thức chuyên môn, kiến thức về các phản ứng hóa học và phân tích hóa học.

Quang phổ cận hồng ngoại được ứng dụng đầu tiên để đo độ ẩm trong ngũ cốc và hạt giống từ năm 1965 bởi Norris và cộng sự [3]. Đến những năm 2000, các nhà khoa học trên thế giới đã bắt đầu ứng dụng kỹ thuật này để xác định một số loại cá cũng như thành phần dinh dưỡng của chúng. Ứng dụng quang phổ NIR kết hợp với thuật toán PLS-DA để thiết lập mô hình phân tích dự đoán độ dai, khả năng đàn hồi, khả năng giữ nước, lực cắt, độ đàn hồi và độ cứng của cá Mè Trắng [4]. Dựa trên công nghệ VIS-NIRS để thiết lập mô hình phân biệt fillet cá ngừ tươi sống với đông lạnh / rã đông [5]. Sự kết hợp giữa quang phổ NIR và thuật toán PLS-DA được áp dụng để phân biệt bảy loại cá, bao gồm cá Trắm cỏ, cá Chép, cá Vèn, cá Mè Trắng, cá Mè Hoa, cá Trắm đen, và cá Diếc với tỷ lệ đúng 100% [6].

¹ The University of Danang - University of Science and Technology, Vietnam (Tran Thi Anh Tuyet, Nguyen Thanh Trong, Mac Thi Ha Thanh, Dang Minh Nhat)

2. Vật liệu và phương pháp nghiên cứu

2.1. Vật liệu

2.1.1. Mẫu cá và chuẩn bị mẫu

Các đối tượng mẫu thử bao gồm 13 mẫu cá Ngừ sọc dưa, 16 mẫu cá Nục sò, 3 mẫu cá Chuồn xanh, 2 mẫu cá Phèn một sọc, 3 mẫu cá Căng cát. Mẫu kiểm định bao gồm 6 mẫu cá Ngừ sọc dưa, 4 mẫu cá Nục sò. Tất cả các mẫu cá được lấy ngẫu nhiên tại cảng cá Thọ Quang, Đà Nẵng.

Bảng 1. Mã hóa vị trí đo trên mẫu cá

STT	Vị trí đo	Mã hóa
1	Ngoài da, gáy	a
2	Ngoài da, lưng	b
3	Ngoài da, đuôi	c
4	Ngoài da, bụng	d
5	Mang 1	e
6	Mang 2	f
7	Mắt 1	g
8	Mắt 2	h
9	Trong thịt, gáy	i
10	Trong thịt, lưng	j
11	Trong thịt, đuôi	k
12	Trong thịt, bụng	l

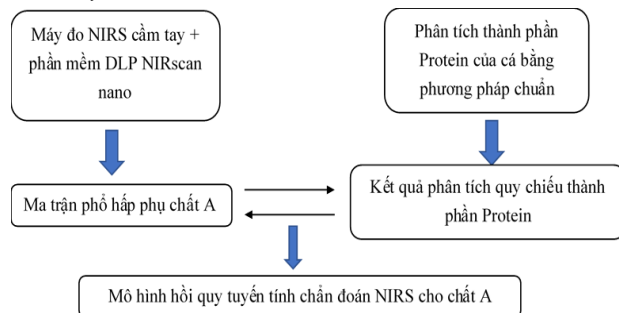
Mẫu cá sau đó sẽ được đưa về phòng thí nghiệm hóa sinh trường Đại học Bách khoa tiến hành rửa sạch rồi đo độ hấp thụ quang các mẫu cá tại 12 vị trí như Bảng 1, tại mỗi vị trí lặp lại kết quả 5 lần. Sau khi đo cá được fillet loại bỏ hoàn toàn da và xương chỉ giữ lại phần thịt đem xay và xác định chỉ tiêu protein theo AOAC 2012 (950.46).

2.1.2. Hóa chất

Các hóa chất để phân tích thành phần protein gồm H_2SO_4 , K_2SO_4 , $CuSO_4$, $NaOH$, H_3BO_3 là của Xilong (Trung Quốc). Dung dịch chuẩn HCl 0,1N là của Merck (Đức). Tất cả các hóa chất đều đạt tiêu chuẩn tinh khiết phân tích.

2.2. Phương pháp nghiên cứu

2.2.1. Kỹ thuật NIRS



Hình 1. Sơ đồ xây dựng mô hình chẩn đoán thành phần Protein trên thiết bị NIRS

2.2.2. Phương pháp phân tích Protein

Hàm lượng protein tổng số của các loài cá trong nghiên cứu này được xác định theo AOAC (2001.11).

2.2.3. Phương pháp làm mịn dữ liệu bằng Savitzky - Golay

Bộ lọc Savitzky – Golay là một bộ lọc kỹ thuật số có thể được áp dụng cho một tập hợp các điểm dữ liệu kỹ thuật

số nhằm mục đích làm mịn dữ liệu, tức là để tăng độ chính xác của dữ liệu mà không làm sai lệch xu hướng tín hiệu. Điều này được thực hiện thông qua quá trình tích chập, bằng cách sử dụng phương pháp bình phương tối thiểu tuyến tính để phù hợp với các tập con liên tiếp của các điểm dữ liệu kề nhau với một đa thức bậc thấp. Thuật toán sẽ giúp ta ước tính bộ lọc tại tâm và đưa ra sự phù hợp với bậc đa thức ở điểm trung tâm [7].

2.2.4. Phương pháp làm mịn dữ liệu bằng hàm Deresolve

Hàm Deresolve được sử dụng để thay đổi độ phân giải biểu kiến của thiết bị đo, chuyển một phổ có độ phân giải cao thành một phổ có độ phân giải thấp bằng cách giảm mẫu tín hiệu. Deresolve cũng có thể được sử dụng để giảm nhiễu. Deresolve sử dụng bộ lọc hạt nhân tam giác để làm mịn dữ liệu bằng cách tích chập các phổ với một hàm độ phân giải, nhằm làm cho chúng trông như đã được thu thập trên một thiết bị có độ phân giải thấp hơn. Đây là phương pháp làm giảm tín hiệu nhiễu nhưng ít thay đổi đường cong cơ sở nhất [8].

2.2.5. Phương pháp lấy đạo hàm

Đạo hàm là một kỹ thuật cổ điển, được sử dụng rộng rãi cho các ứng dụng quang phổ. Một số thông tin "ẩn" trong quang phổ có thể dễ dàng được tiết lộ hơn khi làm việc trên đạo hàm bậc nhất hoặc bậc 2.

Đạo hàm bậc nhất của quang phổ chỉ đơn giản là số đo độ dốc của đường cong quang phổ tại mọi điểm. Độ dốc của đường cong không bị ảnh hưởng bởi các hiệu số đường cơ sở hoàn toàn cộng tính trong phổ, và do đó đạo hàm bậc nhất là một phương pháp rất hiệu quả để loại bỏ các hiệu số đó. Tuy nhiên, các đỉnh trong phổ thô thường trở thành điểm cắt 0 trong phổ đạo hàm bậc nhất, điều này có thể khó giải thích.

Đạo hàm bậc hai là thước đo sự thay đổi độ dốc của đường cong. Ngoài việc loại bỏ độ lệch ra khỏi đường nền, nó không bị ảnh hưởng bởi bất kỳ "độ lệch" tuyến tính nào có thể tồn tại trong dữ liệu, và do đó là một phương pháp rất hiệu quả để loại bỏ cả phần độ lệch đường nền và độ dốc từ một phổ. Đạo hàm bậc hai có thể giúp phân giải các đỉnh lân cận và làm sắc nét các đặc điểm quang phổ [8].

2.2.6. Phương pháp PCR

Phương pháp hồi quy thành phần chính (PCR) là một phương pháp để liên hệ phương sai của một biến phản hồi (biến Y) với phương sai của một số yếu tố dự đoán (biến X), với mục đích giải thích hoặc dự đoán.

PCR là quy trình gồm hai bước, đầu tiên phân tách ma trận biến dự đoán X bằng PCA (phân tích thành phần chính), sau đó phù hợp với mô hình MLR (hồi quy tuyến tính nhiều biến), sử dụng điểm PC (thành phần chính) thay vì các biến X ban đầu làm yếu tố dự đoán. Phương pháp này thực hiện đặc biệt tốt khi các biến X khác nhau thể hiện thông tin chung, tức là khi có một lượng lớn tương quan, hoặc thậm chí là cộng tác [8].

2.2.7. Phương pháp PLS

Hồi quy bình phương một phần nhỏ nhất (PLSR) đôi khi cũng được gọi là phép chiếu cấu trúc tiềm ẩn hoặc chỉ đơn giản gọi là PLS, phương pháp này bao gồm đồng thời cả 2 mô hình ma trận X và ma trận Y để tìm các biến tiềm

ân (hoặc ản) trong X dự đoán tốt nhất các biến tiềm ẩn trong Y. Các thành phần PLS này tương tự như các thành phần chính của phương pháp PCR, nhưng sẽ được gọi là yếu tố. PLSR có thể được thực hiện với một hoặc nhiều biến Y, có nghĩa là có thể sử dụng nhiều phản hồi Y trong mô hình hồi quy [8].

2.2.8. Phương pháp xử lý số liệu

Các số liệu thí nghiệm được xử lý bằng phần mềm Microsoft Excel 2019 và phần mềm Unscrambler v10.4.

3. Kết quả và thảo luận

3.1. Kết quả chạy PCR và PLS trên dữ liệu thô

Khi chạy PCR và PLS trên dữ liệu thô chưa qua các phương pháp tiền xử lý dữ liệu cho kết quả hầu hết có R-square dưới 0,9, duy nhất ở vị trí k cho kết quả R-square Cal là 0,9286 tuy nhiên R-square Val lại khá thấp khi chỉ đạt 0,6702. Các vị trí khác cho kết quả tương đối cao là vị trí a, b, c, d, i, l. Tuy nhiên, giá trị R-square chưa đạt như mong muốn ban đầu. Các vị trí e, f, g, h cho kết quả rất thấp và gần như có thể xem là không có tương quan, do đây là các vị trí ở mang và mắt khi đo cho số liệu nhiều rất nhiều.

Bảng 2. Kết quả của phương pháp PCR và PLS trên dữ liệu thô

STT	Vị trí	PC	Phương pháp PCR				Phương pháp PLS				
			R-square		RMSE		F	R-square		RMSE	
			Cal	Val	Cal	Val		Cal	Val	Cal	Val
1	a	9	0,8509	0,7624	0,6659	0,8835	6	0,8563	0,7431	0,6537	0,9257
2	b	4	0,7675	0,7210	0,8318	0,9587	6	0,8795	0,7370	0,5985	0,8873
3	c	9	0,8327	0,7364	0,7055	0,9143	8	0,8642	0,7207	0,6356	0,9563
4	d	7	0,7321	0,5932	0,8929	1,1513	5	0,7602	0,6589	0,8447	1,0504
5	e	4	0,6536	0,5923	1,0152	1,1523	4	0,7083	0,5719	0,9316	1,1088
6	f	4	0,6921	0,6512	0,9571	1,0569	3	0,7271	0,6640	0,9011	1,0406
7	g	5	0,5440	0,4580	1,1647	1,3405	3	0,5780	0,2682	1,1205	1,4564
8	h	6	0,6152	0,4974	1,0699	1,2752	4	0,6848	0,4682	0,9684	1,2690
9	i	4	0,6784	0,6053	0,9782	1,1303	4	0,7329	0,6030	0,8912	1,1278
10	j	5	0,6890	0,5346	0,9619	0,5346	3	0,6737	0,5449	0,9853	1,2255
11	k	8	0,7703	0,5633	0,8267	1,1286	11	0,9286	0,6702	0,4610	1,0402
12	l	7	0,7994	0,6496	0,7725	1,0749	4	0,7823	0,6129	0,8048	1,1320

PC: Số nhân tố chính cho phương pháp PCR;

F: Số yếu tố tối ưu cho phương pháp PLS;

R-square: Hệ số tương quan của mô hình hồi quy;

RMSE: Lỗi bình phương trung bình gốc;

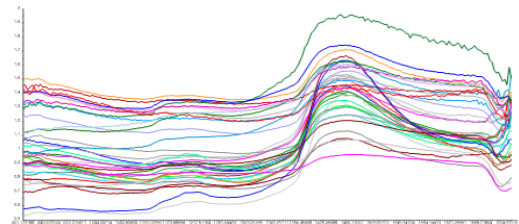
Cal: Mô hình dùng để hiệu chuẩn sai số của phép đo;

Val: Mô hình dùng để thẩm định sai số của phép đo.

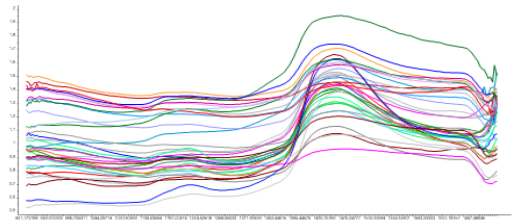
3.2. Kết quả chạy PCR và PLS kết hợp phương pháp Savitzky – Golay

Hình 2, có thể thấy kết quả sau khi làm mịn thì đồ thị phổ cho các đường cong trơn hơn và ít bị nhiễu như ban đầu. Việc làm mịn sẽ giúp ta loại bỏ các tín hiệu gây nhiễu và làm ổn định bộ dữ liệu. Tuy nhiên phương pháp này không giúp ta loại bỏ các đường có độ lệch lớn cũng như độ nhiễu của các bước sóng ở đầu và cuối đồ thị phổ vẫn chưa được giải quyết.

Tiền hành chạy PCR và PLS dựa vào bộ dữ liệu đã được xử lý theo phương pháp làm mịn bằng bộ lọc Savitzky – Golay trên 12 vị trí của 37 mẫu cá thu được kết quả như Bảng 3.



(a)



(b)

Hình 2. Đồ thị phổ trước (a) và sau (b) khi làm mịn bằng bộ lọc Savitzky- Golay

Bảng 3. Kết quả của phương pháp PCR và PLS trên dữ liệu được làm mịn bằng bộ lọc Savitzky - Golay

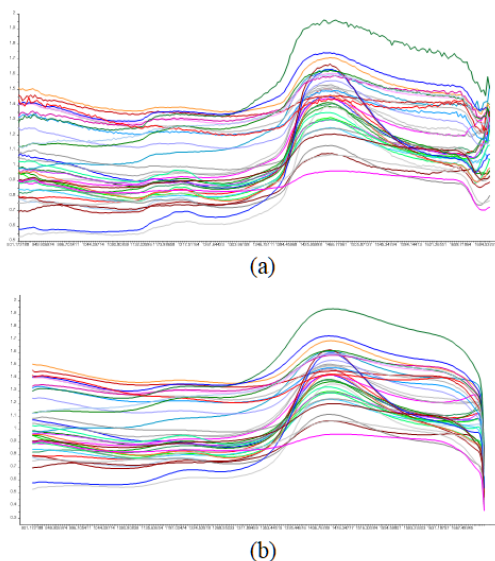
STT	Vị trí	PC	Phương pháp PCR				F	Phương pháp PLS			
			R-square		RMSE			R-square		RMSE	
			Cal	Val	Cal	Val		Cal	Val	Cal	Val
1	a	7	0,8373	0,7240	0,6956	0,9156	7	0,8686	0,7199	0,6486	0,9188
2	b	7	0,8547	0,7392	0,6575	0,8873	6	0,8794	0,7440	0,5989	0,9147
3	c	10	0,8374	0,7032	0,6955	0,9685	7	0,8483	0,6952	0,6718	0,9213
4	d	7	0,7396	0,6311	0,8801	1,0642	5	0,7574	0,6638	0,8495	1,0216
5	e	4	0,6471	0,5949	1,0100	1,1312	4	0,7052	0,6016	0,9365	1,1407
6	f	3	0,6973	0,6703	0,9494	1,0366	3	0,7271	0,6398	0,9010	1,0433
7	g	5	0,5464	0,3344	1,1618	1,4485	3	0,5735	0,2920	1,1264	1,4302
8	h	5	0,6256	0,5065	1,0554	1,2745	4	0,6786	0,4347	0,9778	1,2719
9	i	7	0,7478	0,6136	0,8663	1,0677	4	0,7306	0,6253	0,8953	1,1247
10	j	5	0,6865	0,5077	0,9657	1,2404	3	0,6729	0,5541	0,9864	1,2021
11	k	5	0,6732	0,5812	0,9860	1,1295	4	0,7754	0,6221	0,8174	1,1192
12	l	6	0,8094	0,6553	0,7530	1,0760	4	0,8266	0,5992	0,7182	1,1022

Nhìn chung kết quả chạy PCR và PLS trên dữ liệu đã được làm mịn bằng bộ lọc Savitzky-Golay khá tương đương so với chạy trên dữ liệu thô. Một vài vị trí có R-square val tăng lên một ít nhưng không cao như vị trí a, b, c, e, g, i. Các vị trí d, f, h, k, l có sự giảm nhẹ. Đồng thời một vài vị trí còn tăng RMSE Val. Cụ thể tại vị trí a cho kết quả R-square Cal giảm từ 0,8509 xuống còn 0,8373, R-square Val cũng giảm từ 0,7626 còn 0,7240, hệ số RMSE Cal lại tăng lên từ 0,6659 thành 0,6956 và RMSE Val cũng tăng từ 0,8835 lên thành 0,9156. Điều này cho thấy bộ lọc làm giảm tương quan của mô hình hồi quy và tăng độ sai số của mô hình tại vị trí a.

3.3. Kết quả chạy PCR và PLS kết hợp phương pháp Deresolve

Sau khi chạy Deresolve, đồ thị phổ ít bị thay đổi đường cong cơ sở (Hình 3). Độ mịn của các đường cong trơn hơn so với ban đầu như phương pháp làm mịn bằng bộ lọc Savitzky – Golay, và hơn thế phương pháp này giải quyết được các tín hiệu nhiễu của các bước sóng ở đầu và cuối đồ thị phổ.

Sau khi chạy phương pháp PCR và PLS cho 12 vị trí trên 37 mẫu cá với bộ dữ liệu đã được xử lý bằng phương pháp Deresolve thì thu được kết quả như Bảng 4:



Hình 3. Đồ thị phổ trước (a) và sau (b) khi chạy Deresolve

Bảng 4. Kết quả của phương pháp PCR và PLS trên dữ liệu xử lý bằng hàm Deresolve

STT	Vị trí	Phương pháp PCR				F	Phương pháp PLS				
		R-square		RMSE			R-square		RMSE		
		Cal	Val	Cal	Val		Cal	Val	Cal	Val	
1	a	8	0,8518	0,7193	0,6639	0,8798	7	0,8574	0,7287	0,6513	0,9315
2	b	8	0,8738	0,7876	0,6127	0,8513	6	0,8717	0,7685	0,6176	0,8860
3	c	8	0,8373	0,7221	0,6958	0,9234	6	0,8248	0,7188	0,7219	0,9736
4	d	9	0,7617	0,5772	0,8419	1,1319	6	0,7634	0,6123	0,8389	1,0460
5	e	4	0,6666	0,5877	0,9959	1,1316	7	0,7846	0,6423	0,8004	1,1144
6	f	3	0,7034	0,6734	0,9393	1,0272	3	0,7200	0,6331	0,9126	1,0563
7	g	10	0,6791	0,4442	0,9771	1,3182	7	0,6926	0,4327	0,9564	1,3414
8	h	9	0,7014	0,4883	0,9426	1,2886	4	0,6659	0,4628	0,9970	1,2919
9	i	4	0,7258	0,6531	0,9032	1,0410	4	0,7381	0,6388	0,8826	1,0673
10	j	5	0,6644	0,5044	0,9992	1,2661	8	0,8494	0,6266	0,6694	1,1038
11	k	12	0,9017	0,7893	0,5406	0,8005	7	0,8618	0,7544	0,6410	0,8935
12	l	6	0,8197	0,6956	0,7323	0,9974	4	0,8187	0,7222	0,7343	0,9736

Nhận xét chung thấy so với Chạy PCR và PLS trên dữ liệu thô và so với tiền xử lý bằng bộ lọc Savitzky – Golay thì chạy PCR và PLS sau khi xử lý Deresolve cho kết quả cao trên hầu hết vị trí, cả R-square Cal lẫn R-square Val đều tăng nhưng không làm tăng RMSE thậm chí còn làm giảm RMSE chứng tỏ phương pháp này giúp tăng hệ số tương quan và làm giảm sai số của phép đo.

Sau khi chạy PCR và PLS đã được xử lý bằng 2 phương pháp trên. Ta nhận thấy rằng khi sử dụng hàm Deresolve để loại tín hiệu nhiễu rồi sau đó mới chạy PCR và PLS sẽ làm tăng hệ số tương quan R-square và giảm sai số của phép đo khi giảm RMSE. Tuy nhiên R-square Val còn thấp (tất cả vị trí đều nhỏ hơn 0,8) cho thấy dữ liệu thô chưa mô tả được sự tương quan giữa quang phổ cận hồng ngoại và thành phần Protein của thủy sản. Do đó ta sẽ tiến hành lấy đạo hàm bậc 1 và bậc 2 sau khi đã xử lý tín hiệu nhiễu bằng hàm Deresolve và so sánh với nhau.

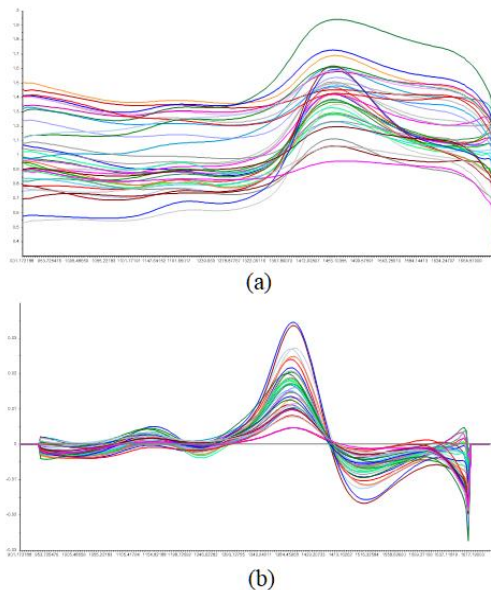
3.4. Kết quả chạy PCR và PLS kết hợp phương pháp lấy đạo hàm bậc nhất

Sau khi lấy đạo hàm bậc nhất, đồ thị phổ thể hiện rõ độ dốc của những đường cong, thể hiện những đỉnh bị ẩn trong dữ liệu thô ban đầu. Và các đỉnh ở dữ liệu ban đầu sau khi đạo hàm thành những điểm cắt 0. Các đỉnh thể hiện rõ ở

các bước sóng 1142nm, 1232nm, 1396nm, 1402nm, 1505nm và 1670nm.

Sau khi chạy phương pháp PCR và PLS trên 12 vị trí của 37 mẫu cá với bộ dữ liệu đã được xử lý bằng phương pháp Deresolve và lấy đạo hàm bậc nhất thu được kết quả như Bảng 5.

Các vị trí có kết quả cao khi chạy PCR và PLS ở các trường hợp trước như vị trí a, b, c thì trong trường hợp này đều tăng giá trị R-square Val và RMSE không thay đổi nhiều. Đối với phương pháp PLS thì các vị trí này không chỉ tăng giá trị R-square cao mà còn giảm giá trị RMSE. Các vị trí này có mối tương quan cao đối với thành phần protein, đó là các vị trí gáy, lưng và đuôi. Kết quả cao nhất đối với phương pháp PLS tại vị trí a, với số yếu tố chính là 16, R-square Cal là 0,968, R-square Val là 0,8592, RMSE Cal là 0,3082, RMSE Val là 0,6948.



Hình 4. Đồ thị phổ trước (a) và sau (b) khi lấy đạo hàm bậc nhất

Bảng 5. Kết quả của phương pháp PCR và PLS kết hợp lấy đạo hàm bậc nhất

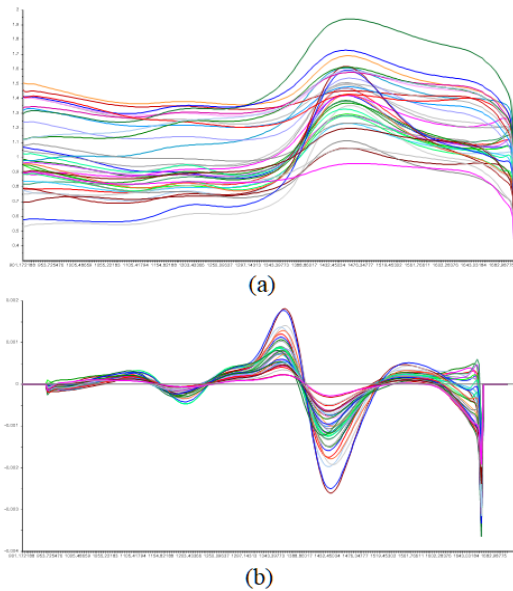
STT	Vị trí	Phương pháp PCR				F	Phương pháp PLS				
		R-square		RMSE			R-square		RMSE		
		Cal	Val	Cal	Val		Cal	Val	Cal	Val	
1	a	10	0,8562	0,7246	0,6539	0,9253	16	0,9680	0,8592	0,3082	0,6948
2	b	7	0,8594	0,7582	0,6467	0,8874	4	0,8299	0,7193	0,7113	0,9261
3	c	16	0,9198	0,7001	0,4884	0,9674	12	0,9245	0,7201	0,4738	0,9398
4	d	8	0,7165	0,4899	0,9183	1,2310	5	0,7329	0,5539	0,8914	1,2508
5	e	4	0,6795	0,6351	0,9765	1,1231	3	0,6932	0,6080	0,9554	1,1290
6	f	4	0,7104	0,5589	0,9283	1,1462	3	0,7154	0,5753	0,9201	1,1670
7	g	5	0,5589	0,3995	1,1456	1,3741	4	0,6340	0,3390	1,0436	1,4494
8	h	4	0,5677	0,4411	1,1341	1,3068	3	0,6509	0,4073	1,0191	1,3338
9	i	3	0,7445	0,7259	0,8718	0,9585	3	0,7583	0,6956	0,8481	0,9827
10	j	13	0,8741	0,7096	0,6120	0,9424	7	0,8662	0,7250	0,6307	0,9538
11	k	6	0,8288	0,7529	0,7135	0,8941	5	0,8585	0,7369	0,6488	0,8981
12	l	11	0,8587	0,6682	0,6483	1,0260	7	0,8758	0,7900	0,6077	0,8519

Các vị trí có kết quả thấp khi chạy PCR và PLS ở các trường hợp trước như d, e, f, g, h thì trong trường hợp này giá trị R-square đều giảm. Chứng tỏ những vị trí này không có mối tương quan cao đối với thành phần protein. Đó là các vị trí bụng, mang và mắt.

Các vị trí i, j khi chạy PCR và PLS trong các trường hợp trước cho kết quả không quá cao cũng không quá thấp nhưng trong trường hợp này thì tăng cao cho thấy khi lấy đạo hàm thì các biến tiềm ẩn trong bộ dữ liệu lộ ra, các vị trí k và l có độ thay đổi thấp trong các trường hợp. Điều này cho thấy các vị trí này có mối tương quan với thành phần Protein. Đó là các vị trí thịt gáy, thịt lưng, thịt đuôi, và thịt bụng.

3.5. Kết quả chạy PCR và PLS kết hợp phương pháp lấy đạo hàm bậc hai

Vì đạo hàm bậc 2 là hệ số góc của đạo hàm bậc nhất nên những điểm cực đại trong đạo hàm bậc nhất lại là cực tiểu trong đạo hàm bậc 2 và trường hợp này sẽ dẫn đến cực đại của dữ liệu thô ban đầu là cực tiểu trong đạo hàm bậc 2.



Hình 5. Đồ thị phổ trước (a) và sau (b) khi đạo hàm bậc 2

Đồ thị phổ của đạo hàm bậc 2 xuất hiện thêm nhiều đỉnh tại các bước sóng 1124nm, 1158nm, 1206nm, 1210nm, 1250nm, 1310nm, 1374nm, 1381nm, 1419nm, 1422nm, 1658nm. Sau khi chạy phương pháp PCR và PLS trên 12 vị trí của 37 mẫu cá với bộ dữ liệu đã được xử lý bằng phương pháp Deresolve và lấy đạo hàm bậc 2 thu được kết quả như Bảng 6.

Bảng 6. Kết quả của phương pháp PCR và PLS kết hợp lấy đạo hàm bậc 2

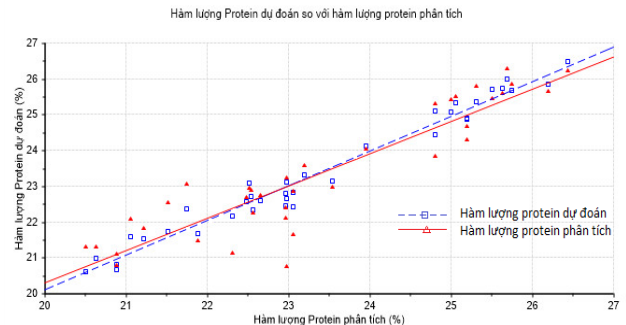
STT	Vị trí	Phương pháp PCR				F	Phương pháp PLS				
		R-square		RMSE			R-square		RMSE		
		Cal	Val	Cal	Val		Cal	Val	Cal	Val	
1	a	5	0,8215	0,7808	0,7287	0,8323	13	0,9645	0,8539	0,3250	0,6820
2	b	4	0,7801	0,7381	0,8087	0,8969	4	0,8298	0,7262	0,7115	0,9134
3	c	17	0,9253	0,7445	0,4712	0,9050	7	0,8873	0,7665	0,5790	0,8881
4	d	4	0,6838	0,5981	0,9698	1,1488	4	0,7154	0,6769	0,9201	1,0291
5	e	11	0,8029	0,6659	0,7656	1,0710	3	0,6846	0,5615	0,9686	1,1556
6	f	10	0,8015	0,5530	0,7689	1,1355	7	0,8148	0,5118	0,7422	1,2329
7	g	4	0,5677	0,4127	1,1341	1,3301	3	0,6509	0,4129	1,0191	1,3646
8	h	10	0,7000	0,4120	0,9447	1,3854	6	0,7000	0,3411	0,9447	1,4420
9	i	3	0,7445	0,7061	0,8718	0,9690	3	0,7582	0,6419	0,8481	0,9925
10	j	9	0,8364	0,6528	0,6975	1,0330	3	0,7618	0,6611	0,8418	1,0425
11	k	7	0,8160	0,7520	0,7397	0,8847	6	0,8412	0,7493	0,6874	0,9077
12	l	10	0,8567	0,7390	0,6528	0,8619	6	0,8721	0,7472	0,6168	0,9139

So sánh với kết quả sau khi lấy đạo hàm bậc nhất thì đạo hàm bậc 2 cho kết quả thấp hơn. Các vị trí có kết quả cao ở đạo hàm bậc 1 như vị trí a, b, c thì ở đạo hàm bậc hai R-square giảm có giảm đôi chút, tuy nhiên RMSE có giảm nhẹ chứng tỏ tuy hệ số tương quan thấp hơn 1 tí nhưng bù lại sai số của phép đo được giảm.

Các vị trí khác d, e, f, g, h ở đạo hàm bậc nhất đã được đánh giá không có tương quan thì ở đạo hàm bậc 2 có tăng R-square Cal nhưng R-square Val không tăng, chứng tỏ khi lấy đạo hàm bậc hai thì các biến ẩn bên trong được làm lộ rõ ra khiến chúng tương quan với các biến tiềm ẩn nhưng chúng vẫn không tương quan với thành phần Protein.

3.6. Đánh giá kết quả thu được

Kết quả chạy PCR và PLS sau khi lấy đạo hàm bậc nhất và bậc 2 cho thấy rằng trong 2 phương pháp, phương pháp PLS cho kết quả tốt hơn so với phương pháp PCR khi cho các giá trị R-square trong ma trận chuẩn tốt và đạt yêu cầu ban đầu đề ra khi hệ số tương quan R-square > 0,9. Cụ thể vị trí cho kết quả tốt nhất là vị trí a – tương ứng vị trí gáy, ngoài da. Cho kết quả R-square Cal = 0,9645 và R-square Val = 0,8539, RMSE Cal = 0,3250, RMSE Val = 0,6820 với số yếu tố chính bằng 16. Do vậy nhóm nghiên cứu quyết định định lượng Protein của cá bằng phương pháp phổ hồng ngoại quét tại vị trí gáy, ngoài da kết hợp với thuật toán hồi quy đa biến PLS với số yếu tố chính bằng 16.



Hình 6. Đường hồi quy giữa thành phần Protein phân tích bằng phương pháp chuẩn trong phòng thí nghiệm và dự đoán trên NIRS

3.7. Kiểm định lại mô hình hồi quy

Tiến hành sử dụng 10 mẫu chuẩn làm ma trận kiểm định như theo quy trình chuẩn bị, xử lý mẫu và xử lý dữ liệu, kết quả ma trận kiểm định có kích thước 10x230. Sau đó tiến hành sử dụng mô hình hồi quy PLS đã chọn như trên tiến hành dự đoán hàm lượng Protein của ma trận kiểm định

Sau khi chạy mô hình hồi quy PLS đã chọn để dự đoán hàm lượng Protein 10 mẫu kiểm định thì cho ra kết quả như Bảng 7.

Từ các kết quả thu được, nhóm tác giả nhận thấy hàm lượng Protein dự đoán theo mô hình hồi quy đa biến PLS so với phương pháp phân tích tiêu chuẩn sai khác đều dưới 10%. Sai số tương đối hàm lượng Protein giữa phương pháp cận hồng ngoại và phương pháp đối chứng là từ 0,06 – 8,75%. Từ sự chênh lệch giữa các phương pháp phân tích cho ta thấy rằng phương pháp phổ cận hồng ngoại kết hợp với thuật toán hồi quy đa biến tỏ ra khá ưu việt khi áp dụng vào phân tích định lượng Protein.

Bảng 7. Hàm lượng Protein (%) dự đoán bằng mô hình hồi quy PLS so với phương pháp đối chứng

Mẫu kiểm định	Loại cá	Phương pháp cận hồng ngoại		Phương pháp đối chứng		Sai số tương đối giữa 2 phương pháp (%)
		Hàm lượng	Phương sai	Hàm lượng	Độ lệch chuẩn (%)	
417	Nục	24,2336	0,9681639	24,65	1,83	1,69
568	Ngừ	22,18286	0,4294998	24,31	1,83	8,75
652	Nục	23,05189	0,4551244	24,25	1,83	4,94
672	Ngừ	23,08494	0,5410158	23,88	1,83	3,33
779	Ngừ	22,30868	0,8891348	24,25	1,83	8,01
833	Nục	22,92109	0,4271629	23,45	1,83	2,26
638	Ngừ	24,32461	0,6872587	24,31	1,83	0,06
834	Ngừ	23,09506	0,6189303	23,19	1,83	0,41
839	Ngừ	24,74111	0,5686936	23	1,83	7,57
406	Nục	22,471	0,657811	21,35	1,83	5,25

4. Kết luận

Nghiên cứu đã xác định được mô hình dự đoán tối ưu hàm lượng Protein trong cá. Cụ thể đầu tiên xử lý bằng phương pháp Deresolve, sau đó lấy đạo hàm bậc nhất và chạy mô hình hồi quy PLS. Nghiên cứu cho thấy, kết quả sau khi lấy đạo hàm bậc nhất cao hơn so với khi lấy đạo hàm bậc 2. Phương pháp PLS cho hệ số tương quan cao hơn so với phương pháp PCR.

Từ sự chênh lệch giữa các phương pháp phân tích cho ta thấy rằng phương pháp phổ cận hồng ngoại kết hợp với thuật toán hồi quy đa biến tỏ ra khá ưu việt khi áp dụng vào phân tích định lượng protein khi sai số so với phương pháp chuẩn chỉ nằm trong khoảng 0,06 – 8,75%.

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: ĐTDL.CN-33/20, Bộ Khoa học & Công nghệ.

TÀI LIỆU THAM KHẢO

- [1] D. I. Givens, J. L. D. Boever, E. R. Deaville. “The principles, practices and some future applications of near infrared spectroscopy for predicting the nutritive value of foods for animals and humans”, *Nutr Res Rev*, vol. 10, no. 1, pp. 83-114, 1997.
- [2] G. McL Dryden, *Near infrared reflectance spectroscopy: applications in deer nutrition*, Rural Industries Research and Development Corporation, 2003.
- [3] K. H. Norris and J. R. Hart, *Direct spectrophotometric determination of moisture content of grain and seeds*. In Humidity and Moisture, Measurement and Control in Science and Industry. Van Nostrand-Reinhold, New York, 1965.
- [4] J. Zhou, X. Wu, J. You, and S. Xiong, “Rapid determination of the textural properties of silver carp (*Hypophthalmichthys molitrix*) using near-infrared reflectance spectroscopy and chemometrics”. *LWT*, vol. 129, p. 109545, 2020.
- [5] M.M. Reis, E. Martínez, E. Saitua, R. Rodríguez, I. Pérez, I. Olabarrieta, “Non-invasive differentiation between fresh and frozen/thawed tuna fillets using near infrared spectroscopy (Vis-NIRS)”, *LWT*, vol. 78, pp. 129-137, 2017.
- [6] H. Lv, W. Xu, J. You, and S Xiong, “Classification of freshwater fish species by linear discriminant analysis based on near infrared reflectance spectroscopy”. *Journal of Near Infrared Spectroscopy*, vol. 25, no. 1, pp. 54-62, 2017.
- [7] A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures”, *Anal. Chem*, vol. 36, no. 8, pp. 1627-1639, 1964.
- [8] CAMO Software. “The Unscrambler® X v10.3 User Manual”, 2014.