

ỨNG DỤNG PHỔ CẬN HỒNG NGOẠI NIR KẾT HỢP VỚI CÁC PHƯƠNG PHÁP XỬ LÝ SỐ LIỆU CHEMOMETRICS ĐỂ XÁC THỰC NHANH CÀ PHÊ ĐẠT CHUẨN CHỈ DẪN ĐỊA LÝ Ở VÙNG BUÔN MA THUỘT

Tô Phan Chiêu Đan^{1,2}, Nguyễn Quốc Cường^{1,2}, Lê Nguyễn Đoàn Duy^{3,*}

TÓM TẮT

Cà phê là một mặt hàng quan trọng, chiếm 3% GDP của Việt Nam và kim ngạch xuất khẩu cà phê của Việt Nam đã đạt trên 3 tỷ đô la Mỹ trong nhiều năm. Trong những năm gần đây, mặc dù kinh tế thế giới trải qua nhiều giai đoạn khó khăn dẫn đến sức mua giảm nhưng xuất khẩu cà phê của Việt Nam vẫn duy trì được tốc độ tăng trưởng đáng khích lệ, do đó để đảm bảo chuỗi cung ứng cà phê cần có biện pháp bảo vệ thương hiệu cà phê và ngăn chặn các hành vi gian lận về chất lượng sản phẩm. Nghiên cứu nhằm sử dụng phổ cận hồng ngoại NIR cùng với các phương pháp xử lý số liệu Chemometrics để xác thực và bảo vệ thương hiệu của cà phê đạt chỉ dẫn địa lý (GI) vùng Buôn Ma Thuột. Trong nghiên cứu, số mẫu cà phê nhân đạt chuẩn GI (n = 49) và cà phê nhân không đạt chuẩn GI (n = 103) bằng cách phân tích trực tiếp mà không cần phá hủy mẫu. Sau đó, PLS-DA và SIMCA được sử dụng để xây dựng các mô hình. Mô hình kết hợp PLS-DA với tiền xử lý MC, SNV cho kết quả tốt nhất, độ chính xác của mô hình 87,96%, độ nhạy 81,25% và độ đặc hiệu 95,23%. Đồng thời, mô hình kết hợp SIMCA với tiền xử lý MC, 2nd Der, SNV đạt được độ chính xác của mô hình chỉ 85,76%. Do đó, các phương pháp đề xuất có thể hữu ích cho cả người tiêu dùng và cơ quan quản lý trong việc đăng ký GI cho sản phẩm vì nó khẳng định tiêu chuẩn chất lượng của cà phê đặc sản Đắk Lắk (GI), ngăn ngừa gian lận nhãn mác, phương pháp thân thiện với môi trường, kỹ thuật nhanh chóng, tiết kiệm thời gian và tiền bạc.

Từ khóa: Cà phê, chemometrics, chỉ dẫn địa lý, phổ cận hồng ngoại.

1. ĐẶT VẤN ĐỀ

Việt Nam là nước đứng thứ hai trên thế giới về sản xuất cà phê, đóng góp 3% GDP cả nước, xuất khẩu trung bình 3 tỷ USD mỗi năm. Cà phê Việt Nam đã xuất khẩu đến hơn 80 đất nước và vùng lãnh thổ, chiếm 14,2% tổng khối lượng cà phê nhân xuất khẩu trên toàn thế giới, đứng thứ hai sau Brazil. Đa phần cà phê ở Việt Nam là cà phê Robusta, có vị đắng hơn, màu sẫm hơn so với Arabica. Chất lượng cà phê chịu ảnh hưởng bởi giống, cách canh tác và cách thu hoạch [1].

Ở một quốc gia có nhiều lợi thế về các mặt hàng nông sản nhưng lại chưa có nhiều phương thức chế biến đa dạng như Việt Nam, chỉ dẫn địa lý đã trở thành một trong những yếu tố quan trọng giúp nâng cao giá trị cho nông sản trên thị trường quốc tế. Chỉ

dẫn địa lý (Geography Indicator - GI) chứa thông tin về nguồn gốc của hàng hoá: từ ngữ, dấu hiệu, biểu tượng; hình ảnh để chỉ: một quốc gia, một vùng lãnh thổ, một địa phương mà hàng hoá được sản xuất ra từ đó. Chất lượng, uy tín, danh tiếng của hàng hoá là do nguồn gốc địa lý tạo nên. Tính đến tháng 1 năm 2019, 346 GI đã được đăng ký ở các nước Đông Nam Á, Việt Nam đã đăng ký 69 GI cho các sản phẩm, ví dụ như: cà phê Buôn Ma Thuột (2005), trà tuyết san Mộc Châu (2010), nước mắm Phú Quốc (2012), cà phê Sơn La (2017), hạt điều Bình Phước (2018)...[2].

Để được bảo hộ chỉ dẫn địa lý thì các sản phẩm phải có danh tiếng, chất lượng hoặc đặc tính mà chủ yếu là do điều kiện địa lý mang lại. Điển hình như chất lượng cà phê Buôn Ma Thuột đăng ký GI phải đảm bảo các yếu tố: màu hạt: màu xám, màu xanh hoặc xám lục nhẹ; cỡ hạt: dài 10-11 mm, rộng 6-7 mm, dày 3-4 mm; độ axit thấp, hàm lượng caffein từ 2,2-2,4%. Chỉ các hạt cà phê đạt chuẩn về trồng trọt và thu hoạch ở khu vực Cư M'gar, Ea H'leo, Krông Ana, Cư Kưin, Krông Buk, Krông Năng, Krông Pak, Buôn Hồ; thành phố Buôn Ma Thuột (Cư Kưin được tách từ Krông An; Buôn Hồ được tách từ Krông Buk)

¹ Bộ môn Công nghệ Thực phẩm, Khoa Kỹ thuật Hóa học, Trường Đại học Bách khoa thành phố Hồ Chí Minh

² Đại học Quốc gia thành phố TP. Hồ Chí Minh

³ Khoa Công nghệ Thực phẩm, Trường Đại học Công nghiệp thực phẩm thành phố TP. Hồ Chí Minh

*Email: duyln@hufi.edu.vn

đạt chuẩn về chất lượng nêu trên mới được cấp chỉ dẫn địa lý GI [3].

Tuy nhiên, chất lượng cà phê có chỉ dẫn địa lý có thể bị giả mạo trong quá trình thương mại và nhằm hỗ trợ nhanh trong bước định danh các hạt cà phê thuộc các vùng địa lý trên trong quá trình đăng kí làm hồ sơ chỉ dẫn địa lý, cần có một phương pháp để xây dựng mô hình xác định nhanh cà phê đạt chuẩn GI so với các loại cà phê khác.

Việc sử dụng kỹ thuật phổ hồng ngoại để giải quyết vấn đề xác thực nguồn gốc và phát hiện thực phẩm giả mạo có được ưu thế nhờ nhanh chóng, rẻ tiền, không sử dụng hóa chất nên không ảnh hưởng đến môi trường, cũng như không đòi hỏi việc chuẩn bị mẫu [4].

Mục tiêu của nghiên cứu là sử dụng thiết bị quang phổ cận hồng ngoại cầm tay kết hợp với phương pháp xử lý đa biến để xác thực nhanh, phân biệt cà phê đạt tiêu chuẩn chất lượng GI so với các loại cà phê trên thị trường trong quá trình đăng kí chỉ dẫn địa lý và trong quá trình thương mại. Đây là một nghiên cứu hoàn toàn mới ở Việt Nam và có thể hỗ trợ cho việc xây dựng thương hiệu cà phê Việt, góp phần gia tăng giá trị của cà phê Việt Nam.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Vật liệu nghiên cứu

2.1.1. Vật liệu thí nghiệm

Lấy 152 mẫu cà phê nhân khô Robusta đại diện cho các vùng địa lý của các tỉnh ở Tây Nguyên như: Đắk Lắk, Đắk Nông, Gia Lai, Lâm Đồng. Các mẫu lấy trong khu vực địa phương rải đều các thôn trong xã,

khoảng cách giữa các điểm lấy mẫu phải cách nhau ít nhất 2 km, tại mỗi điểm lấy 3 mẫu lặp lại. Điều kiện về phương pháp chế biến, công đoạn phơi/sấy nguyên liệu, quả cà phê (đối với phương pháp chế biến khô)/hạt cà phê nhân (đối với phương pháp chế biến nửa ướt) được phơi trong thời tiết nắng nhẹ, trời làm râm, không mưa, nhiệt độ không khí dưới 35°C/hoặc sấy không quá 75°C và điều chỉnh trong quá trình sấy cho đến khi đạt độ ẩm khoảng từ 12-12,5%. Dùng máy xát để loại bỏ vỏ quả cà phê sau khi phơi/sấy và thu phần hạt cà phê nhân. Để biết hạt cà phê nhân có đạt độ ẩm mức như trên, tiến hành lấy mẫu, đem xay và đưa vào máy đo độ ẩm nông sản để đo. Các mẫu cà phê có điều kiện xử lý tương tự nhau, khác nhau đến từ phương pháp canh tác: điều kiện đất, nước, khí hậu ở địa phương...Điều kiện mẫu cà phê nhân thu thập: giống cà phê Robusta, thu hoạch từ tháng 10/2021 đến tháng 1/2022, đạt tiêu chuẩn về bề ngoài: kích cỡ 7 mm, cỡ sàng 18, lỗi loại 2 dưới 10 (phân loại theo hạt lỗi). Mỗi mẫu lấy lặp lại 3 lần, mỗi lần lấy 300 g cà phê nhân, bảo quản trong túi hút chân không tránh ánh sáng, bỏ thêm silicagel để tránh hút ẩm. Danh sách số lượng mẫu được trình bày ở bảng 1. Mẫu được chuẩn bị trong vòng 2 tháng và trong vòng 7 ngày chuyển đến Phòng thí nghiệm Công nghệ thực phẩm, Trường Đại học Bách khoa thành phố Hồ Chí Minh. Cà phê gửi đến được bảo quản ở nhiệt độ phòng, dao động từ 25-30°C. Mẫu được chia theo 2 nhóm, mẫu chuẩn và mẫu kiểm chứng tương ứng với tỉ lệ 3:1, số mẫu chuẩn là 115 mẫu, số mẫu kiểm chứng là 37 mẫu.

Bảng 1. Số lượng mẫu cà phê thu thập được

STT	Tỉnh	Huyện	Số mẫu chuẩn	Số mẫu kiểm chứng
1	Đắk Lắk	Buôn Ma Thuột	30	10
		Ea Hleo	7	2
2	Đắk Nông	Đắk Mil	10	3
3	Gia Lai	Chư Prông	5	1
		Chư Sê	14	4
		Pleiku	15	5
4	Lâm Đồng	Đức Trọng	17	6
		Đà Lạt	18	6
Tổng cộng		152 mẫu	115 mẫu	37 mẫu

2.1.2. Thiết bị nghiên cứu

Thiết bị cảm biến phổ cận hồng ngoại SCiO™ được dùng để quét phổ cận hồng ngoại của các mẫu, được Công ty Consumer Physics (Mỹ) chế tạo với dải phổ từ 740 nm đến 1070 nm ở độ phân giải 1 nm.

Thiết bị đã được Moli và cs (2018) [5] dùng để xây dựng các mô hình dự báo chất lượng của các loại quả kiwi, táo, bơ..., kết quả cho độ chính xác lên đến 100%. Bằng cách sử dụng những nguồn bức xạ có độ chọn lọc cao và các đầu dò đủ nhạy, phổ cận hồng ngoại đã thể hiện được vai trò của nó trong phân tích

với những ưu điểm về định lượng, định tính và cả nghiên cứu cấu tạo của vật chất. Thiết bị SCiO™ có giá thành rẻ, nhỏ gọn dễ sử dụng, tạo ra tiềm năng ứng dụng kỹ thuật NIR trong công nghiệp rộng rãi hơn, có khả năng phân loại và sàng lọc nhanh chóng để hỗ trợ trong việc dự đoán chất lượng, góp phần vào quy trình ra quyết định trong toàn bộ chuỗi cung ứng.

Nguyên lý hoạt động của thiết bị đo phổ cận hồng ngoại SCiO™: ánh sáng từ nguồn sáng được chiếu xuyên qua bộ lọc. Bộ lọc tách ánh sáng thành các bước sóng vùng cận hồng ngoại. Ánh sáng cận hồng ngoại sau đó được chiếu trực tiếp lên sản phẩm được đo ở bộ phận đèn chiếu như hình 1. Ánh sáng phản chiếu của sản phẩm sẽ được thu bởi bộ cảm biến phân tử SCiO và được chụp bằng một tấm gương cầu tích hợp (integrating sphere) và được tập trung vào một đầu dò (detector). Tín hiệu phổ ra từ đầu dò sau đó được xử lý nhờ vào mô hình hiệu chuẩn của thiết bị và sẽ được hiển thị trên màn hình ở dạng số hay dạng phổ.



Hình 1. Thiết bị quang phổ cận hồng ngoại cầm tay SCiO™

2.2. Phương pháp nghiên cứu

2.2.1. Nguyên lý ứng dụng phổ cận hồng ngoại NIR

Vùng quang phổ cận hồng ngoại là vùng quang phổ tồn tại từ cuối tia đỏ trong vùng ánh sáng nhìn thấy và bắt đầu vùng trung hồng ngoại, có bước sóng từ 750- 2500 nm ($12500-4000\text{ cm}^{-1}$). Quang phổ cận hồng ngoại (NIRS) dựa trên nguyên tắc đo sự hấp thụ, truyền qua và phản xạ khuếch tán toàn bộ vùng bước sóng 750-2500 nm. Phương pháp sử dụng quang phổ cận hồng ngoại là một phương pháp phân tích đã được ứng dụng trong phân tích thực phẩm, hóa chất từ những năm 1950 [6].

Vùng phổ cận hồng ngoại (NIR) xảy ra sự hấp thụ, là một chuỗi các rung động ở tần số thích hợp hoặc kết hợp của sự hấp thụ cơ bản. Do đó, vùng NIR thường được gọi là vùng âm bội (overtone). Các loại

hấp thụ chi phối NIR là các loại hấp thụ hydro như rung động $-OH$, $-NH$ và $-CH$. Những loại hấp thụ này được chi phối bởi độ ẩm và hầu như tất cả các thành phần chính khác của thực phẩm. Do các nhóm phân tử khác nhau được tạo thành từ các nguyên tử có khối lượng và cường độ liên kết khác nhau nên mỗi phân tử thể hiện một tập hợp hấp thụ duy nhất ở tần số dao động hoặc bước sóng cụ thể [6].

Kỹ thuật này kết hợp với sự phát triển của kỹ thuật chemometrics đã trở thành công cụ phân tích hiệu quả, nhanh chóng, có độ chính xác cao và đặc biệt không phá hủy mẫu, có thể dùng để nghiên cứu tính chất, cấu tạo hoặc định lượng các hợp chất hữu cơ.

2.2.2. Chemometrics

Chemometrics là một kỹ thuật thông tin chiết xuất hóa học có liên quan đến các dữ liệu được tạo ra trong các thí nghiệm hóa học với việc sử dụng các phương pháp toán học và thống kê. Trong kỹ thuật này, sử dụng phương pháp xử lý dữ liệu đa biến như sự phân tích thành phần cơ bản (Principal component analysis-PCA), phương pháp hồi quy bình phương tối thiểu từng phần (Partial least square regression-PLS), phương pháp hồi quy đa biến (Multiple linear regression-MLR) để xác định mối liên quan giữa phổ hồng ngoại và các đặc tính của thực phẩm.

2.2.3. Phương pháp tiền xử lý phổ cận hồng ngoại

Một số phương pháp tiền xử lý phổ hồng ngoại:

Trung bình (Mean centering – MC): phương pháp này tiến hành tính toán bằng cách trung bình hóa các phổ của bộ dữ liệu, sau đó tiến hành loại trừ giá trị trung bình này cho mỗi phổ đặc thù [7].

Biến đổi chuẩn hóa (Standard Normal Variate-SNV): sử dụng để loại bỏ sự hấp thụ hóa học của ánh sáng tán xạ. SNV thay đổi đường nền phổ và khoảng cách ánh sáng hồng ngoại đi qua một mẫu trong quá trình phân tích. Phép biến đổi SNV đã được chứng minh là một phương pháp hiệu quả, có thể loại bỏ các phổ và độ dốc gây ra bởi sự tán xạ ánh sáng bên trong đối với các mẫu rắn [8].

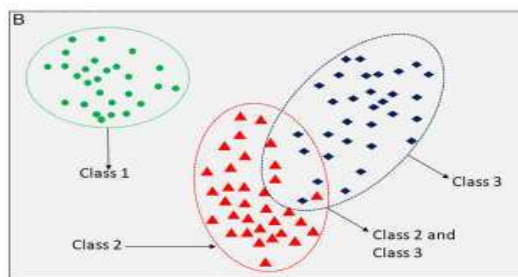
Đạo hàm Savitzky Golay: được áp dụng nhằm mục đích làm mịn dữ liệu, nghĩa là để tăng độ chính xác của dữ liệu mà không làm sai lệch xu hướng tín hiệu. Điều này đạt được trong một quá trình được gọi

là tích chập, bằng cách lắp các tập con liên tiếp của các điểm dữ liệu liên kế với một đa thức bậc thấp bằng phương pháp bình phương tuyến tính nhỏ nhất. Khi các điểm dữ liệu cách đều nhau, có thể tìm thấy một giải pháp phân tích cho các phương trình bình phương nhỏ nhất, dưới dạng một tập hợp "hệ số tích chập" duy nhất có thể được áp dụng cho tất cả các tập con dữ liệu, để đưa ra các ước lượng của tín hiệu tại điểm trung tâm của mỗi tập con. Phương pháp này đã được mở rộng để xử lý dữ liệu 2 và 3 chiều [9].

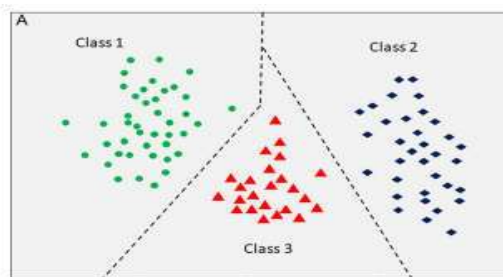
2.2.4. Phương pháp xử lý số liệu

- Xây dựng mô hình bằng SIMCA:

Đối với phương pháp xây dựng mô hình phân loại SIMCA, dựa vào tìm các đặc tính giống nhau của các thành phần trong bộ mẫu, điều này có nghĩa là



mô hình sẽ tập hợp những mẫu có cùng đặc điểm và gọi đó là một nhóm. Do đó, công thức để phân vùng cho mỗi nhóm sẽ được xây dựng bằng thuật toán đặc biệt và phản hồi liệu mẫu có thuộc nhóm được xem xét hay không. Điều này sẽ dẫn đến việc một số mẫu xuất hiện trong các nhóm khác nhau cùng một lúc (có nhiều điểm chung với nhiều nhóm mẫu) hoặc thậm chí không có nhóm nào cả (không có cùng đặc điểm với bất kỳ nhóm nào được liệt kê, vì mô hình phân loại chỉ tập trung vào việc nhóm các mẫu có cùng đặc điểm; so với phương pháp xây dựng mô hình phân biệt, trong đó các mẫu thuộc một nhóm và không gian mẫu được chia hoàn toàn riêng biệt và không xảy ra sự chồng chéo [10].



Hình 2. Mô hình phân loại SIMCA và mô hình phân biệt PLS-DA

SIMCA (Soft Independent Modelling of Class Analogies) là quá trình xây dựng không gian mô hình bằng cách tính khoảng cách d của mỗi đối tượng đến không gian nhóm, tổng hợp hai yếu tố khác nhau: khoảng cách từ không gian mô hình (khoảng cách trực giao) và khoảng cách vào không gian điểm (khoảng cách điểm). Khoảng cách trực giao và khoảng cách điểm tương ứng với Q và T^2 , được tính bằng tổng bình phương của phần dư mô hình và khoảng cách Mahalanobis từ tâm của không gian điểm [11].

- Xây dựng mô hình bằng PLS-DA:

Đối với phương pháp xây dựng mô hình phân biệt, phương pháp này tập trung vào việc mỗi nhóm được chỉ định khác nhau như thế nào, từ đó thuật toán sẽ được áp dụng để tìm các đường phân biệt (thẳng hoặc cong) để chia không gian mẫu của chúng thành các nhóm. Phản hồi sẽ cho biết liệu mẫu thuộc nhóm nào, ở phía nào của ranh giới phân biệt, có nghĩa là mỗi mẫu chỉ có thể thuộc một nhóm. Ví dụ cho các phương pháp này là phân tích phân biệt tuyến tính của Fishers (1936) [12], hoặc phương pháp phân biệt bình phương cực tiểu từng phần

(PLS-DA), Orthogonal PLS-DA [13], hoặc các phương pháp phi tham số như K-Nearest Neighbor...[14].

Phân tích phân biệt bình phương cực tiểu từng phần (PLS-DA) là một công cụ phân loại tuyến tính ứng dụng của thuật toán hồi quy bình phương cực tiểu (PLS). Kỹ thuật chính đằng sau là tìm số lượng biến tiềm ẩn (LVs) tối ưu, sự kết hợp tuyến tính của các biến quan sát, được gọi là biến biểu hiện (MV), phương sai cực đại với phản hồi Y . Điều này được biểu diễn trên một biểu đồ trực quan hóa và hiểu mô hình dữ liệu theo điểm T tiềm ẩn và P tải trọng. Các P tải trọng đại diện cho các hệ số của kết hợp tuyến tính xác định LVs và điểm T đại diện cho tọa độ của các mẫu trong không gian chiếu LV [15].

2.2.5. Đánh giá mô hình

Trong việc xây dựng mô hình, cần phải xác thực mô hình hiệu chuẩn để đo lường khả năng dự đoán của mô hình. Tập dữ liệu được chia thành hai tập: tập huấn luyện để xây dựng mô hình hiệu chuẩn và tập thử nghiệm để kiểm tra khả năng dự đoán của mô hình đã xây dựng. Tùy thuộc vào việc các biến được kiểm tra có khác với các biến trong mô hình hiệu

chuẩn hay không, có 2 loại đánh giá: đánh giá nội bộ bằng mẫu huấn luyện và đánh giá kiểm chứng bằng mẫu thử nghiệm (hoặc đánh giá chéo) [10].

2.3. Trình tự thí nghiệm

Các mẫu cà phê từ các tỉnh: Đắk Lắk, Đắk Nông, Lâm Đồng và Gia Lai được lấy từ thùng bảo quản. Mẫu (50 g) được đựng trong vật chứa bằng thủy tinh (đĩa Petri) và quét 6 lần. Phổ của từng mẫu cà phê được thu thập bằng cách sử dụng máy quang phổ cầm tay (SCiO™). Toàn bộ quá trình được thực hiện ở môi trường nhiệt độ phòng. Sau đó dữ liệu phổ sẽ được trích xuất từ hệ thống dữ liệu đám mây của SciO và tiến hành các quá trình tiền xử lý theo phương pháp Chemometrics với phần mềm SIMCA 15 của Công ty Sartorius (Đức).

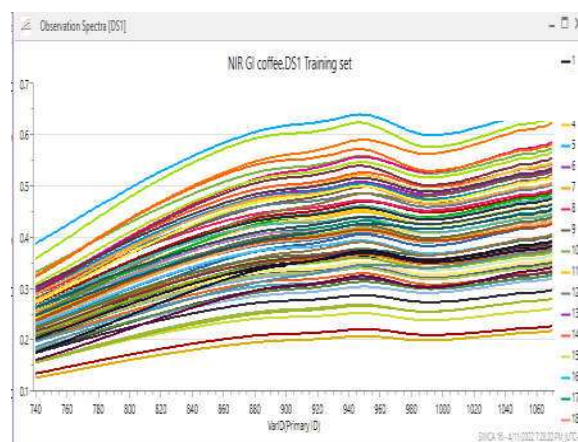
Phổ xử lý là phổ trung bình của 6 lần lặp lại. Các phổ thu được sẽ được tiến hành tiền xử lý theo các phương pháp tiền xử lý khác nhau của phần mềm SIMCA. Sử dụng PLS-DA và SIMCA trong phần mềm SIMCA để xây dựng mô hình phân biệt nguồn gốc và tách loại các loại cà phê bằng dữ liệu phổ cận hồng ngoại thu được từ máy đo quang phổ cận hồng ngoại Scio; so sánh các phương pháp tiền xử lý phổ cận hồng ngoại là phương pháp trung bình trung tâm, biến đổi chuẩn hóa và đạo hàm theo Savitzky-Golay kết hợp.

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Tiền xử lý phổ cận hồng ngoại

Khi sử dụng máy quang phổ cận hồng ngoại quét lên hạt cà phê nhân tạo ra các dải quang phổ và các đỉnh phổ, các dải này tạo thành các âm bội và kết hợp các rung động cơ bản tương ứng với tính chất hóa học, từ đó cung cấp đặc tính riêng biệt hoặc “finger-print” cho các mẫu cà phê khác nhau. Hình 3 cho thấy, nhóm phổ thu được chia sẻ như biểu đồ hấp phụ trong bước sóng từ 740 - 1070 nm. Việc phân biệt các biểu đồ hấp phụ này là không dễ dàng nếu sử dụng dữ liệu thô. Để ứng dụng phân tích dữ liệu

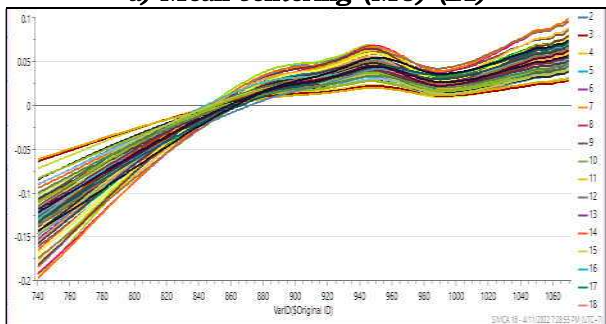
đa biến đối với các dữ liệu thô này, đầu tiên cần tiền xử lý trước bằng phương pháp phù hợp. Các phổ thu được sau khi quét mẫu bằng thiết bị cận hồng ngoại cho ra rất nhiều dải bằng tần (Hình 3). Các dải bằng này được tạo thành từ sự kết hợp của các sóng bậc cao, sự rung động tương ứng với các thành phần hóa học của thực phẩm như caffein, axit chlorogenic, nước. Do đó, nghiên cứu tiến hành đánh giá thử nghiệm sơ bộ với nhiều kỹ thuật xử lý phổ khác nhau và cuối cùng chọn được 3 kỹ thuật tiền xử lý là: Standard Normal Variant (SNV), đạo hàm theo Savitzky Golay, Mean Centering (MC) sử dụng kết hợp các phương pháp, sau đó tiến hành phân tích SIMCA và PLS-DA. Dù có sử dụng bất kỳ phương pháp tiền xử lý nào, tất cả dữ liệu đều được tính trung bình (Mean centering) trước khi đưa vào xử lý số liệu đa biến.



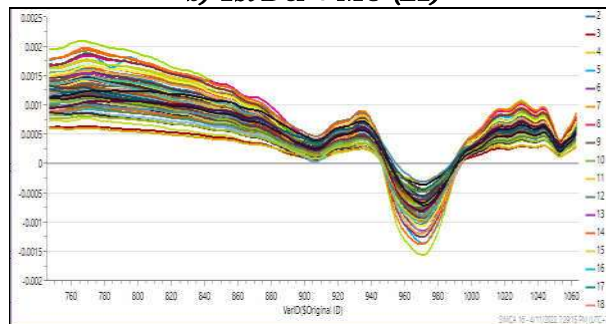
Hình 3. Phổ NIR thu được từ máy Scio cầm tay

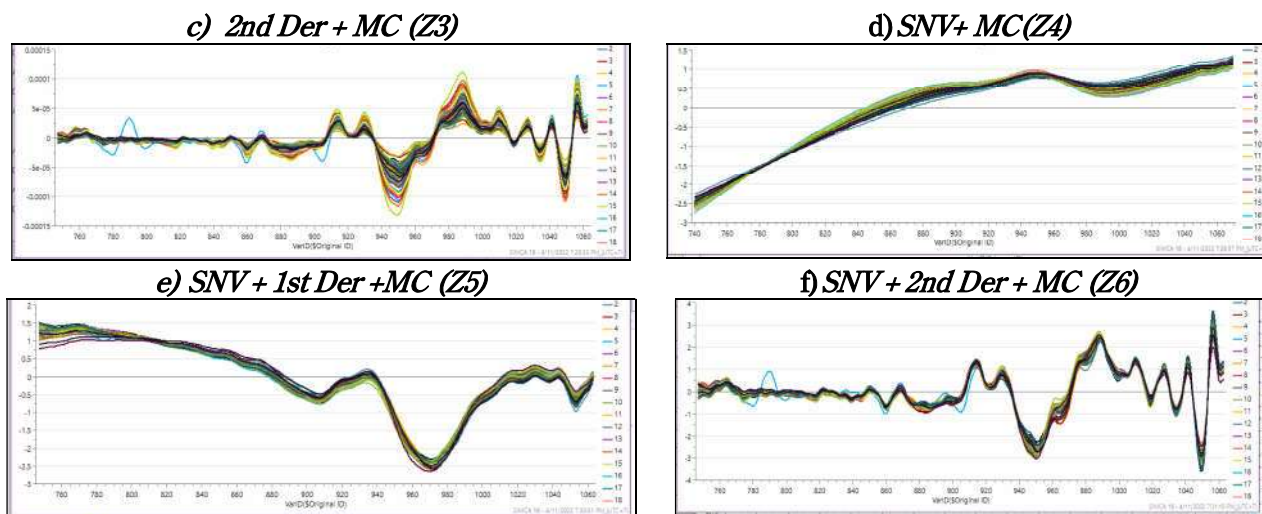
Đặc điểm của phổ NIR là các ảnh hưởng nhiễu và có thể tạo ra sự gia tăng đáng kể của các bước hấp thụ của từng dãy sóng đơn, điều này có thể giấu đi các thông tin quan trọng từ dữ liệu thô. Do đó, các phương pháp tiền xử lý được dùng để giảm đi các sóng nhiễu và cho thấy hiệu quả giảm thiểu đáng kể so với dữ liệu thô (Hình 4).

a) Mean centering (MC) (Z1)



b) 1st Der + MC (Z2)





Hình 4. Phổ thu được sau khi tiền xử lý a) MC, b) 1st Der +MC, c) 2nd Der +MC, d)SNV + MC, e) SNV+ 1st Der + MC, f) SNV + 2nd Der+MC

Đối với phổ được tiền xử lý SNV và MC, phổ được loại bỏ sự hấp thụ hóa học của ánh sáng tán xạ, thay đổi đường nền phổ, loại bỏ các yếu tố nhiễu, đưa các dữ liệu về cùng 1 tỉ lệ.

Đối với đạo hàm bậc 1 và bậc 2 dữ liệu phổ được loại bỏ các đỉnh bị trùng, tăng độ cao của các đỉnh hấp thụ và loại bỏ độ dốc của đường cơ sở.

Nghiên cứu tiến hành kết hợp các phương pháp tiền xử lý để loại bỏ nhiễu và đánh giá cục bộ để chọn ra mô hình có phương pháp tiền xử lý hiệu quả nhất.

3.2. Xây dựng mô hình phân biệt và phê đạt chuẩn chất lượng GI và non- GI bằng SIMCA

3.2.1. Mô hình phân biệt và phê đạt chuẩn chất lượng GI và non- GI

Mô hình xây dựng bằng SIMCA được đánh giá chéo nội bộ bằng 7 bộ mẫu còn lại, tương ứng với 7 địa điểm gửi mẫu đến. Bảng 2 cho thấy, so sánh kết quả phân biệt và phê GI và non - GI trên 8 mẫu hỗn hợp, tất cả các tiền xử lý được thử nghiệm trên các mẫu chuẩn, kết quả được thể hiện trên số lượng biến tiềm ẩn (LVs), thông tin biến tiềm ẩn thể hiện cho tập dữ liệu (Explained Var) và tỉ lệ % phân loại lỗi (Error rate CV) của mô hình.

Kết quả cho thấy, khi sử dụng SIMCA kết hợp với phương pháp tiền xử lý MC, đạo hàm bậc 2 (Z3) và MC, đạo hàm bậc 2, SNV (Z6) cho tỉ lệ lỗi thấp nhất tương ứng với 14,54% và 10,34%. Mô hình Z6 có phương pháp tiền xử lý tốt nhất với tỉ lệ lỗi thấp nhất, chỉ 10,34%, thể hiện bởi 7 biến tiềm ẩn giải thích được 91,47% thông tin của biến đầu vào. Bên cạnh đó,

mô hình Z3 cũng cho tỉ lệ lỗi chỉ 14,54%, thể hiện bởi 21 biến tiềm ẩn (LVs), giải thích được 98,92% thông tin biến đầu vào. Tuy nhiên, để tối ưu, sẽ chọn mô hình có tỉ lệ lỗi thấp nhất

Bảng 2. Đánh giá nội bộ mô hình phân loại các loại và phê bằng SIMCA

	Tiền xử lý	Số biến	Tỉ lệ lỗi (%)	Giải thích tập dữ liệu (%)
Z1	MC	8	47,23	99,76
Z2	1st Der + MC	6	49,12	99,43
Z3	2nd Der +MC	21	14,54	98,92
Z4	SNV + MC	8	41,32	99,64
Z5	SNV+ 1st Der + MC	9	37,98	99,22
Z6	SNV+ 2nd Der + MC	7	10,34	91,47

3.2.2. Kiểm định mô hình phân biệt và phê đạt chuẩn chất lượng GI và non-GI bằng mẫu kiểm chứng bằng SIMCA

Để kiểm chứng mô hình xây dựng bằng SIMCA kết hợp với các phương pháp tiền xử lý MC, đạo hàm bậc 2 (mô hình Z3) và phương pháp tiền xử lý MC, đạo hàm bậc 2, SNV (mô hình Z6), sử dụng dữ liệu của mẫu kiểm chứng để đánh giá mô hình, kết quả được thể hiện qua các thông số về tỉ lệ lỗi (Error rate), độ nhạy (Sensitivity- thể hiện khả năng của mô hình nhận đúng các mẫu thuộc vào nhóm) và độ đặc hiệu (Specificity-mô tả khả năng của mô hình loại bỏ đúng các mẫu không thuộc vào nhóm) như bảng 3. Độ nhạy được tính dựa trên tỉ lệ dương tính thực sự trong mẫu và phê GI. Độ đặc hiệu của mô hình thể

hiện bởi khả năng xác định các mẫu non-GI một cách chính xác. Độ chính xác được tính toán dựa trên tỉ lệ giữa dương tính thực sự và âm tính thực sự trong tất cả các mẫu được đánh giá.

Bảng 3. Đánh giá mô hình phân loại các loại cà phê bằng SIMCA sử dụng mẫu kiểm chứng

	Tỉ lệ lỗi (%)	Độ nhạy (%)	Độ đặc hiệu (%)
Z3	23,22	78,01	82,23
Z6	16,78	82,56	89,08

Kết quả mô hình Z6 cho độ chính xác 85,76%, với độ nhạy là 82,56% và độ đặc hiệu là 89,08%. Trong khi đó, mô hình Z3 cho độ chính xác thấp hơn: 80,09%, với độ nhạy của mô hình là 78,01% và độ đặc hiệu là 82,23%. Với độ nhạy của phép thử là khả năng xác định chính xác mẫu cà phê GI.

Đối với 37 mẫu kiểm chứng bao gồm 12 mẫu GI và 27 mẫu không phải GI, mô hình Z6 nhận đúng 10 mẫu/12 mẫu GI và 24 mẫu/27 mẫu, tỉ lệ lỗi của mô hình khi nhận mẫu GI là 16,78% và khi trên mẫu không GI là 11,11%; có thể thấy mô hình phân nhóm hiệu quả các mẫu có đặc trưng hóa học giống nhau.

Kết quả của mô hình phân loại được thể hiện bằng biểu đồ của Cooman cho thấy khoảng cách trực giao bình phương từ các điểm dữ liệu đến mô hình SIMCA (Z6) đã chọn, cũng như các giới hạn khoảng cách tới hạn cho các điểm thuộc mô hình. Khoảng cách tới hạn được xây dựng bằng cách sử dụng phân phối chi bình phương của các giá trị khoảng cách của các điểm trong cùng nhóm, với mô hình Z6 khoảng cách tới hạn của các điểm thuộc nhóm cà phê GI là 1,276, nhóm cà phê non- GI là 1,21. Trong đó, các dữ liệu cùng một nhóm, tập trung ở cùng 1 khu vực và được thể hiện cùng màu với nhau.



Hình 5. Biểu đồ Cooman's thể hiện các nhóm phân biệt bằng mô hình

Kết quả thu được của nghiên cứu cũng tương thích với nghiên cứu phân loại cà phê nhân chất lượng cao (Gourmet) so với các loại cà phê truyền thống (Traditional) của Taynna Kevla và cs (2021) [16], với mục tiêu phân biệt, đảm bảo giá trị và chất lượng của các loại cà phê cao cấp nhất Brazil không bị giả mạo, bảo vệ thương hiệu cà phê ở đất nước có tỉ trọng xuất khẩu cà phê cao nhất thế giới. Với số mẫu cà phê chất lượng cao (Gourmet) 44, các loại cà phê còn lại (Traditional) 46 mẫu, tiến hành thu thập phổ cận hồng ngoại không phá hủy mẫu. Nghiên cứu đã thành công sử dụng SIMCA kết hợp với các phương pháp tiền xử lý số liệu và cho ra được mô hình SIMCA kết hợp với đạo hàm và SNV với độ nhạy 100% chỉ 7 biến đại diện.

3.3. Xây dựng mô hình phân biệt cà phê đạt chuẩn chất lượng GI và non- GI bằng PLS-DA

3.3.1. Mô hình phân biệt cà phê đạt chuẩn chất lượng GI và non- GI

Khi sử dụng phương pháp xử lý bằng PLS-DA, kết hợp với các phương pháp tiền xử lý, mô hình được đánh giá nội bộ chéo (LOO-CV) bởi 7 nhóm còn lại, thì tỉ lệ lỗi của mô hình Z4 là thấp nhất (11,44%) và thể hiện được nhiều thông tin của dữ liệu tốt nhất (94,92%).

Bảng 4. Đánh giá nội bộ mô hình phân loại các loại cà phê bằng PLS-DA

	Tiền xử lý	Số biến	Tỉ lệ lỗi (%)	Giải thích tập dữ liệu (%)
Z1	MC	8	29,80	83,20
Z2	1st Der + MC	8	44,83	69,34
Z3	2nd Der +MC	8	25,89	60,88
Z4	SNV + MC	8	11,44	94,92
Z5	SNV+ 1 st Der + MC	3	26,78	74,98
Z6	SNV+ 2 nd Der + MC	6	47,61	51,37

3.3.2. Kiểm định mô hình phân biệt và phê đạt chuẩn chất lượng GI và non-GI bằng mẫu kiểm chứng

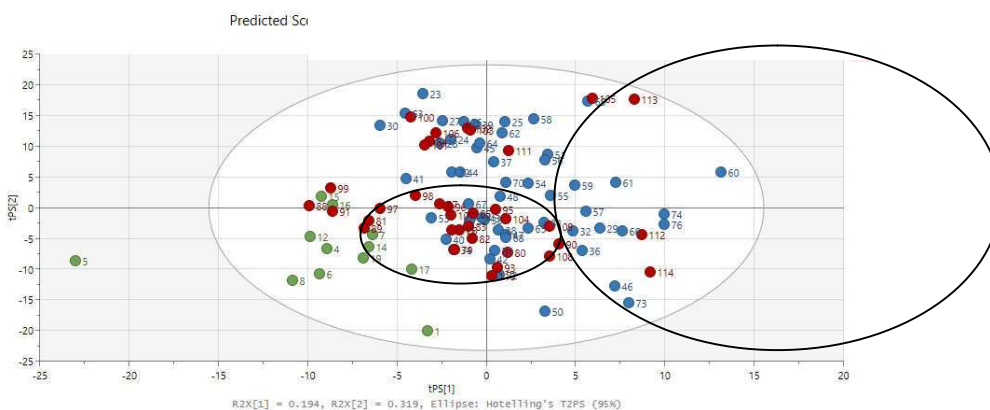
Với mô hình Z4 được xây dựng bằng PLS-DA kết hợp với tiền xử lý SNV, mô hình được đánh giá trên mẫu kiểm chứng: với 14 mẫu GI, mô hình nhận đúng 13 mẫu, không nhận ra được 1 mẫu; với 23 mẫu non-GI, mô hình nhận đúng 20 mẫu, không nhận được 2 mẫu và sai 1 mẫu. Mô hình phân biệt hiệu quả hơn đối với các mẫu cả phê đạt GI với tỉ lệ nhận lỗi (7,14%) thấp hơn so với các mẫu cả phê không đạt GI (13,04%). Điều này chứng tỏ, các mẫu cả phê GI có

tính đặc trưng tương tác khá chặt chẽ hơn các mẫu cả phê không đạt GI.

Bảng 5. Kết quả phân loại các loại cả phê bằng PLS-DAs

Kiểm chứng	GI	non-GI	Không nhận được
GI	13	0	1
Non-GI	1	20	2

Kết quả mô hình được chiếu lên mặt phẳng như hình 6, các mẫu kiểm chứng được phân bố thành 2 nhóm thuộc về khu vực nhóm GI và non-GI như đã khoanh tròn.



Hình 6. Kết quả phân loại các loại cả phê bằng PLS-DA sử dụng mẫu kiểm chứng

Mô hình Z4 cho kết quả độ nhạy 81,25%, độ đặc hiệu 95,23%, tỉ lệ lỗi 12,03% và độ chính xác 87,96%.

Bảng 6. Đánh giá mô hình phân loại các loại cả phê bằng PLS-DA sử dụng mẫu kiểm chứng

	Tỉ lệ lỗi (%)	Độ nhạy (%)	Độ đặc hiệu (%)
Z4	12,03	81,25	95,23

Kết quả này phù hợp với nghiên cứu của A. Giraudo và cs (2018) [17] về xác định nguồn gốc địa lý của cả phê nhân sử dụng dữ liệu phổ cận hồng ngoại kết hợp với các phương pháp xử lý số liệu đa biến. Với số mẫu tổng cộng 191 mẫu từ 2 châu lục và 9 quốc gia, nghiên cứu đã sử dụng PLS-DA kết hợp với MC và SNV cho ra mô hình với độ nhạy lên đến 98% khi phân biệt theo châu lục và 100% khi phân biệt theo quốc gia.

Giữa các châu lục cách xa nhau về địa lý, nên mô hình phân biệt dữ liệu đầu vào có thể cho kết quả cao khi phân biệt các mẫu. Trong khi đó, các mẫu cùng tập trung ở khu vực Tây Nguyên, điều kiện khí hậu khác biệt không quá lớn, vì vậy mô hình có khả năng phân biệt với độ nhạy 87,96%, là công cụ để giúp phân loại nhanh các loại cả phê theo chất lượng.

3.4. So sánh mô hình phân biệt cả phê đạt chuẩn chất lượng GI và non-GI bằng SIMCA và PLS-DA

Đầu tiên, so sánh về các phương pháp tiền xử lý, đối với phương pháp sử dụng SIMCA, phương pháp tiền xử lý tốt nhất là kết hợp đạo hàm bậc 2, SNV và MC, sẽ trích xuất nhiều thông tin quan trọng và giảm nhiễu dữ liệu thô hơn so với các phương pháp tiền xử lý khác. Trong khi đó, phương pháp tiền xử lý SNV kết hợp trung bình hiệu chuẩn (MC) kết hợp với PLS-DA tỉ lệ lỗi chỉ 11,44%. Có thể giải thích bởi bản chất của xử lý SNV trên dữ liệu phổ là sự thay đổi đường cơ sở chính xác theo mức độ đặc hiệu của từng phổ hấp thụ, đặc biệt là đối với mô hình PLS; trong khi đối với mô hình SIMCA, SNV loại bỏ các yếu tố là đặc điểm chung của các mẫu để căn cứ phân loại nhóm. Đối với đạo hàm bậc nhất, phương pháp loại bỏ các xu hướng tuyến tính, giảm tính đặc trưng và giảm tỉ lệ lỗi từ 10-20% phụ thuộc vào mô hình. Đối với đạo hàm bậc 2, phương pháp nâng cao hiệu suất của mô hình do giảm nhiễu tín hiệu, do đó giá trị của tỉ lệ lỗi cũng thay đổi nhiều hơn. Do đó, SNV kết hợp MC cho kết quả tốt ở mô hình PLS-DA

và cần kết hợp với đạo hàm bậc 2 để tăng đỉnh hấp thụ của phổ để hiệu quả hơn ở SIMCA.

Để so sánh hiệu quả của việc xây dựng mô hình bằng 2 phương pháp trong chemometrics: SIMCA và PLS-DA, nghiên cứu thực hiện so sánh các thông số tỉ lệ lỗi, độ nhạy, độ chính xác, độ đặc hiệu của 2 mô hình sau khi đánh giá bằng mẫu kiểm chứng. Kết quả cho thấy, mô hình xây dựng bằng PLS-DA kết hợp với các phương pháp tiền xử lý cho độ nhạy, độ chính xác, độ đặc hiệu tốt hơn, đồng thời tỉ lệ lỗi thấp so với SIMCA khi dùng để phân loại cà phê GI và non-GI.

Bảng 7. So sánh hiệu quả phân loại của 2 mô hình xây dựng bằng SIMCA và PLS-DA

	SIMCA	PLS-DA
Tỉ lệ lỗi (EV)	16,78	12,03
Độ chính xác (%)	85,76	87,96
Độ nhạy (%)	82,56	81,25
Độ đặc hiệu (%)	89,08	95,23

Kết quả trên phù hợp với nghiên cứu của O.Galtier và cs (2010)[18] về so sánh các phương pháp chemometric: SIMCA và PLS-DA trên dữ liệu phổ cận hồng ngoại để phân loại các loại dầu olive nguyên chất bằng NIR. Với tổng cộng 317 mẫu dầu olive thu thập tại 5 tỉnh ở Pháp, nhằm phân biệt chất lượng dầu olive có ít axit béo nhất và hàm lượng tryglycerol nhất đến từ tỉnh Haute. Nghiên cứu đã thành công xây dựng mô hình phân biệt 5 loại dầu olive nguyên chất khác nhau bằng PLS-DA với độ nhạy đạt 100%. Nghiên cứu cũng đã so sánh giữa SIMCA và PLS-DA, kết quả cho thấy PLS-DA cho ra mô hình có độ nhạy tốt hơn, trong khi mô hình xây dựng bằng SIMCA có độ nhạy chỉ lên đến 91%.

4. KẾT LUẬN

Mô hình xây dựng bằng phương pháp xử lý số liệu PLS-DA kết hợp phương pháp tiền xử lý MC và SNV cho độ nhạy đến 87,96%. Đây là một kết quả rất triển vọng cho việc phát triển phương pháp xác thực nhanh bằng thiết bị quang phổ cầm tay để hỗ trợ cho việc chống hàng giả mạo, đảm bảo tính xác thực về chất lượng của cà phê đã được đăng kí hồ sơ chỉ dẫn địa lý và góp phần hỗ trợ đánh giá nhanh nguồn nguyên liệu đầu vào trong quá trình đánh giá hồ sơ đăng kí chỉ dẫn địa lý. Bên cạnh đó, ngoài phục vụ cho quy trình và sản phẩm cà phê, nghiên cứu có thể ứng dụng vào các loại sản phẩm khác như yến sào, tiêu đen.

TÀI LIỆU THAM KHẢO

1. Nguyen Thi My Hang, Nguyen Thi Minh Thuy (2020). The status quo and solutions for developing Vietnam's coffee industry. Industry and trade magazine: 10-20.
2. Asean Intellectual Property Rights (2020). A booklet on Asean geographical indications procedure and products. Chapter 1.
3. UBND tỉnh Đắk Lắk (2010). Quyết định ban hành "Quy chế quản lý, sử dụng chỉ dẫn địa lý cà phê Buôn Ma Thuột đối với sản phẩm cà phê nhân Robusta".
4. Mezza-Marquez, O. G., Gallardo, T. V., Osorio-Revilla, G. (2010). Application of mid-infrared spectroscopy with multivariate analysis soft independent modelling of class analogies (SIMCA) for the detection of adulterants in minced beef. *Meat Sciences* 86: 511-519.
5. Li, M., Qian, Z., Shi, B., Medlicott, J., East, A. (2018). Evaluating the performance of a consumer scale SCiO™ molecular sensor to predict quality of horticultural products. *Postharvest Biology and Technology*, 145, 183-192.
6. Cozzolino, D. (2014). An overview of the use of infrared spectroscopy and chemometrics in authenticity and traceability of cereals. *Food Research International*, 60: 262-265.
7. Rinnan, A., Van den Berg, F., Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* 28 (10): 1201-1222.
8. Dhanoa, M., Lister, S., Sanderson, R., Barnes, R. (1994). The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformation of NIR spectra. *Journal of Near Infrared Spectroscopy* 2 (1): 43 – 47.
9. Savitzky, A., Golay, M.J.E (1964). *Analytical Chemistry journal*. Chapter 1.
10. Banicollio, A., Ma°ge, I., Ns, T (2015) Combining SO-PLS and linear discriminant analysis for multi-block classification, *Chemom. Intell. Lab. Syst.* 141, 58-67.
11. Wold, S., Sjostrom, M. (1977). SIMCA: a method for analysing chemical data in terms of similarity and analogy. In: Kowalski, B.R. (Ed.), *Chemometrics, Theory and Application*, American Chemical Society Symposium Series No. 52. *American Chemical Society*, Washington, DC, pp. 243-282.

12. Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugen.* 7, 179-188.
13. Bylesjo, M., Cloarec, O., Rantalainen, M. (2009) *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Oxford, Elsevier, Vol. 2, pp. 109-125.
14. Jiang, L, *et al.* (2007). Survey of improving k-nearest-neighbor for classification. in *Fuzzy Systems and Knowledge Discovery*, 2007. FSKD 2007. Fourth International Conference on. 2007. IEEE.
15. Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. *PLS-DA. Analytical Methods*, 5(16), 3790.
16. De Araujo, T.K. L, Nobrega, O., De Souza Fernandes, D.D., De Araujo, M.C.U., Dias Diniz, P.H.G., Da Silva, E.C. (2021). Non- destructive authentication of Gourmet ground roasted coffees using NIR spectroscopy and digital images. *Food chemistry* 135(2012) 1828-1835.
17. Giraudo, A., Grassi, S., Savorani, F., Gavoci, G., Casiraghi, E., Geobaldo, F. (2019). Determination of the geographical origin of green coffee beans using NIR spectroscopy and multivariate data analysis. *Food control* 78-80 (2019).
18. Galtier, O. Abbas, O., Y. Le Dreau, Y., Rebufa, C., Kister, J., Artaud, J., N. Dupuy, N. (2010). Comparison of PLS-DA and SIMCA for classification by origin of crude petroleum oils by MIR and virgin olive oils by NIR for different spectral regions. *Vibrational Spectroscopy* 55(2011) 132-140.

STUDY ON THE USING NEAR-INFRARED (NIR) SPECTROSCOPY COMBINE WITH CHEMOMETRICS TO AUTHENTICATE THE GEOGRAPHY INDICATOR COFFEE IN BUON MA THUOT.

To Phan Chieu Dan^{1,2}, Nguyen Quoc Cuong^{1,2}, Le Nguyen Doan Duy^{3*}

¹*Department of Food Technology, Ho Chi Minh city University of Technology (HCMUT)*

²*Vietnam National University Ho Chi Minh city*

³*Faculty of Food Science and Technology, Ho Chi Minh city University of Food Industry*

**Email: duyind@hufi.edu.vn*

Summary

Coffee is an important commodity, accounting for 3% of Vietnam' GDP, and the Vietnam' export turnover of coffee has reached over US \$3 billion for many years. In recent years, although the world economy has experienced many difficult times leading to a decrease in purchasing power, Vietnam's coffee exports have maintained a very encouraging growth rate, there is an unmet need to protect coffee brand and quality to prevent the frauds in coffee supply chain. Thus, this work proposed the suitability of NIR spectroscopy coupled with chemometrics methods for the non- destructive authentication of Geography indicator coffees. For this, GI coffees beans (n=49) and non-GI coffees beans (n=103) by directly analyzing without any sample preparation. Then, PLS-DA and SIMCA were used to construct the models. PLS-DA combine with MC, SNV pre-processing constructing model achieved the best result, model accuracy 87.96%, sensitivity 81.25% and specificity 95.23%. By the way, SIMCA combine with MC, 2nd Der, SNV pre-processing constructing model achieved the model accuracy just 85.76%. Therefore, the proposed methodologies can be useful for both the consumer and regulatory in registration GI for products because it confirms the elevated standard of Dak Lak specialty coffees (GI), preventing fraudulent labeling, friendly with environmental and quick technique, save time and save money.

Keywords: *Coffees, chemometrics, geography indicator, NIR spectroscopy.*

Người phản biện: PGS.TS. Trần Như Khuyến

Ngày nhận bài: 28/4/2022

Ngày thông qua phản biện: 27/5/2022

Ngày duyệt đăng: 20/6/2022