

## ỨNG DỤNG CÁC PHƯƠNG PHÁP PHÂN TÍCH THỐNG KÊ ĐA BIẾN ĐỂ PHÂN LOẠI NGUỒN GỐC MẬT ONG Ở VIỆT NAM DỰA TRÊN DỮ LIỆU PHỔ $^1\text{H-NMR}$

Đến tòa soạn 10-03-2021

**Nguyễn Thế Anh, Phạm Quang Trung, Tạ Thị Thảo**

Khoa Hóa học, Trường Đại học Khoa học Tự nhiên- Đại học Quốc Gia Hà Nội

**Nguyễn Thị Hà Ly**

Khoa phân tích tiêu chuẩn – Viện Dược liệu

### SUMMARY

*Bee honey is the most well-known and economically important. Honey has long been used as a home remedy to treat coughs and colds for children. Honey diversity is due to many diverse origins. Properties and compositions of bee honey depend strongly on the type of flowers the bees visited, season, on the climatic conditions in which the plants grow and treatment of beekeepers. The aim of this study was to classify honey samples in Vietnam using  $^1\text{H-NMR}$  data combined multivariate statistics (PCA, SIMCA, kNN, LDA and PLS-DA) methods. The similar results of origin classification of un-known eight honey samples by four multivariate statistics methods were obtained and proved the accurate applied methods. The results can help market managers and companies control the origin of honey samples in Vietnam.*

**Key words:** Mật ong, phân loại nguồn gốc, thống kê đa biến, phổ  $^1\text{H-NMR}$ .

### 1. ĐẶT VẤN ĐỀ

Mật ong là một vị thuốc quý được sử dụng từ lâu trong y học cổ truyền (YHCT) và được qui định trong Dược điển Việt Nam [1] với các công dụng chính như chữa ho, cảm lạnh, bổ trung, nhuận táo, giải độc. Y học hiện đại đã chứng minh mật ong có tác dụng tăng sức đề kháng, phòng chống nhiễm trùng, bồi bổ,... Về thành phần hoá học, mật ong là hỗn hợp của các loại đường và một số thành phần khác. Cacbohydrat trong mật ong chủ yếu là fructose (khoảng 38,5%) và glucose (khoảng 31,0%), maltose, sucrose và cacbohydrat hỗn hợp [4]. Mật ong cũng chứa một lượng nhỏ các vitamin, khoáng chất và các hợp chất chức năng như chất chống oxy hóa bao gồm chrysin, pinobanksin, vitamin C, catalase và pinocembrin. Tuy nhiên, thành phần cụ thể của mật ong phụ thuộc vào hoa mà ong hút mật [5]. Theo số liệu năm 2013, sản lượng mật ong trung bình là 42.000 tấn/năm, xuất khẩu 38.000

tấn, tiêu thụ nội địa khoảng 4.000 tấn [2]. Phần lớn mật ong của Việt Nam là mật ong đa hoa, ngoài ra, còn có nhiều loại mật ong đơn hoa khác như mật ong cà phê, nhãn, táo, vải, bạc hà, cao su, hạt điều... Theo Tiêu chuẩn Việt Nam về mật ong tự nhiên, mật ong được phân biệt theo nguồn gốc thực vật thành ba loại là mật ong hoa, mật ong dịch lá và mật ong hỗn hợp. Ngoài ra, nguồn gốc thực vật của mật ong còn chia theo các loại cây như mật ong bạc hà, mật ong nhãn, mật ong cà phê [3]. Theo nguồn gốc địa lý, mật ong được phân biệt từ các vùng, các địa phương khác nhau như mật ong Đắk Lắk, Gia Lai, Bình Phước, Bắc Giang, một số tỉnh Đông Nam Bộ như: Lâm Đồng (cây cà phê), Đồng Nai (cà phê, cao su và chôm chôm), Bình Phước (hạt điều, cao su), Hưng Yên (nhãn, keo)...[2]. Hiện nay trên thị trường sản phẩm mật ong rất đa dạng về xuất xứ, mùi vị. Một số đơn vị sản xuất thường pha trộn nhiều nguồn mật ong khác nhau để có

được màu sắc và khối lượng cần thiết, đồng thời làm giảm giá thành sản phẩm, điều này gây ảnh hưởng đến quyền lợi người tiêu dùng. Vì vậy cần thiết có phương pháp nhằm xác định được chính xác nguồn gốc của mật ong ở Việt Nam.

Phương pháp phân tích thống kê đa biến (*multivariate analysis*) là phương pháp xử lý số liệu nhằm tìm ra mối quan hệ giữa các biến trong tập số liệu. Trên cơ sở bộ số liệu (về hóa học) của mẫu phân tích, các phương pháp thống kê đa biến sẽ tìm ra điểm giống nhau và khác nhau giữa các mẫu, từ đó phân loại chúng vào các nhóm khác nhau. Với một mẫu bất kỳ khi sử dụng thống kê đa biến hoàn toàn có thể nhận dạng xem chúng thuộc nhóm nào trong các nhóm đã được phân loại. Một số phương pháp thống kê đa biến thường được sử dụng trong phân nhóm mẫu gồm: phân nhóm không chỉ dẫn (PCA, CA) và phân nhóm có chỉ dẫn (PLS-DA, LDA, kNN, SIMCA) [10-11]. Đối với các phương pháp phân nhóm không chỉ dẫn, dữ liệu đầu vào không được dán nhãn và không có kết quả đã biết, một mô hình phân cụm sẽ được đề xuất dựa trên thông tin thu được có trong dữ liệu đầu vào. Đối với các phương pháp phân nhóm có chỉ dẫn, dữ liệu đầu vào (gọi là dữ liệu luyện) có nhãn hoặc kết quả đã biết, một mô hình phân nhóm sẽ được đề xuất dựa trên dữ liệu luyện đó, và có thể đưa ra dự đoán về nhóm đối với những dữ liệu cần kiểm tra [6-7]. Từ thực tiễn sự đa dạng về nguồn gốc của mật ong ở Việt Nam, cùng với thực trạng chất lượng mật ong trên thị trường Việt Nam còn nhiều tồn tại, chưa được kiểm soát chặt chẽ. Nghiên cứu này được thực hiện tiến hành áp dụng mô hình thống kê đa biến kết hợp các dữ liệu hoá học để bước đầu phân loại nguồn gốc mật ong ở Việt Nam.

## **2. THỰC NGHIỆM**

### **2.1. Hóa chất, thiết bị**

#### **2.1.1 Hóa chất**

Dung môi đo NMR là hỗn hợp của D<sub>2</sub>O và nước cất 2 lần (tỷ lệ 90:10, v/v); acetonitrin (ACN), magie sunfat (MgSO<sub>4</sub>), natri acetat (CH<sub>3</sub>COONa).

#### **2.1.2 Thiết bị, dụng cụ**

Thiết bị phổ cộng hưởng từ hạt nhân của hãng Bruker, Advance 500 MHz; các dụng cụ gồm có: micropipet 100μl; ống đo NMR khô và sạch; ống fancel dung tích 15 ml.

### **2.2. Đối tượng nghiên cứu**

Các mẫu mật ong được thu thập từ thị trường là các cửa hàng, siêu thị, hoặc trực tiếp lấy trong kho công ty và nhà dân... Các mẫu sau khi thu thập được chuyển về phòng thí nghiệm và bảo quản trong tủ mát (20°C). Quy trình lấy mẫu: 57 mẫu mật ong được thu thập từ các nguồn khác nhau, trong đó có 18 mẫu mật ong nhãn, 10 mẫu mật ong từ hoa khác và 29 mẫu chưa biết rõ nguồn gốc. Sau khi lấy mẫu, tiến hành tách tập mẫu này ra làm 2 phần: *mẫu luyện* gồm 49 mẫu mật ong đã có nguồn gốc, và *mẫu kiểm tra* gồm 8 mẫu chưa rõ nguồn gốc.

### **2.3. Phương pháp nghiên cứu**

#### **2.3.1. Phương pháp phổ cộng hưởng từ hạt nhân (NMR) phân tích mẫu mật ong**

Mẫu đo được chuẩn bị bằng cách hòa tan 0,10 ml mẫu mật ong trong 0,3 ml hỗn hợp dung môi đo NMR. Tiến hành đo phổ <sup>1</sup>H-NMR trên thiết bị Bruker Advance 500 MHz 11,7T; số điểm dữ liệu: 32K; thời gian: 2,04s; độ rộng phổ: 8012,820 Hz. Phép chuyển hóa Fourier các phổ và trừ nền được tiến hành tự động bằng phần mềm Topspin. Để so sánh cường độ pic của các mẫu, ta dịch chuyển pic của α-glucose về vị trí 5,12 ppm (vị trí chuẩn với phổ <sup>1</sup>H-NMR), khi đó các pic còn lại sẽ được dịch chuyển tương đối. Ghi toàn bộ phổ từ 0,2- 5,4 ppm. Từ thiết bị NMR, ma trận số liệu xuất dưới dạng .txt và nhập vào Matlab.

#### **2.3.2. Phân nhóm mẫu dựa trên các phương pháp thống kê**

Trong nghiên cứu này, các phương pháp thống kê (PCA, kNN, SIMCA, LDA) đều được thực hiện trên phần mềm “PCA toolbox for MATLAB”, là sản phẩm miễn phí của Milano Chemometrics and QSAR Research Group [8-9] bao gồm một tập hợp các modul MATLAB để phân tích cấu trúc dữ liệu.

Phân tích thành phần chính (Principal Component Analysis, PCA): PCA là một

phương pháp thống kê thường được sử dụng để phân nhóm không giám sát các mẫu trong một tập mẫu.

Thao tác: Chọn đường dẫn đến thư mục PCA toolbox for MATLAB => Mở giao diện sử dụng của PCA toolbox bằng lệnh: `pca_gui`; Trong thẻ File, load các ma trận dữ liệu gốc ( $X_{train}$ ), ma trận nhóm (Class\_train) và các ma trận nhãn (Labels); Trong thẻ Calculate, ta chọn phương pháp phân tích cần dùng, cụ thể là PCA; Số PCs cũng như cách tiền xử lý số liệu có thể lựa chọn khảo sát.

Để xem các đồ thị kết quả, vào thẻ Results; để xuất một đồ thị, chọn Export. Để kiểm tra một bộ số liệu mới, lưu lại mô hình đã lập được bằng File – Save Model sau đó load data bằng ma trận số liệu mới. Vào thẻ Predict, chọn Predict Sample. Xem kết quả dự đoán với bộ số liệu mới ở Results.

Bộ công cụ ICOSHIFT: Bộ công cụ ICOSHIFT cho Matlab là một chương trình mã nguồn mở được thiết kế đặc biệt để giải quyết các vấn đề liên kết tín hiệu trong phân tích dữ liệu NMR. Trong nghiên cứu này, ICOSHIFT được sử dụng để xác định các pic và các vùng pic trong phổ, từ đó đưa ra các hiệu chỉnh cần thiết cho ma trận tín hiệu đo.

Thao tác:

Bước 1: Nhập bộ công cụ ICOSHIFT và các ma trận phổ NMR.

Bước 2: Nhập câu lệnh của ICOSHIFT

Bước 3: Xuất kết quả

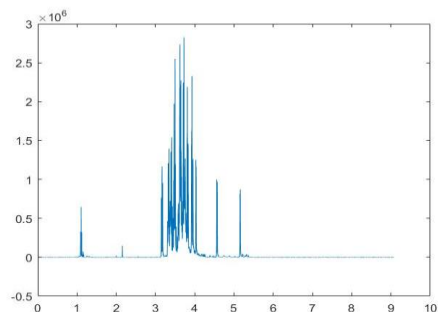
Các phương pháp khác kNN, LDA, SIMCA:

Thao tác: chọn đường dẫn đến thư mục Classification toolbox for MATLAB; mở giao diện sử dụng của Classification toolbox bằng lệnh: `class_gui`; Trong thẻ File, load các ma trận dữ liệu gốc ( $X_{train}$ ), ma trận nhóm (Class\_train) và các ma trận nhãn (Labels). Để thực hiện thuật toán nào, chọn thuật toán đó ở thẻ Calculate. Sau khi khảo sát số PCs cần thiết cũng như các phương pháp tiền xử lý số liệu, chọn Calculate. Xem kết quả ở thẻ Results, kiểm tra với bộ số liệu mới sau khi đã load ở thẻ Predict.

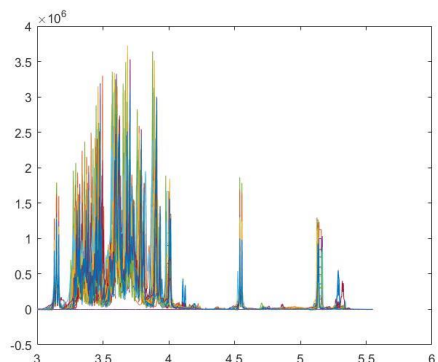
### 3. KẾT QUẢ NGHIÊN CỨU

#### 3.1. Phân tích dữ liệu $^1\text{H-NMR}$ của các mẫu mật ong

Tiến hành đo phổ  $^1\text{H-NMR}$  theo điều kiện trình bày trong mục 2.3, thu được phổ  $^1\text{H-NMR}$  của 1 mẫu mật ong (Hình 1) và hình ảnh chồng phổ  $^1\text{H-NMR}$  của 57 mẫu mật ong thu thập được (Hình 2).

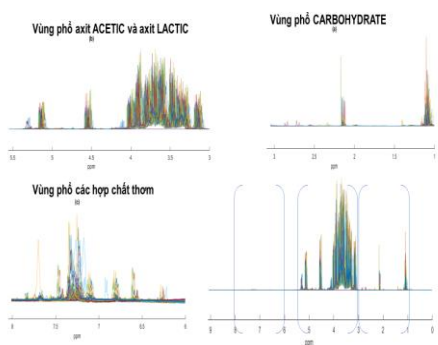


Hình 1. Phổ  $^1\text{H-NMR}$  của một mẫu mật ong



Hình 2. Phổ  $^1\text{H-NMR}$  của 57 mẫu mật ong khi xếp cạnh nhau

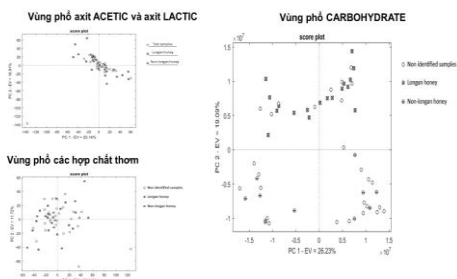
Nhìn vào phổ đồ Hình 2, có thể chia phổ  $^1\text{H-NMR}$  của các mẫu mật ong thành 3 vùng khác nhau theo độ dịch chuyển hoá học (số ppm) như Hình 3.



Hình 3. Ba vùng phổ chính của trong  $^1\text{H-NMR}$  của mẫu mật ong

### 3.2. Phân cụm các mẫu mật ong dựa trên mô hình PCA kết hợp công cụ ICOSHIFT

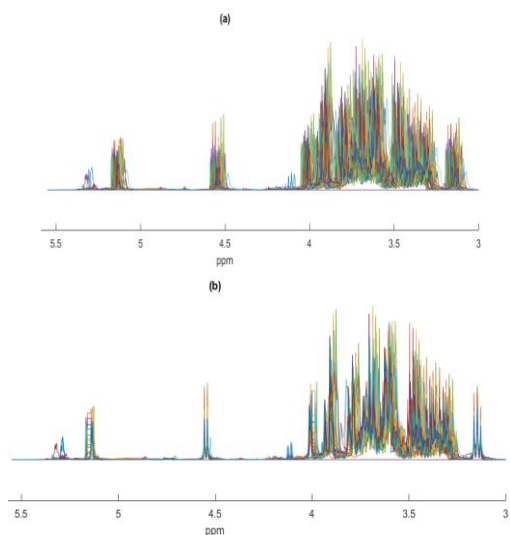
Sau khi thu nhận tín hiệu  $^1\text{H-NMR}$  và thực hiện thuật toán PCA với lần lượt 3 vùng phổ, thu được kết quả như Hình 4.



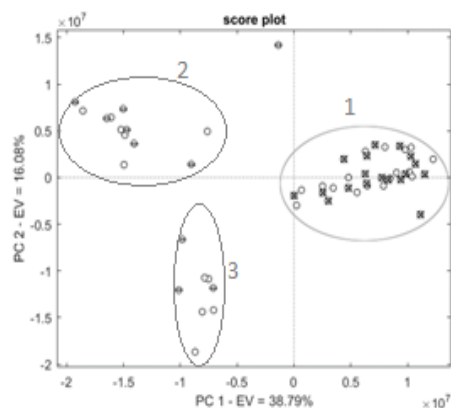
Hình 4. Đồ thị các trọng số (score plot) trong PCA của 3 vùng phổ

Kết quả thu được cho thấy vùng tín hiệu phổ  $^1\text{H-NMR}$  tương ứng với các nhóm chất carbohydrat có ý nghĩa về mặt phân loại nhất. Tuy nhiên, độ chụm của các mẫu vẫn còn chưa cao, vì vậy, cần tối ưu hóa tập số liệu bằng cách sắp xếp lại dữ liệu, sử dụng bộ công cụ ICOSHIFT (Hình 5).

Từ tập số liệu sau khi hiệu chỉnh bằng ICOSHIFT, thực hiện lại thuật toán PCA với 49 mẫu luyện này, thu được đồ thị trọng số như Hình 6.



Hình 5. Phổ  $^1\text{H-NMR}$  sau khi tiền xử lý số liệu bằng ICOSHIFT

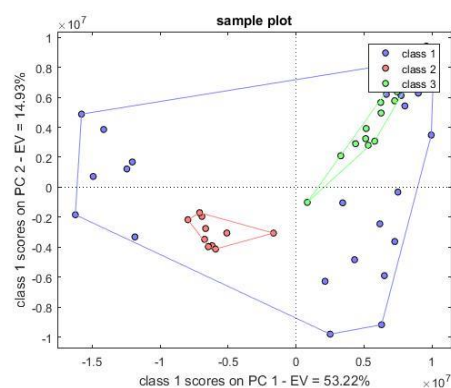


Hình 6. Đồ thị PCA phân nhóm 49 mẫu luyện  
Nhận thấy phần trăm phương sai của 2 PC đầu tiên đã được cải thiện, tổng giá trị lớn hơn 55% nên kết quả phân loại là có ý nghĩa. Đồng thời, từ kết quả thu được ở Hình 6, có thể thấy 49 mẫu của bộ số liệu luyện được chia thành 3 nhóm rõ rệt (ký hiệu lần lượt là các nhóm 1, 2, 3), trong đó 1 có khả năng cao là nhóm mật ong nhãn.

### 3.3. Phân nhóm mẫu dựa trên các thuật toán thống kê đa biến khác

Kết quả phân nhóm bằng SIMCA

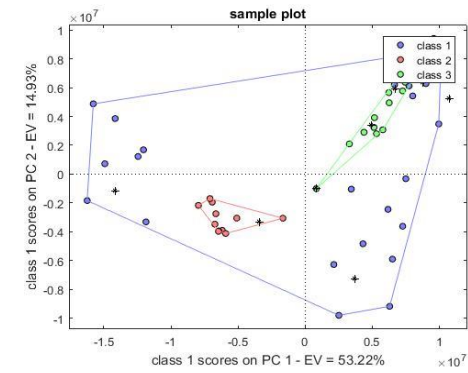
Sau khi khảo sát các điều kiện tiền xử lý số liệu, số PCs cần thiết để chọn phương án thực hiện tối ưu, thu được đồ thị SIMCA-score như Hình 7.



Hình 7. Đồ thị SIMCA-score phân nhóm 49 mẫu luyện

Với tổng phần trăm phương sai lớn hơn 65%, kết quả này là hoàn toàn chấp nhận được. Từ biểu đồ trên, dễ thấy các nhóm có sự phân biệt khá rõ rệt, đặc biệt sự phân nhóm của các mẫu theo SIMCA có sự tương đồng so với sự phân nhóm theo PCA. Áp dụng đồ thị SIMCA để

kiểm tra 8 mẫu chưa biết, thu được kết quả như Hình 8.

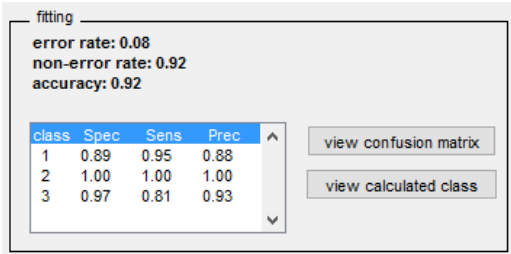


Hình 8. Đồ thị SIMCA kiểm tra mẫu

Kết quả thu được cho thấy trong số 8 mẫu này mật ong chưa có nguồn gốc, có 3 mẫu có nguồn gốc là mật ong nhãn và 5 mẫu không phải mật ong nhãn.

#### Kết quả phân nhóm bằng kNN

Sau khi khảo sát các điều kiện tiền xử lý số liệu, giá trị K cần thiết để chọn phương án thực hiện tối ưu, ta được kết quả phân nhóm kNN như Hình 3.9.

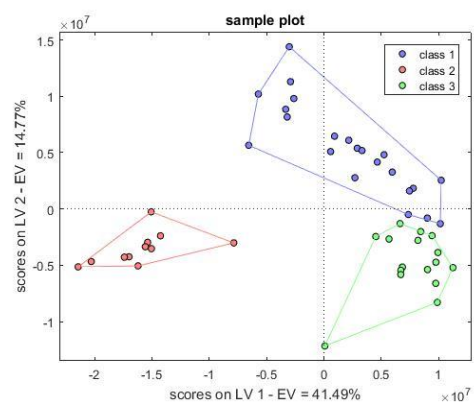


Hình 9. Kết quả phân nhóm kNN

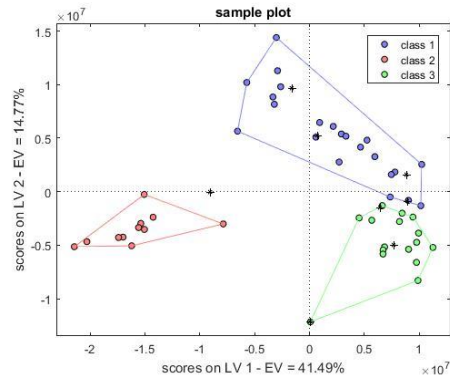
Giá trị sai số là 0,08 là một giá trị chấp nhận được. Kết quả kiểm tra 8 mẫu mật ong chưa biết nguồn gốc cho thấy 4 mẫu có nguồn gốc là mật ong nhãn, 4 mẫu còn lại không phải mật ong nhãn.

#### Kết quả phân nhóm bằng PLS-DA

Sau khi khảo sát các điều kiện tiền xử lý số liệu, số LVs cần thiết để chọn phương án thực hiện tối ưu, thu được đồ thị PLS-DA phân nhóm mẫu luyện và mẫu kiểm tra như Hình 10 và Hình 11.



Hình 10. Đồ thị PLS-DA phân nhóm 49 mẫu luyện



Hình 11. Đồ thị PLS-DA kiểm tra mẫu

Với tổng phần trăm phương sai lớn hơn 55%, kết quả này là chấp nhận được. Từ biểu đồ Hình 3.10, dễ thấy các nhóm có sự phân biệt khá rõ rệt.

#### Kết quả thực hiện LDA dựa trên dữ liệu phân cụm PCA

Từ kết quả phân cụm 49 mẫu luyện bằng PCA, tiến hành kiểm tra 8 mẫu chưa rõ nguồn gốc bằng thuật toán LDA thu nhận được ma trận P như Bảng 3.1.

Bảng 1. Kết quả phân tích LDA

| Mẫu | Nhóm 1 | Nhóm 2 | Nhóm 3 |
|-----|--------|--------|--------|
| 1   | 0      | 0      | 1      |
| 2   | 1      | 0      | 0      |
| 3   | 1      | 0      | 0      |
| 4   | 0      | 0      | 1      |
| 5   | 0      | 0      | 1      |
| 6   | 0      | 1      | 0      |
| 7   | 1      | 0      | 0      |
| 8   | 1      | 0      | 0      |

Kết quả kiểm tra 8 mẫu mật ong chưa biết nguồn gốc cho thấy 4 mẫu có nguồn gốc là mật

ong nhãn, 4 mẫu còn lại không phải mật ong nhãn (không thuộc nhóm 1).

### 3.4. So sánh kết quả phân loại nguồn gốc các mẫu mật ong kiểm tra

Kết quả phân loại nguồn gốc các mẫu mật ong kiểm tra bằng 4 phương pháp khác nhau (SIMCA, kNN, PLS-DA và LDA) được trình bày tóm tắt trong Bảng 2.

*Bảng 2. So sánh kết quả phân loại nguồn gốc các mẫu mật ong kiểm tra bằng các phương pháp thống kê đa biến khác nhau*

| Mẫu | SIMCA | kNN | PLS-DA | LDA |
|-----|-------|-----|--------|-----|
| 1   | 3     | 3   | 3      | 3   |
| 2   | 1     | 3   | 1      | 1   |
| 3   | 1     | 1   | 1      | 1   |
| 4   | 3     | 3   | 3      | 3   |
| 5   | 3     | 3   | 3      | 3   |
| 6   | 2     | 2   | 2      | 2   |
| 7   | 1     | 1   | 1      | 1   |
| 8   | 1     | 1   | 1      | 1   |

Kết quả thu được từ bốn phương pháp là tương đối phù hợp với nhau, chỉ riêng mẫu số 2 có kết quả phân nhóm khác nhau. Đối chiếu lại kết quả ở Hình 9 và Hình 12 cho thấy, mẫu này nằm ở vùng biên, có nghĩa là xác suất thuộc nhóm 1 hoặc 3 là tương đối xấp xỉ nhau; sai số này là chấp nhận được.

#### Bàn luận

Việc sử dụng bốn phương pháp phân tích nhóm SIMCA, kNN, LDA và PLS-DA để kiểm tra sự phân nhóm không chỉ dẫn của PCA cũng như kiểm tra nhóm của các mẫu chưa biết là thành công. Trong khi với kết quả PCA, chúng ta chỉ có thể gán nguồn gốc phân nhóm bằng mắt thường (không có xác suất thống kê), các phương pháp SIMCA, kNN, LDA và PLS-DA cung cấp kết quả phân nhóm có kèm đồ thị và xác suất, từ đó đưa ra các nhận xét chính xác hơn cũng như dễ dàng đánh giá nguyên nhân gây ra sai số.

Kết quả thu được cho thấy có thể xây dựng một công cụ kiểm tra nhanh các mẫu mật ong trên thị trường, là một công cụ hữu hiệu cho các cơ quan quản lý thị trường cũng như cho các công ty sản xuất mật ong tự kiểm định chất lượng sản phẩm đầu ra – đầu vào. Độ chính xác của phương pháp có thể dễ dàng được cải thiện khi bộ số liệu luyện được tăng thêm số mẫu đến khoảng vài nghìn hoặc lớn hơn.

### 4. KẾT LUẬN

Kết quả thu được cho thấy việc áp dụng các phương pháp thống kê đa biến kết hợp phân tích tín hiệu  $^1\text{H-NMR}$  của các mẫu mật ong khả thi, có thể phân loại các mẫu mật ong ở Việt Nam. Từ 49 mẫu luyện, thông qua sự phân nhóm không chỉ dẫn bằng PCA, đã phân được thành 3 nhóm lớn có sự phân cụm rõ rệt. Kết quả kiểm tra sự phân nhóm của PCA và gán nhóm cho 8 mẫu kiểm tra bằng các phương pháp SIMCA, kNN, PLS-DA và LDA cho kết quả phù hợp, chứng tỏ các mô hình thống kê áp dụng trong nghiên cứu này phù hợp để phân loại nguồn gốc mật ong ở Việt Nam.

### TÀI LIỆU THAM KHẢO

1. Bộ Y tế (2017), Dược điển Việt Nam V - chuyên luận Mật ong, trang 1246.
2. Vũ Thục Linh (2015), Báo cáo thị trường mật ong EU, Cục xúc tiến thương mại, Bộ công thương, trang 13-20.
3. TCVN 5269-1990, Mật Ong Tự Nhiên - Phương Pháp Xác Định Hàm Lượng Đường Sacaroza.
4. Stefan Bogdanov et al (2008), Carbohydrates and the Sweetness of Honey: a review, *American Journal of the College of Nutrition*, 27: 677-689.
5. Rainer Krell (1996), *Value- Added Products from Beekeeping*, Fao Agricultural Services Bulletin No. 124, chapter 2, Food and Agriculture Organization of the UN.
6. Diniz P. H. G. D. et al (2014), Simultaneous classification of teas according to their varieties and geographical origins by using NIR spectroscopy and SPA-LDA, *Food Analytical Methods* 7(8), pp. 1712-1718.
7. Diniz P.H.G. et al (2016), Using UV-Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup, *Food Chemistry* 192, pp. 374-379.
8. <http://michem.disat.unimib.it/chm/>
9. Trần Quang Khánh (2013), Matlab ứng dụng Tập 1+2, NXB khoa học và kỹ thuật, Hà Nội.
10. Marc Spiteri et al (2016) Combination of  $^1\text{H}$  NMR and chemometrics to discriminate Manuka honey from other floral honey types from Oceania *Food Chemistry*, 217:766-772.
11. Ioannis K. et al (2017), Characterization and geographical discrimination of commercial Citrus honeys produced in different Mediterranean countries based on minerals, volatile compounds and physicochemical parameters, using chemometrics, *Food Chemistry*, 217:445-455.