CÁC PHƯƠNG PHÁP PHÂN TÍCH THỐNG KÊ ĐA BIẾN SỐ LIỆU NGHIÊN CỨU LÂM NGHIỆP BẰNG SAS

Bùi Mạnh Hưng

Trường Đại học Lâm nghiệp

TÓM TẮT

Phân tích đa biến đã và đang chứng minh được nhiều ưu điểm nổi trội như: khai thác triệt để số liệu, kết quả phân tích toàn diện và khách quan hơn. SAS có thể thực hiện được nhiều phân tích đa biến khác nhau. Đầu tiên phải kể đến là phân tích thành phần chính. Phương pháp này có thể được áp dụng để phân tích mối quan hệ giữa các loài trong rừng tự nhiên. Các loài sẽ được phân thành 3 nhóm chính: đối kháng, đối kháng ít và không đối kháng. Phân tích thứ hai là tương quan chính tắc. Phân tích này có thể phân tích được mối tương quan giữa hai nhóm biến (nhóm X, nhóm Y). Điều này vượt trội hơn hẳn các phân tích tương quan đơn biến thường được áp dụng trước đây. Phân tích thứ ba là phân tích tương đồng. Phân tích tương đồng có thể tìm ra các loài ưu thế ở mỗi ô, đồng thời phân loại các ô có mức tương đồng về mức độ đa dạng sinh học loài thành các nhóm. Đây là cơ sở quan trọng để điều tiết tổ thành và nâng cao đa dạng sinh học tại khu vực nghiên cứu. Phân tích cuối cùng là phân tích phân nhóm. Phân tích này sẽ tạo thành các nhóm loài tương đồng, ít đối kháng. Ngoài ra nó sẽ cho biết phức độ biến động có thể được giải thích bởi các nhóm. Đó là cơ sở tốt để khẳng định độ tin cậy của các nhóm.

Từ khóa: Phân tích nhóm, phân tích thành phần chính, phân tích tương đồng, Sas, tương quan chính tắc.

I. ĐẶT VẤN ĐỀ

Việc xử lý số liệu trong nghiên cứu nói chung và trong Lâm nghiệp nói riêng là điều cực kỳ quan trọng. Bởi lẽ, phân tích số liệu là cơ sở để giúp các nhà nghiên cứu có những kết luận đúng đắn, chính xác, từ đó có những nhận định, cách nhìn và đề xuất phù hợp trong việc quản lý và phát triển tài nguyên rừng một cách bền vững (B.M. Hưng, 2016; S. Wagner, 2016).

Trong những năm gần đây, có nhiều phân tích thống kê đa biến đã được áp dụng như: phân tích tương quan đa biến, phân tích thành phần chính, phân tích hệ số đường ảnh hưởng, phân tích tương đồng, phân tích phân nhóm… đã được áp dụng nhiều trong các lĩnh vực nghiên cứu sinh thái học nói chung, trong đó có lâm nghiệp (S. Wagner, 2014; S. Wagner, 2016; U. Berger, 2008). Tuy nhiên, tại Việt Nam, việc ứng dụng các phương pháp phân tích này trong lĩnh vực lâm nghiệp còn rất hạn chế. Một nguyên nhân chính dẫn đến hạn chế này là thiếu các tài liệu hướng dẫn khai thác và ứng dụng các phần mềm thống kê mạnh cho phân tích số liệu nghiên cứu lâm nghiệp (B.M. Hưng và cộng sự, 2013; B.M. Hưng và cộng sự, 2017).

Phân tích đa biến đã và đang chứng minh được những ưu điểm nổi trội hơn các phương pháp đơn biến thường được áp dụng trước kia trong các nghiên cứu lâm nghiệp. Trước hết, nó khai thác được tổng hợp toàn bộ các biến, các số liệu mà chúng ta có, tránh việc lãng phí số liệu và công sức thu thập. Thứ hai, kết quả phân tích phản ánh toàn diện và khách quan hơn đối tượng mà các nhà nghiên cứu cần phân tích. Và vì thế, nó dẫn đến một ưu điểm cuối cùng là các đề xuất, kết luận sẽ trở lên chính xác và hiệu quả hơn.

Trong phân tích số liệu nói chung, có nhiều phần mềm tin học hỗ trợ rất mạnh cho việc xử lý số liệu nghiên cứu nói chung và số liệu lâm nghiệp nói riêng như: SPSS, Stata, R, M.S. Excel, Irristat, Minitab, Statgraphics... Tuy nhiên, qua quá trình nghiên cứu và sử dụng phần mềm SAS đã chứng minh được nhiều chức năng mới có giá trị cao trong phân tích số liệu nghiên cứu lâm nghiệp như: lập phân bố thực nghiệm cho đại lượng liên tục, hệ thống tiêu chuẩn phi tham số để so sánh các mẫu, hệ thống phân tích tương quan phi tuyến và đặc biệt là phân tích đa biến, đa mẫu (M. Marasinghe, 2008; C.Y. Joanne Peng, 2009; L.Q. Hưng, 2009; B.M. Hưng, 2011). Một ưu

điểm nội trội khác của SAS là việc viết và tạo lập các dòng lệch để phân tích số liệu. Điều này sẽ giúp việc phân tích số liệu lần tiếp theo, hoặc lặp lại ở một ô tiêu chuẩn khác được thực hiện một cách rất dễ dàng và nhanh chóng.

Với những lý do như trên, bài báo này sẽ trình bày một cách cụ thể các phương pháp phân tích thống kê đa biến với sự hỗ trợ bởi SAS; qua đó cho thấy sự cần thiết và hữu ích trong việc ứng dụng phần mềm này trong phân tích số liệu nghiên cứu lâm nghiệp, giúp việc phân tích số liệu được hiệu quả, nhanh chóng và chính xác. Phương pháp phân tích thống kê đa biến sẽ khắc phục được những nhược điểm của Excel và một số phần mềm khác.

II. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Phương pháp nghiên cứu tài liệu và số liệu chọn lọc

Một số tài liệu hướng dẫn sử dụng SAS cũng như phân tích thống kê đa biến trong SAS được thu thập, phân tích một cách có chọn lọc. Các tài liệu phân tích về lĩnh vực lâm nghiệp được ưu tiên hàng đầu. Sau đó tới các lĩnh vực gần gũi hơn như quản lý tài nguyên rừng, quản lý môi trường, chế biến gỗ và kinh tế lâm nghiệp. Các tài liệu được tập hợp và phân tích theo cơ sở lý thuyết về phân tích bằng SAS, thành tựu và những kết quả đã đạt được trong lĩnh vực phân tích số liệu nghiên cứu lâm nghiệp bằng SAS (V.C. Đàm, 1999).

Số liệu được kế thừa từ những nghiên cứu trước, với sự đồng ý của các tác giả giữ quyền sở hữu các bộ số liệu đó. Số liệu tập trung chủ yếu về các lĩnh vực trong lâm nghiệp như: Điều tra quy hoạch, Lâm học, Lâm nghiệp xã hội…

2.2. Phương pháp thử nghiệm và so sánh

Từ việc thống kê, phân tích các trình lệnh, quy trình được sử dụng để phân tích đa biến với sự hỗ trợ của SAS, các trình lệnh cho phân tích số liệu lâm nghiệp được xây dựng một cách tỉ mỉ, chính xác. Tiếp đó, các trình lệnh được chạy thử với các bộ số liệu lâm nghiệp. Sau đó, kết quả xuất ra được kiểm tra, đánh giá và so sánh với kết quả xuất ra của các phần mềm khác như Spss, Stata và R. Từ đó, chọn ra được quy trình chính xác, hiệu quả cho phân tích đa biến số liệu lâm nghiệp (B.M. Hưng và cộng sự, 2013).

III. KẾT QUẢ NGHIÊN CỨU

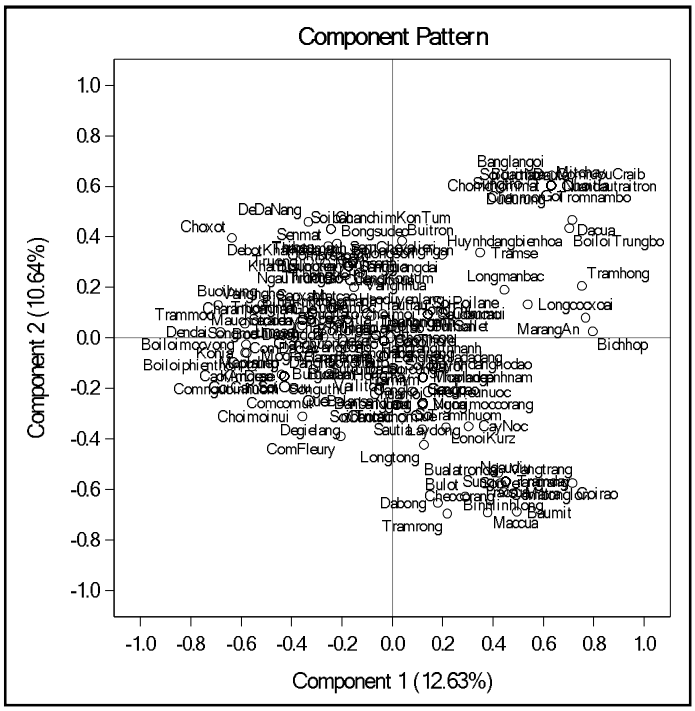
3.1. Phân tích thành phần chính (Principal Component Analysis)

Phân tích thành phần chính (PCA) là một phân tích đa biến rất quan trọng trong phân tích số liệu. Đây là phương pháp nhóm các đối tượng phân tích. Phân tích thành phần chính rất hữu ích khi bảng dữ liệu có nhiều biến tham gia. Phương pháp này sẽ giúp tìm ra được các thành phần nào là chính trong bảng dữ liệu. Những nhân tố này sẽ đóng góp phần lớn vào sự biến động của tập dữ liệu. Nguyên lý của PCA khá đơn giản, trước hết PCA sẽ dò ra hướng nào có biến động nhiều nhất trong tập dữ liệu. Sau đó PCA sẽ xoay trục hoành theo hướng đó và trục tung theo hướng vuông góc còn lại (A.M.C. Davies và cộng sự, 2017). Đây là cơ sở để chúng ta có thể loại bớt các biến, các nhân tố không cần thiết, không quan trọng trong tập dữ liệu. Đồng thời phân loại được nhóm các nhân tố đối kháng, ít đối kháng và đối kháng mạnh.

PCA có nhiều ứng dụng, tuy nhiên một ứng dụng khá phổ biến là để phân tích quan hệ giữa các loài trong rừng tự nhiên. Để chạy được ứng dụng này, các lệnh sau được thực hiện:

```
proc princomp data=WORK.IMPORT5 plots(only ncomp=2)=(pattern);
        var“Tên biến của các loài”;
run;
```

Ứng dụng sau đây cho thấy PCA có thể phân loại được các loài cây ra thành các nhóm: đối kháng, đối kháng ít và đối kháng mạnh. Ví dụ như Chò xót và Dẻ đà nẵng thường chung sống cùng nhau và không đối kháng. Chúng đối kháng ít với Da cua, Bởi lời trung bộ, Chòi mòi núi và Côm Fleury. Tuy nhiên, chúng rất đối kháng với các loài: Côi rào, Bầu mít, Mặc cưa hay Trâm rộng… Vì vậy, khi gây tạo rừng trồng với các loài tự nhiên, cần tránh các loài đối kháng và cần tập trung vào các loài không đối kháng, đó là cơ sở sinh lý tự nhiên rút ra được từ các quần thể thực vật. Điều này được thể hiện trong biểu đồ PCA (hình 01).



Hình 01. Biểu đồ phân tích PCA cho các loài rừng tự nhiên

3.2. Phân tích tương quan chính tắc (Canonical Correlation)

Phân tích tương quan chính tắc (CC) được sử dụng để phân tích mối quan hệ giữa hai tập biến. Tuy nhiên, CC không xác định đâu là tập biến độc lập, đâu là tập biến phụ thuộc. CC sẽ lập một tập biến chính tắc (canonical variates). Đây là tập hợp tuyến tính các biến để giải thích tốt nhất cho mối quan hệ giữa hai tập biến, tập gọi là tập biến X và tập biến Y. CC sẽ tạo ra hai biến chính tắc đầu tiên, thường ký hiệu là W_1 và V_1. Trong đó: W_1 là tổ hợp tuyến tính của các biến trong nhóm X và V_1 là tổ hợp tuyến tính của các biến trong nhóm Y. Sau đó CC sẽ tạo tiếp các biến chính tắc tiếp theo. Số lượng biến chính tắc bằng với số lượng biến trong tập biến nhỏ hơn. Kết quả phân tích tương quan chính tắc sẽ cho chúng ta thấy mối quan hệ chặt hay không chặt giữa hai nhóm biến X và Y nhờ vào hệ số tương quan bình phương giữa W_1 và V_1, đồng thời kiểm định sự tồn tại của mô hình thông qua tiêu chuẩn F. Biểu đồ tương quan giữa biến chính tắc W_1 và V_1 cũng được tạo ra để có cái nhìn trực quan hơn về mối quan hệ giữa hai tập biến X và Y (Robert M. Thorndike, 2000). Ngoài ra, CC còn cho chúng ta thấy được mối quan hệ giữa các biến trong từng nhóm biến và giữa các nhóm biến khác nhau (Rodrigo Loureiro Malacarne, 2014; Richard A. Johnson and Dean W. Wichern, 2007).

Quy trình thực hiện trong SAS để thực hiện phân tích tương quan chính tắc như sau:

```
proc cancorr data=WORK.IMPORT4 out=Work._tempout;
        /*** The VAR statement defines Variable set 1 ***/
        var dtnn dtln tuoi songuoi;
        /*** The WITH statement defines Variable set 2 ***/
        with thunhap hocluc;
run;
proc sgrender data=Work._tempout template="squareplot";
run;
proc delete data=Work._tempout;
run;
```

Trong ứng dụng dưới đây, từ số liệu điều tra xã hội học của các hội gia đình, muốn phân tích mối quan hệ giữa tập biến Y bao gồm: thu nhập bình quân của hộ gia đình và trình độ học vấn của hộ với tập biến X gồm: diện tích đất nông nghiệp, diện tích đất lâm nghiệp, độ tuổi và số người lao động trong gia đình. Kết quả phân tích mối quan hệ giữa hai nhóm biến được như bảng 01.

Bảng 01. Kết quả phân tích hồi quy chính tắc giữa hai nhóm biến X, Y

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0,343989	0,295846	0,082941	0,118329	0,1342	0,1136	0,8667	0,8667
2	0,142187	0,092902	0,092170	0,020217	0,0206		0,1333	1,0000

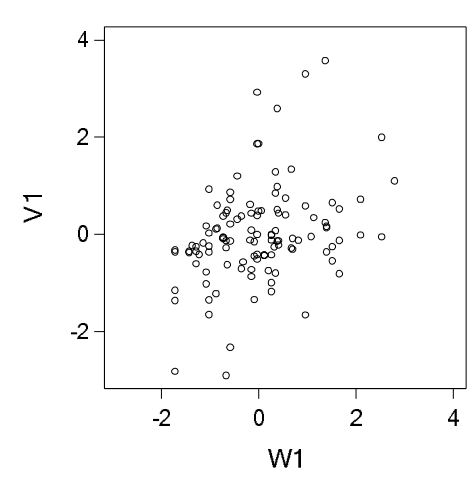
Kết quả bảng trên cho thấy tương quan giữa hai nhóm biến X và Y không chặt. Kết quả R2 là 0,11. Tức là chỉ 11% biến động của nhóm Y được diễn tả bởi nhóm X.

Kết quả phân tích mối quan hệ giữa các biến thuộc nhóm X được trình bày trong bảng sau. Kết quả bảng sau cho thấy rằng mối tương quan giữa các biến là rất lỏng lẻo. Chỉ duy nhất giữa diện tích đất nông nghiệp và số người trong gia đình là tương đối lớn (R = -0,4247). Tuy nhiên, quan hệ giữa hai biến này lại nghịch biến, tức là nếu số người tăng lên trong mỗi gia đình thì diện tích đất nông nghiệp lại giảm đi. Lý do cho kết quả này là nhiều lao động trong các hộ gia đình không làm nông nghiệp mà làm các ngành nghề khác.

Bảng 02. Kết quả phân tích hồi qui giữa các biến thuộc nhóm X

Correlations Among the Regression Coefficient Estimates				
	dtnn	**dtln**	**tuoi**	**songuoi**
dtnn	1,0000	-0,0025	-0,2617	-0,4247
dtln	-0,0025	1,0000	-0,0292	-0,0669
tuoi	-0,2617	-0,0292	1,0000	0,0008
songuoi	-0,4247	-0,0669	0,0008	1,0000

Biểu đồ tương quan giữa hai biến chính tắc đầu tiên được tạo ra trong các nhóm X và Y được trình bày như sau. Biểu đồ một lần nữa cho thấy tương quan giữa hai nhóm biến là lỏng lẻo, không thực sự chặt. Bởi lẽ, các điểm lằm rải rác, không tập trung và hình thành một xu hướng nào cả.



Hình 02. Biểu đồ thể hiện mối tương quan giữa hai biến chính tắc đầu

3.3. Phân tích tương đồng (Correspondence Analysis)

Phân tích tương đồng (CA) là một phương pháp phân tích đa biến. Phương pháp này được phát triển bởi Hirschfeld, sau đó được kế thừa và phát triển tiếp bởi Jean-Paul Benzécri. CA thường được áp dụng cho các biến rời rạc, thứ bậc, hơn là các biến liên tục.

Các bước cơ bản của phân tích tương đồng là (P.M. Yelland, 2010; J.C. Epidemiol, 2010):

- Bước 1: Thành lập bảng số liệu bao gồm hai nhóm biến X và Y. Sau đó sẽ tính toán giá trị tần số ở mỗi tổ của nhóm biến X và nhóm biến Y.

- Bước 2: Tính toán giá trị khoảng cách giữa hai biến cho từng ô, theo dòng, tạo nên ma trận khoảng cách bằng công thức sau:

$$K(X,Y) = \sqrt{\sum_{j=1}^{i}\left(\frac{\left(F_{ij}-F_{i'j}\right)^2}{F_j}\right)} \qquad (1)$$

Trong đó:

K(X,Y) là giá trị khoảng cách giữa hai nhóm biến X và Y;

F_{ij} là giá trị lũy tích tương ứng dòng thứ i và cột j;

$F_{i'j}$ là giá trị lũy tích tương ứng dòng thư i' và cột j;

F_j là tổng giá trị tương ứng ở cột thứ j.

- Bước 3: Tính điểm cho các dòng. Phân tích tương đồng sẽ sử dụng phương pháp biểu đồ để thể hiện ma trận khoảng cách tính toán ở bước 2. Trong đó, các dòng biểu thị bởi các điểm. Vì vậy, khoảng cách giữa các điểm chính là giá trị khoảng cách giữa các dòng. Sau đó, từ tọa độ của các điểm sẽ tính toán được điểm cho mỗi dòng.

- Bước 4: Vẽ biểu đồ. Hai thành phần đầu tiên của mỗi dòng điểm được sử dụng để vẽ biểu đồ dạng 2 chiều. Biểu đồ sẽ phân xác biến trong nhóm X và Y thành 4 nhóm, nằm tại 4 cung phần tư. Từ thông tin thu được ở 4 cung phần tư, cho phép kết luận về mối quan hệ giữa các biến trong nhóm X với từng biến trong nhóm Y, cũng như các biến trội trong nhóm X tương ứng với từng biến trong nhóm Y. Đồng thời, có thể kết luận về các biến trong từng nhóm X và Y có tính tương đồng cao hơn.

Để thực hiện phân tích tương đồng thì các lệnh sau cần được thực hiện trong SAS:

```
proc corresp data=WORK.IMPORT1 dimens=2 plots;
        varTên các biến;
        idTên biến loài;
run;
```

Ví dụ dưới đây được áp dụng cho việc phân tích mối quan hệ giữa hai nhóm biến là ô tiêu chuẩn I (OTC) và nhóm biến tên loài. Từ đó có thể tìm được loài ưu thế tại mỗi ô, cũng như phân nhóm được các ô có mức độ tương đồng về đa dạng sinh học cao hơn.

Phương pháp này ưu điểm hơn những phân tích truyền thống ở chỗ kết quả sẽ phản ánh toàn một cánh toàn diện hệ trạng thái, vì dựa

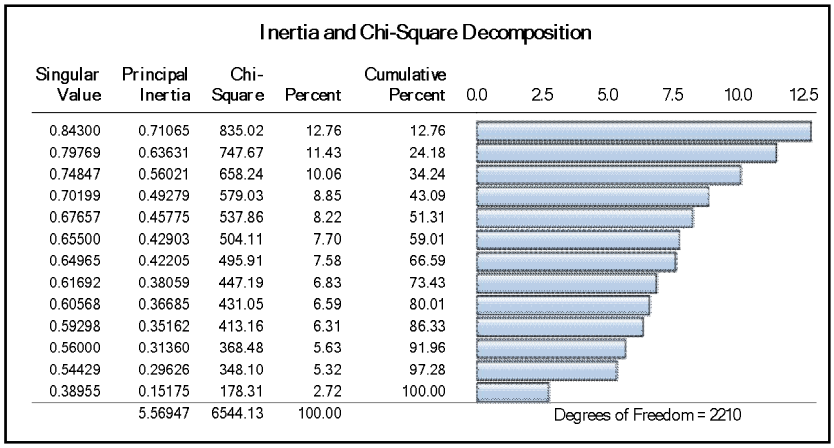
vào số liệu của nhiều ô. Ngoài ra, các phân tích truyền thống dựa vào số cây và tỷ lệ số cây của mỗi loài không phân loại được các ô có mức độ đa dạng sinh học tương tự nhau (Palmer, 2017; Murtagh, 2016).

Kết quả tính toán tiêu chuẩn χ^2 để phân tích mối quan hệ giữa hai biến OTC và loài cây được thể hiện trong bảng sau. Kết quả cho thấy rằng giữa hai biến thực sự tồn tại mối quan hệ.

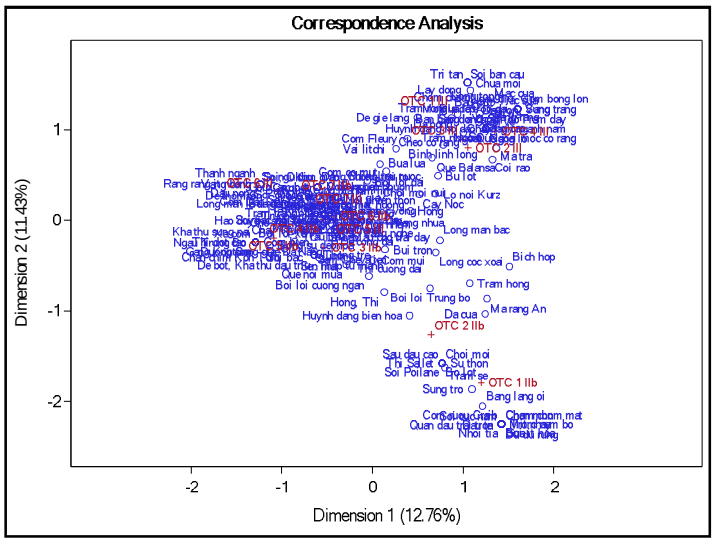
Bảng 03. Kết quả tính toán tiêu chuẩn χ^2

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.84300	0.71065	835.02	12.76	12.76
0.79769	0.63631	747.67	11.43	24.18
0.74847	0.56021	658.24	10.06	34.24
0.70199	0.49279	579.03	8.85	43.09
0.67657	0.45775	537.86	8.22	51.31
0.65500	0.42903	504.11	7.70	59.01
0.64965	0.42205	495.91	7.58	66.59
0.61692	0.38059	447.19	6.83	73.43
0.60568	0.36685	431.05	6.59	80.01
0.59298	0.35162	413.16	6.31	86.33
0.56000	0.31360	368.48	5.63	91.96
0.54429	0.29626	348.10	5.32	97.28
0.38955	0.15175	178.31	2.72	100.00
	5.56947	6544.13	100.00	



Ngoài ra, CA còn cung cấp biểu đồ phân loại các loài, các ô và tương quan giữa loài và ô như trong hình 03.



Hình 03. Kết quả phân nhóm loài, ô và các ô tương đồng

Kết quả trong hình 03 cho thấy rằng các loài ưu thế của trạng thái IIb chủ yếu là Sung trổ, Bằng lăng ổi, Bời lời trung bộ, Da cu, Bời lời cuống ngắn, Dẻ đà nẵng… trong khi đó, với rừng III loài ưu thế chủ yếu là: Mặc cưa, Côi trào, Ma trá, Côm đắk lắk…

Về mức độ đa dạng sinh học không thực sự có sự khác biệt giữa hai trạng thái, kết luận này một lần nữa được khẳng định, bởi lẽ có nhiều ô của rừng IIb và rừng III có cùng loài ưu thế và tổ thành loài cũng tương tự nhau, do vậy chúng cùng được phân vào một nhóm, điều này có thể thấy được trong góc phần tư thứ II.

Tuy nhiên, nếu xét ở nhóm nhỏ hơn, chi tiết, dựa vào loài ưu thế và tổ thành loài của các ô thì mức độ đồng nhất giữa các ô cùng trạng thái lớn hơn nhiều so với các ô khác trạng thái. Điều này được chứng minh trong 3 góc phần tư I, III và IV trong hình trên. Như vậy, có thể thấy rằng, việc lựa chọn vị trí điều tra của các ô trong cùng một trạng thái là tương đối tốt về mặt loài cây và vì thế số liệu thu thập đạt độ tin cậy cao.

3.4. Phân tích nhóm (Cluster Variables)

Phương pháp phân tích phân nhóm (CV) dựa vào ma trận khoảng cách giữa các dòng ở tương ứng ở từng cột. Các bước cơ bản như sau (T. Lee và cộng sự, 2008):

- Bước 1: Tính toán ma trận khoảng cách của các biến;
- Bước 2: Áp dụng thuật toán phân nhóm cho ma trận khoảng cách vừa thu được;
- Bước 3: Phân thành các nhóm. Mỗi nhóm sẽ bao gồm các biến đồng nhất với nhau;
- Bước 4: Tính toán thành phần nhóm thứ nhất cho mỗi nhóm.

Phân tích nhóm thứ bậc có thể được sử dụng để phân tích mối quan hệ giữa các loài. Về nguyên lý, phân tích thứ bậc sẽ phân các loài xuất hiện cùng nhau và có số lượng cá thể tương đương nhau vào cùng một nhóm. Dựa vào số liệu cá thể của mỗi loài ở các ô, phân tích thứ bậc sẽ tạo ra ma trận khoảng cách, các loài có khoảng cách trung bình với các loài khác nhỏ thì được xếp vào một nhóm, các loài có khoảng trung bình lớn thì sẽ tách thành một nhóm khác (Oksanen và các cộng sự, 2016). Để thực hiện nội dung phân tích này trong SAS thì các lệnh sau được thực hiện.

Lệnh chạy trong SAS:

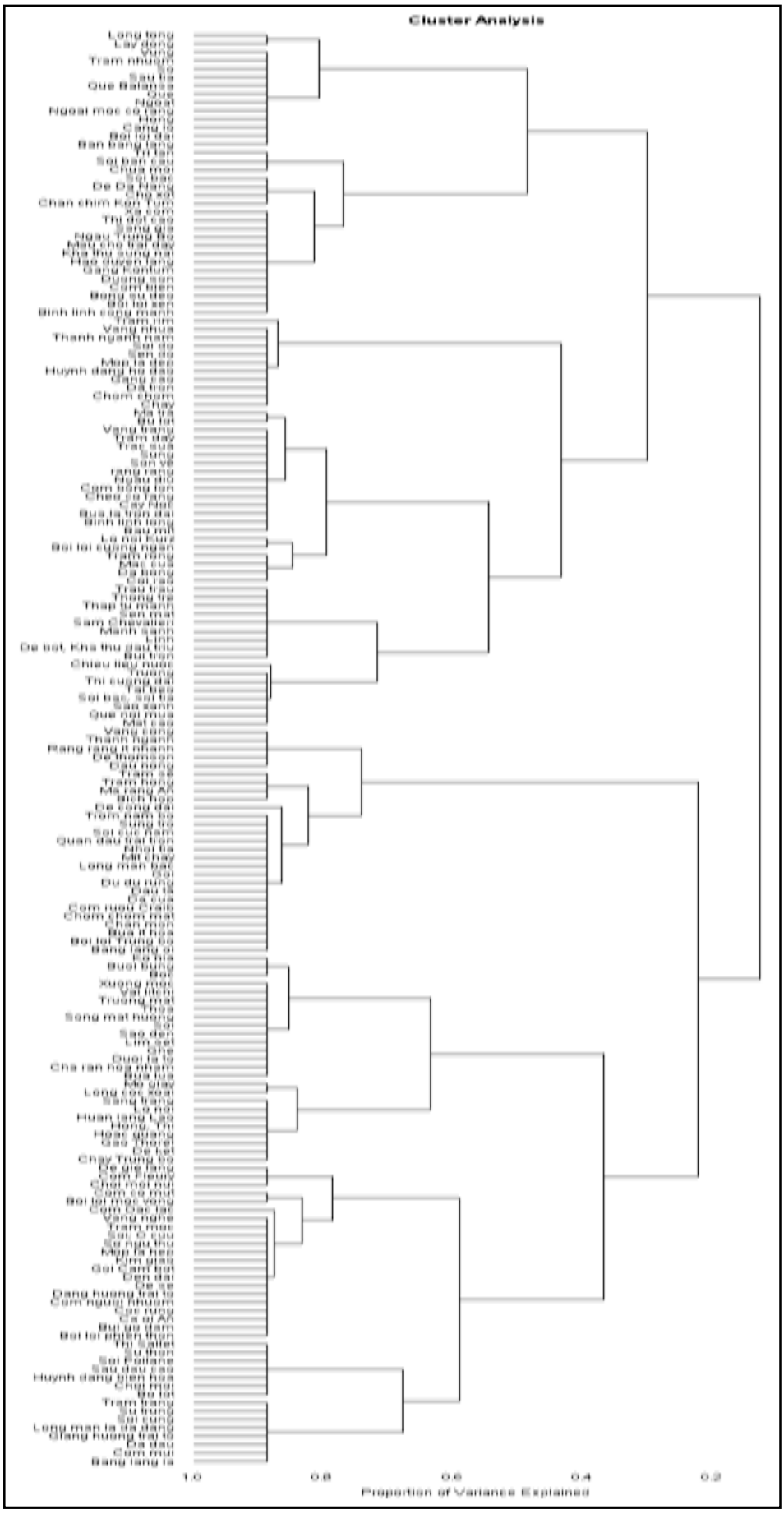
```
proc varclus data=WORK.IMPORT hierarchy plots;
        var Tên các loài;
run;
```

Với số liệu đầu vào là các loài trong rừng tự nhiên, thì kết quả phân nhóm và tỷ lệ biến có thể được giải thích bởi các nhóm như bảng 04. Như vậy, với 28 nhóm có thể giải thích tới 88,63% số liệu thực cần kiểm tra, do vậy độ tin cậy của các nhóm là rất cao.

Bảng 04. Kết quả tính toán phương sai được giải thích bằng các nhóm

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	21.598738	0.1263	0.1263	18.192509	0.0000	
2	37.962014	0.2220	0.2150	14.349161	0.0015	0.9990
3	51.202034	0.2994	0.2150	13.282871	0.0042	1.0326
4	62.767409	0.3671	0.2636	12.750986	0.0011	1.0359
5	73.931556	0.4323	0.3138	10.362658	0.0011	1.2108
6	82.806384	0.4842	0.3995	10.250255	0.0160	1.3147
7	92.927573	0.5434	0.3995	8.161210	0.0159	1.3147
8	100.831495	0.5897	0.4114	8.045070	0.0159	1.6954
9	108.406836	0.6340	0.4973	7.395710	0.0159	7.2717
10	115.645378	0.6763	0.5218	6.738627	0.0159	8.1875
11	122.323184	0.7153	0.5218	5.272939	0.0159	14.603
12	126.430594	0.7394	0.6399	5.034271	0.0159	14.685
13	131.343500	0.7681	0.6533	3.202249	0.0159	14.685
14	134.306853	0.7854	0.6769	2.137728	0.0283	24.278
15	135.956806	0.7951	0.6769	2.046481	0.0283	24.278
16	137.753808	0.8056	0.6408	1.781546	0.0283	24.278
17	139.176678	0.8139	0.6408	1.777455	0.0283	14.685
18	140.687956	0.8227	0.6408	1.728825	0.0283	14.685
19	142.214079	0.8317	0.6408	1.480002	0.0271	14.685
20	143.444178	0.8389	0.6408	1.442020	0.0271	14.685
21	144.711209	0.8463	0.6408	1.430204	0.0271	14.685
22	145.716423	0.8521	0.7081	1.242435	0.0271	14.685
23	146.780875	0.8584	0.7081	1.179482	0.0271	14.685
24	147.696386	0.8637	0.7081	1.110132	0.0271	14.685
25	148.670995	0.8694	0.7081	1.091138	0.0314	14.685
26	149.609593	0.8749	0.7081	1.081336	0.0314	14.685
27	150.578857	0.8806	0.7081	1.055805	0.0314	14.685
28	151.552876	0.8863	0.7222	0.986217	0.0821	8.1650

Biểu đồ phân nhóm các loài cây được thể hiện trong hình 04.



Hình 04. Biểu đồ phân loại nhóm loài

Hình 04 cho thấy các loài được sắp xếp thành các nhóm nhỏ. Các loài trong cùng một nhóm nhỏ là các loài không đối kháng. Chúng hỗ trợ nhau cùng phát triển và cùng xuất hiện trong một ô. Ví dụ: Vàng nhựa, Mảnh sành, Bùi tròn, Trường, Dẻ đà nẵng, Sến mật và Sầm là một nhóm thường xuất hiện cùng nhau. Hay Thành ngạch nam, Găng cao, Sồi đỏ, Mớp lá đẹp và Côm có mật là một nhóm nhỏ khác thường chung sống cùng nhau. Do vậy, khi phục hồi rừng với mục đích nâng cao đa dạng sinh học thì cần tập trung chọn các loài tại các nhóm khác nhau, đó là cơ sở tốt cho phục hồi rừng, nâng cao đa dạng sinh học.

IV. KẾT LUẬN

Trong những năm gần đây, rất nhiều các phương pháp phân tích đa biến đã được áp dụng nhiều trong các lĩnh vực nghiên cứu sinh thái học nói chung, trong đó có lâm nghiệp (S. Wagner, 2014; S. Wagner, 2016; U. Berger, 2008). Bởi lẽ phân tích đa biến đã chứng minh được nhiều ưu điểm nổi trội như: khai thác triệt để số liệu, kết quả phân tích toàn diện và khách quan hơn, vì vậy những đề xuất sẽ hiệu quả và chính xác hơn. Tuy nhiên, tại Việt Nam, việc ứng dụng các phương pháp phân tích này trong lĩnh vực lâm nghiệp còn rất hạn chế. Nguyên nhân chính là còn hạn chế trong hướng dẫn và khai thác sử dụng các phần mềm phân tích số liệu mạnh hiện nay.

Trong phân tích số liệu nói chung, có nhiều phần mềm tin học hỗ trợ rất mạnh cho việc xử lý số liệu nghiên cứu nói chung và số liệu lâm nghiệp nói riêng như: Spss, Stata, R, M.S. Excel, Irristat, Minitab, Statgraphics... Tuy nhiên, qua quá trình nghiên cứu và sử dụng phần mềm SAS đã chứng minh được nhiều chức năng mới có giá trị cao trong phân tích số liệu nghiên cứu lâm nghiệp, đặc biệt là phân tích đa biến, đa mẫu (M. Marasinghe, 2008; C.Y. Joanne Peng, 2009; L.Q. Hưng, 2009; B.M. Hưng, 2011).

Kết quả nghiên cứu đã cho thấy rằng SAS có thể thực hiện được phần lớn các phương pháp phân tích đa biến hiện nay. Trước hết, SAS có thể thực hiện phân tích thành phần chính. Phương pháp này có thể được áp dụng để phân tích mối quan hệ giữa các loài trong rừng tự nhiên. Các loài sẽ được phân thành 3 nhóm chính: đối kháng, đối kháng ít và không đối kháng. Phân tích thứ hai có thể thực hiện trong SAS là tương quan chính tắc. Phân tích này có thể phân tích được mối tương quan giữa hai nhóm biến (nhóm X, nhóm Y). Điều này vượt trội hơn hẳn các phân tích tương quan đơn biến thường được áp dụng trước đây. Phân tích thứ ba là phân tích tương đồng. Phân tích này có khả năng ứng dụng cao trong phân tích số liệu rừng tự nhiên. Cụ thể, phân tích tương đồng có thể tìm ra các loài ưu thế ở mỗi ô, đồng thời phân loại các ô có mức tương đồng về mức độ đa dạng sinh học loài thành các nhóm. Đây là cơ sở quan trọng để điều tiết tổ thành và nâng cao đa dạng sinh học tại khu vực nghiên cứu. Phân tích cuối cùng được trình bày trong bài báo này là phân tích phân nhóm. Phân tích phân nhóm sẽ tạo thành các nhóm loài tương đồng, ít đối kháng. Ngoài ra nó sẽ cho biết phức độ biến động có thể được giải thích bởi các nhóm. Đó là cơ sở tốt để khẳng định độ tin cậy của các nhóm.

V. TÀI LIỆU THAM KHẢO

1. Bui Manh Hung and Bui The Doi (2017). Applying linear mixed model (LMM) to analyze forestry data, checking autocorrelation and random effects, using R. *Journal of Forestry Science and technology*, 2(2017): p. 17-26.

2. L.Q. Hưng (2009). *Ứng dụng SAS phân tích số liệu thí nghiệm*. Đại học Nông Lâm TP. Hồ Chí Minh.

3. Ngô Đăng Phong, Huỳnh T. Thùy Trang, Nguyễn Duy Năng, Trần Văn Mỹ, Trần Hoài Thanh (2013). *Hướng dẫn sử dụng Mstatc, Sas và Excel 2007 trong xử lý thí nghiệm cho ngành nông nghiệp và quản lý nước*. Trường Đại học Nông Lâm TP. Hồ Chí Minh.

4. Vũ Cao Đàm (1999). *Phương pháp nghiên cứu khoa học*. NXB. Khoa học và Kỹ thuật.

5. A. M. C. Davies and Tom Fearn (2017). *Back to*

basics: the principles of principal component analysis. Spectroscopy Europe and Asia, pp. 20-23.

6. Robert M. Thorndike (2000). *Canonical correlation analysis*, in *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. Academic Press.

7. Rodrigo Loureiro Malacame (2014). *Canonical Correlation Analysis. The Mathematica Journal*, 16(2014): p. 1-22.

8. Richard A. Johnson and Dean W. Wichern (2007). *Applied Multivariate Statistical Analysis*. Pearson Education, Inc.

MULTIVARIATE ANALYSIS METHODS FOR FORESTRY RESEARCH DATA, USING SAS

Bui Manh Hung

Vietnam National University of Forestry

SUMMARY

Multivariate analysis has been shown to have many outstanding advantages such as full exploitation of data, more comprehensive analysis and more objective results. SAS can perform a variety of multivariate analyzes. First of all, it is the principal component analysis. This method can be applied to analyze relationships among species in natural forests. The species will be classified into three main groups: resistance, minor resistance and non-resistance. The second analysis is a canonical correlation. This analysis can analyze the correlation between two groups of variables (group X, group Y). This surpasses the previous regression analysis. The third analysis is the correspondence analysis. The correspondence analysis can identify dominant species in each plot and classify plots with similar levels of species biodiversity into groups. This is an important basis for regulating and enhancing biodiversity in a region. The final analysis is the cluster analysis. This analysis will form similar, less antagonistic groups. In addition, it will indicate the variation that can be explained by the clusters. That is an excellent basis for asserting the significance of groups.

Keywords: Canonical correlation, cluster analysis, correspondence analysis, principle component analysis, Sas.

Ngày nhận bài	**: 02/8/2017**
Ngày phản biện	**: 30/8/2017**
Ngày quyết định đăng	**: 08/9/2017**