J. Sci. & Devel. 2014, Vol. 12, No. 5: 762-768 Tạp chí Khoa học và Phát triển 2014, tập 12, số 5: 762-768
www.vnua.edu.vn

PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH VÀ PHÂN TÍCH CHÙM TRONG XỬ LÍ SỐ LIỆU THỐNG KÊ NHIỀU CHIỀU

Nguyễn Hữu Du

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Email: namtoha@gmail.com

Ngày gửi bài: 06.05.2014 Ngày chấp nhận: 15.07.2014

TÓM TẮT

Một vấn đề quan trọng đặt ra trong công tác nghiên cứu thực nghiệm là phân tích và xử lí những dữ liệu thu thập được. Nếu bảng số liệu thu thập được lớn thì việc tìm hiểu thông tin từ đó là khá khó khăn và phức tạp. Bài báo trình bày hai phương pháp hiệu quả trong xử lí số liệu nhiều chiều, gồm phân tích thành phần chính và phân tích chùm. Sau đó áp dụng hai phương pháp này để phân tích một bộ dữ liệu cụ thể từ một đề tài khoa học trong nông nghiệp cũng như đưa ra những nhận xét, đánh giá từ dữ liệu đã được xử lí.

Từ khóa: Phân tích thành phần chính, phân tích chùm, xử lí số liệu.

Methods of Principal Component Analysis and Cluster Analysis in Multi-Dimension Statistics Data Processing

ABSTRACT

A crucial issue posed in practical research is analyzing and processing of collected data. If collected data is large, examining the information will be relatively complex and troublesome. This article presented two efficient methods in multi-dimension data processing, principal component analysis and cluster analysis. These methods were successfully tested to analyze a set of data from a research project in agronomy.

Keywords: Cluster analysis, data processing, principal component analysis.

1. ĐẶT VẤN ĐỀ

Mỗi bộ dữ liệu thu thập được khi tiến hành các nghiên cứu, thí nghiệm thường được thể hiện dưới dạng bảng các giá trị số của nhiều cá thể. Chúng tạo thành "đám mây số liệu" khá phức tạp. Các số liệu này cần được phân tích và xử lí để có thể rút ra được những nhận xét, đánh giá thích hợp.

Phân tích thành phần chính là kĩ thuật biểu diễn số liệu dựa theo các tiêu chuẩn về đại số và hình học mà không đòi hỏi một giả thuyết thống kê hay mô hình đặc biệt nào. Lĩnh vực áp dụng của phân tích thành phần chính rất rộng trong nông nghiệp, kinh tế, khoa học cơ bản.

Phân tích chùm là kĩ thuật ghép các điểm quan sát lại thành nhóm theo một tiêu chí nào đó, tương tự như trong cách phân loại trong sinh học. Việc phân tích có thuật toán đơn giản, đồng thời đem lại cái nhìn trực quan của phân loại thu được nên dễ được các nhà chuyên môn trong các ngành khoa học khác nhau chấp nhận.

Bài báo trình bày về hai phương pháp nói trên trong xử lí số liệu thống kê nhiều chiều. Sau đó đưa ra ví dụ phân tích cụ thể số liệu từ một đề tài khoa học nông nghiệp. Đây là hai phương pháp đơn giản nhưng có tính hiệu quả cao trong số nhiều phương pháp phân tích số liệu đã được đưa ra bởi các nhà thống kê, tuy nhiên việc ứng dụng chúng trong nghiên cứu

thực nghiệm, nhất là các đề tài thuộc lĩnh vực nông nghiệp còn hạn chế. Hi vọng bài báo này phần nào giúp các nhà chuyên môn thấy được sự hữu ích của việc áp dụng các kiến thức thống kê trong việc nghiên cứu của mình.

2. PHÂN TÍCH THÀNH PHẦN CHÍNH

2.1. Bảng số liệu

Cho bảng số liệu:

$$X_{n,p} = \begin{bmatrix} x_{11} & x_{12} ..... & x_{1p} \\ x_{21} & x_{22} ..... & x_{2p} \\ ..... & & \\ x_{n1} & x_{n2} ..... & x_{np} \end{bmatrix}$$

trong đó x_{ij} là giá trị mà biến X_j, $j = \overline{i,p}$ nhận trên cá thể thứ $i, i = \overline{1,n}$. Ta có một đám mây n điểm trong không gian R^p trong đó điểm x_i có tọa độ $x_i = \left(x_{i1}, x_{i2}, ..., x_{ip}\right)$, $i = \overline{1,n}$ và gọi là điểm - cá thể i. Không gian R^p gọi là không gian các cá thể.

Tương tự, ta có không gian R^n các điểm - biến, trong đó ta có p điểm biến

$$X_j = \left(x_{1j}, x_{2j}, ..., x_{nj}\right), \ j = \overline{1,\mathrm{p}}$$

Điều này có nghĩa là trong bảng số liệu, các cột là các biến và các dòng là các cá thể, trên đó đo giá trị các biến.

2.2. Tìm các thành phần chính

Mục đích của phân tích thành phần chính là rút ra thông tin chủ yếu chứa trong bảng số liệu bằng cách xây dựng một biểu diễn đơn giản hơn sao cho đám mấy số liệu được thể hiện rõ nhất. Cụ thể hơn, phân tích thành phần chính tức là đi tìm những trục hay mặt phẳng "phản ánh" tốt nhất, trung thực nhất đám mây điểm - biến, điểm - cá thể.

Việc tìm các thành phần chính (trục chính) được thực hiện như sau:

Tìm trục chính thứ nhất là trục mà quán tính nhỏ nhất, tức là đường thẳng qua tâm gần đám mây điểm nhất.

Trục chính thứ hai là trục qua tâm trực giao với trục chính thứ nhất và quán tính của đám mây theo nó là nhỏ nhất.

Trục chính thứ ba là trục qua tâm, trực giao với hai trục chính thứ nhất và thứ hai và gần đám mây nhất sau hai trục trên.

Tiếp tục như vậy đến bước thứ q$\left(q \le p, n\right)$, ta được một hệ q vectơ trực giao tạo thành siêu phẳng q chiều mà đám mây thể hiện trên đó là rõ nhất. Tuy nhiên trong thực tế, khi đã tìm được một số trục chính có tỉ lệ đóng góp tương đối tốt, có thể dừng lại để quan sát. Cách tìm các trục tọa độ được phân tích xây dựng trên cơ sở toán học (Tô Cẩm Tú và cs., 2003).

2.3. Biểu diễn hình học

Sau khi tìm được các thành phần chính, chiếu đám mây số liệu lên các mặt phẳng chính ta sẽ được hình ảnh "rõ nhất" của dữ liệu. Qua hình ảnh thu được, có thể thấy các điểm nào gần nhau, điểm nào xa nhau, giúp quan sát rõ hơn và đưa ra những nhận xét thích hợp.

2.4. Phân tích hình ảnh dữ liệu thu được

Đây là bước quan trọng đòi hỏi người phân tích phải nắm vững không chỉ các kiến thức toán học mà cả về kiến thức chuyên môn. Với hình ảnh trực quan thu được, người phân tích có thể thấy được sự " gần nhau" của các vectơ biến, vectơ cá thể, sự tương quan giữa 2 biến... Từ đó có thể rút ra những nhận xét, đánh giá chuyên môn hữu ích.

Nếu ma trận số liệu là lớn, việc tính toán rất phức tạp. Ngày nay, nhờ có sự hỗ trợ của máy tính và các phần mềm thống kê, việc tính toán, biểu diễn trở nên đơn giản hơn.

3. PHÂN TÍCH CHÙM DỰA VÀO KHOẢNG CÁCH

Xuất phát từ việc coi mỗi phần của tập hợp là một tập con của nó, tìm cách ghép các tập con này thành một số lớp theo các mức khác nhau (Hadle et al., 2003). Hình ảnh thu được sau khi ghép sẽ cho cái nhìn trực quan về mối liên hệ giữa các dữ liệu thu thập được. Có thể hình dung như sau: Coi các điểm như những chiếc lá,

các lá "gần" nhau sẽ ghép lại thành nhánh, các nhánh "gần" nhau sẽ ghép lại thành cành, các cành "gần" nhau sẽ ghép lại thành cây.

Có nhiều phương pháp xác định số đo sự "gần gũi" giữa các lớp. Mỗi cách xác định số đo tương ứng với một cách lập cây phân loại dựa trên số đo đó. Với hai số đo khác nhau, hai cây phân loại lập được có thể sẽ khác nhau, do đó hình ảnh thu được tương ứng cũng khác nhau. Vì vậy việc chọn số đo thích hợp đối với mỗi ngành khoa học là hết sức quan trọng.

Với sự hỗ trợ của máy tính, việc phân lớp và ghép lớp trở nên đơn giản (Nguyễn Đình Hiền, 2008). Chỉ cần chọn khoảng cách thích hợp và số lớp cần phân chia sẽ thu được một hình ảnh trực quan về những thông tin chứa đựng trong các số liệu thu được.

4. ÁP DỤNG PHÂN TÍCH THÀNH PHẦN CHÍNH VÀ PHÂN TÍCH CHÙM TRONG PHÂN TÍCH SỐ LIỆU

Trong phần này, chúng ta sẽ vận dụng phương pháp phân tích thành phần chính và phân tích chùm để phân tích số liệu và chỉ ra mối quan hệ giữa các chỉ tiêu, các giống cây trồng (Phạm Văn Vân và cs., 2009). Các số liệu ở đây được xử lý dựa vào phần mềm Minitab Nguyễn Đình Hiền (2008).

4.1. Thực hiện phép phân tích thành phần chính và phân lớp trên các biến và các cá thể

Principal Component Analysis: Nang suat, GTSX, CPTG, Lao dong, CPLĐ, GTGT, TNHH

Eigenanalysis of the Correlation Matrix

(bảng trang sau)

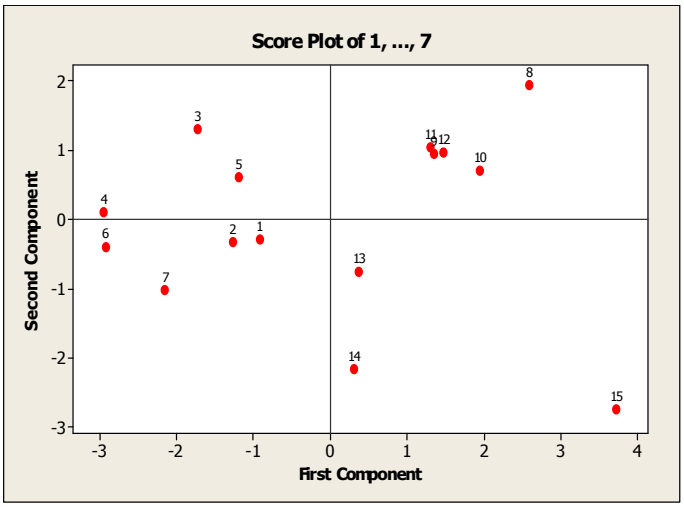
Ba thành phần chính ban đầu chiếm tỉ lệ trên 96% đóng góp, do đó chỉ cần quan sát các cá thể trên hệ trục ba chiều gồm ba trục đầu. Biểu diễn các biến và các cá thể trên mặt phẳng chính mà hai trục là thành phần chính thứ nhất và thành phần chính thứ hai sẽ được biểu đồ trang sau:

Bảng 1. Hiệu quả kinh tế của một số cây trồng chính vùng 1

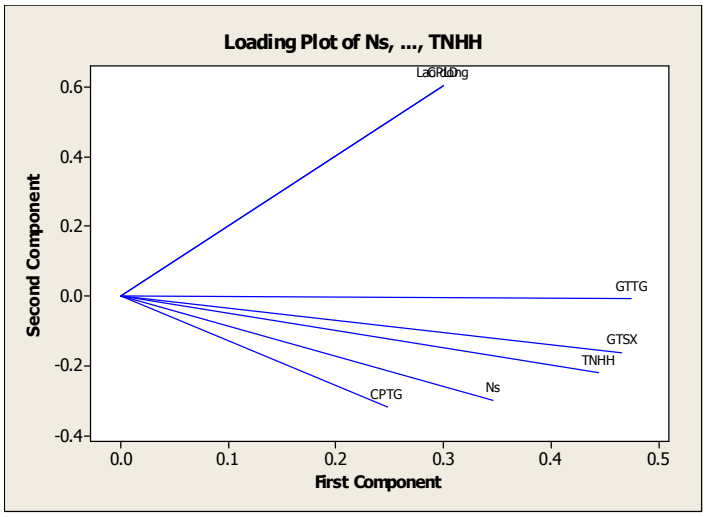
Loại cây trồng	Năng suất (tạ)	GTSX (1.000 đ)	CPTG (1.000 đ)	Lao động (Công)	CPLĐ (1.000 đ)	GTGT (1.000 đ)	TNHH (1.000 đ)
Lúa xuân	61,09	24,436	4,736	310	9,30	20,060	10,760
Lúa mùa	56,98	22,792	4,565	300	9,00	18,227	9,227
Ngô	52,00	17,160	4,712	390	11,70	12,448	0,748
Đậu tương	17,00	12,750	3,290	290	8,70	9,460	0,760
Lạc	22,50	21,735	3,115	350	10,50	18,260	7,760
Khoai lang	80,00	12,800	2,050	260	7,80	10,750	2,950
Khoai tây	66,90	20,070	8,765	260	7,80	11,305	3,505
Cà chua	134,00	40,200	9,498	525	15,75	30,702	14,952
Su hào	198,00	33,660	6,111	435	13,05	27,549	14,449
Bắp cải	218,00	33,880	7,135	435	13,05	31,745	18,695
Dưa chuột	223,00	33,450	8,242	450	13,50	25,208	11,708
Bí đỏ	234,00	33,100	8,766	450	13,50	26,334	12,834
Hành tỏi	105,00	32,550	5,745	310	9,30	26,805	17,505
Cây ăn quả	26,00	39,000	19,061	270	8,10	19,939	11,839
Mía	862,00	47,410	10,273	310	9,30	36,687	27,387

Chú thích: GTSX: Giá trị sản xuất, CPTG: Chi phí trung gian, CPLĐ: Chi phí lao động, GTGT: Giá trị gia tăng, TNHH: Thu nhập hỗn hợp.

Eigenvalue	4.2007	1.6744	0.8569	0.2580	0.0100	0.0000	0.0000
Proportion	0.600	0.239	0.122	0.037	0.001	0.000	0.000
Cumulative	0.600	0.839	0.962	0.999	1.000	1.000	1.000
Variable	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*	*PC7*
Nang suat	0.346	-0.301	0.488	-0.742	-0.030	-0.000	0.000
GTSX	0.466	-0.162	-0.195	0.121	0.839	0.002	-0.000
CPTG	0.247	-0.320	-0.802	-0.268	-0.347	-0.001	0.000
Lao dong	0.299	0.605	-0.078	-0.155	-0.045	0.112	-0.707
CPLĐ	0.299	0.605	-0.078	-0.155	-0.045	0.112	0.707
GTGT	0.475	-0.010	0.157	0.340	-0.276	-0.747	0.000
TNHH	0.444	-0.221	0.208	0.446	-0.307	0.646	-0.000



Biểu đồ 1. Biểu diễn các cá thể trên mặt phẳng chính



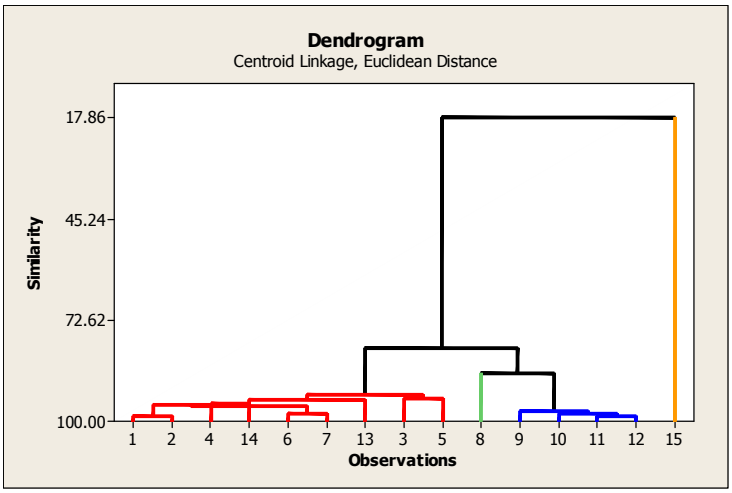
Biểu đồ 2. Biểu diễn các biến trên mặt phẳng chính

4.2. Thực hiện phân lớp trên không gian các biến và cá thể

4.2.1. Phân lớp trên không gian các cá thể

Euclidean Distance, Centroid Linkage

Step	N. of clusters	Similarity level	Distance level	Clusters Joined		New cluster	N. of obs.in new clusters
1	14	98.6854	11.133	11	12	11	2
2	13	98.6775	11.199	1	2	1	2
3	12	98.0590	16.437	6	7	6	2
4	11	97.8526	18.185	10	11	10	3
5	10	97.1732	23.939	9	10	9	4
6	9	95.2008	40.641	4	14	4	2
7	8	95.5819	37.414	1	4	1	4
8	7	95.8451	35.185	1	6	1	6
9	6	94.1917	49.186	1	13	1	7
10	5	94.0041	50.775	3	5	3	2
11	4	92.9648	59.577	1	3	1	9
12	3	86.9926	110.151	8	9	8	5
13	2	80.2355	167.372	1	8	1	14
14	1	17.8597	695.591	1	15	1	15



Biểu đồ 3. Phân lớp trên không gian các cá thể

4.2.2. Phân lớp trên không gian các biến

Correlation Coefficient Distance, Average Linkage

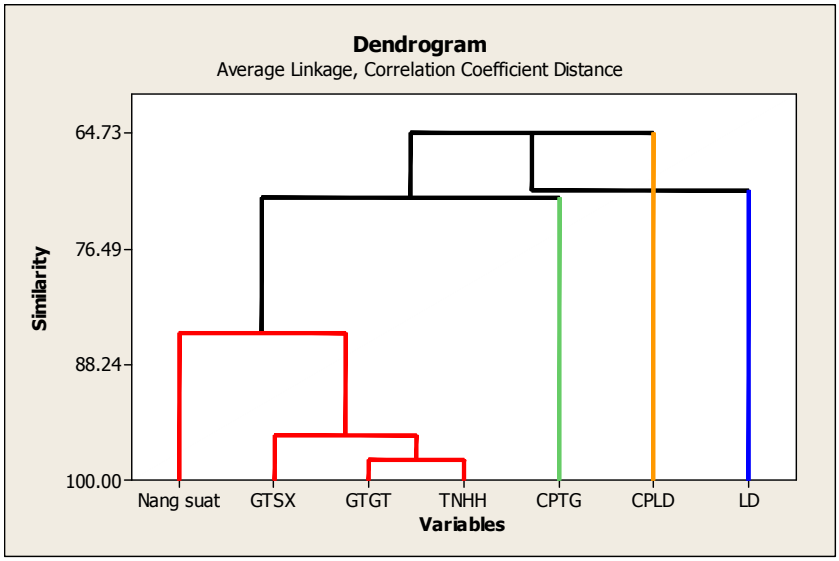
4.3. Phân tích

Từ những thông tin đã được xử lý và các biểu đồ thu được, ta có thể có một số nhận xét đánh giá như sau:

4.3.1. Với các biến

- Trên biểu đồ 4, các biến (GTGT - TNHH) gần nhau nhất (ở cùng một nhóm), gần với biến (GTSX), (Nang suat) và rất xa với biến (Lao dong - CPLĐ). Điều này cũng được thể hiện rõ trong biểu đồ 1, cụm các biến (GTGT - TNHH - GTSX - Nang suat) khá gần nhau và có hệ số tương quan lớn. Có thể hiểu là các chỉ tiêu về

Step	N. of clusters	Similarity level	Distance level	Clusters Joined		New cluster	N. of obs.in new clusters
1	6	97.9030	0.041940	6	7	6	2
2	5	95.5207	0.089587	2	6	2	3
3	4	85.1305	0.297390	1	2	1	4
4	3	71.3245	0.573511	1	3	1	5
5	2	64.7345	0.705309	1	5	1	6
6	1	70.6121	0.587758	1	4	1	7



Biểu đồ 4. Phân lớp trên không gian các biến

năng suất, thu nhập hỗn hợp, giá trị sản suất có mối liên hệ gần gũi với nhau và không chịu nhiều sự tác động của công lao động.

- Do các biến (GTSX - GTGT - CPTG) rất gần biến Nang suat và TNHH nên trong dự báo năng suất và TNHH, có thể chọn một trong các biến nói trên làm biến giải thích, thay vì phải dùng hồi qui bội, tuy hệ số tương quan cao hơn nhưng tính ổn định không cao. Đặc biệt, các nhà chuyên môn nên lưu tâm về sự gần nhau của hai biến GTSX và TNHH.

- Trong biểu đồ 4, CPLĐ và LĐ ở gần nhau và tách thành 2 biến, nhưng khi nhìn vào biểu đồ 2 thấy biểu diễn của 2 biến này là ***trùng nhau,*** quan sát lại bảng số liệu nhận thấy giá trị CPLĐ bằng với giá trị LĐ nhân 3. Trong phân tích thống kê, các nhà chuyên môn ***nên bỏ bớt một trong hai biến này*** vì có đưa thêm vào cũng không có ý nghĩa mà làm cho số liệu thêm phức tạp.

- Hình ảnh từ biểu đồ 2 cho thấy, biến LĐ ở rất xa với cụm biến còn lại. Điều này cũng đặt ra câu hỏi cho các nhà chuyên môn, phải theo dõi vì sao biến LĐ lại ít ảnh hưởng đến các biến khác: Do thống kê không đầy đủ, chưa liên quan nhiều, hay do lao động thủ công không đem lại năng suất cao... Bởi lẽ về nguyên lí nếu càng tiêu tốn lao động thì càng ảnh hưởng đến giá trị sản xuất... Giải thích rõ được điều này sẽ có những kết luận chuyên môn hữu ích.

4.3.2. Với các cá thể

Trước hết nhận xét từ bảng phân tích thành phần chính, đối với trục chính thứ nhất thì hướng về chiều dương giá trị các biến đều

lớn, hướng về phía chiều âm thì giá trị các biến đều nhỏ; với trục chính thứ hai, hướng về chiều dương giá trị biến (LĐ-CPLĐ) lớn, hướng về chiều âm giá trị biến (CPTG - Nang suat) lớn. Tham chiếu điều này vào vị trí các cá thể trên mặt phẳng chính ở biểu đồ 1 ta nhận thấy (ta *chỉ nhận xét một số cá thể có sự khác biệt lớn*):

- Cá thể thứ 15 (mía) khác biệt nhất và ở vị trí có hoành độ dương lớn nhất so với các cá thể khác, do đó có thể kết luận các giá trị của các biến của cây mía lớn hơn các giá trị tương ứng của các cây trồng còn lại. Cây mía đồng thời cũng có tung độ âm lớn nhất lên biến Nang suat và GTSX đặc biệt lớn hơn so với Lao dong, ở đây có thể hiểu là cùng một lao động nhưng tạo ra năng suất và giá trị sản xuất hơn hẳn. Tương tự cá thể 14 tuy có giá trị hoành độ dương nhỏ, nhưng tung độ âm lớn nên giá trị các biến Nang suat, GTSX cũng rất lớn, chứng tỏ hiệu quả năng suất và giá trị sản xuất là cao.

- Cá thể thứ 8 (cà chua) có hoành độ dương lớn nên giá trị các biến nhìn chung là lớn, đặc biệt tung độ dương là lớn nhất nên cà chua là cây có biến LĐ lớn nhất. Điều này có thể hiểu, cây cà chua có Nang suat, TNHH, GTTG…lớn nhưng cần nhiều nhất công lao động

- Đối chiếu vị trí của cá thể 8 và 15 sang cây phân loại ở biểu đồ 3, đây là 2 cá thể ở xa nhất so với các cá thể khác, chứng tỏ kết quả trên cây phân loại và phân tích thành phần chính là khớp nhau và giúp cho hình ảnh thu được là hợp lý và đáng tin cậy

- Nhóm các cá thể 9, 10, 11, 12 (Su hào, bắp cải, dưa chuột, bí đỏ) ở rất gần nhau, gần như cùng vị trí, chứng tỏ các biến này có hoạt động sản suất và hiệu quả kinh tế là như nhau. Điều này được kiểm chứng trong biểu đồ 3 khi các cá thể này ở cùng nhóm với nhau

- Hai cá thể 4 (đậu tương) và 6 (khoai lang) ở gần nhau và ở về vị trí có giá trị hoành độ và tung độ đều nhỏ, chứng tỏ hai loại cây trong này có giá trị các biến là như nhau và đều rất nhỏ so với các biến còn lại. Tức là đối với đậu tương và khoai lang thì lao động, chi phí trung gian là ít, nhưng năng suất, hiệu quả kinh tế cũng thấp như nhau. Kiểm tra lại trong cây phân loại thì thấy hai cá thể 4 và 6 cũng cùng nhóm với nhau ở mức độ thứ hai.

Sau khi nhận xét trên hình ảnh dữ liệu thu được, có thể quay lại kiểm chứng trên bảng số liệu và thấy rất phù hợp với số liệu thực. Rõ ràng, nếu chỉ nhìn vào bảng số liệu không thể đưa ra được những quan sát như vậy.

5. KẾT LUẬN

Phương pháp phân tích thành phần chính và phân tích chùm trong xử lí số liệu giúp cho người nghiên cứu có được hình ảnh gần đúng tốt nhất của bộ dữ liệu thu được. Từ đó có thể đưa ra được những nhận xét rất quan trọng cho công tác nghiên cứu của mình mà nếu chỉ đơn giản là quan sát bảng số liệu thì không thể nhận ra được. Sau đó, các nhà chuyên môn có thể sử dụng các phương pháp phân tích khác của thống kê nhiều chiều: phân tích nhân tố, phân tích phân biệt… để khai thác và sử dụng tối đa những thông tin từ bộ dữ liệu thu được.

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn chân thành và sâu sắc tới thầy Lê Đức Vĩnh, thầy Nguyễn Đình Hiền, cán bộ giảng dạy Khoa Công nghệ thông tin đã động viên tinh thần và giúp đỡ tôi rất nhiều về chuyên môn để tôi hoàn thành bài báo này.

TÀI LIỆU THAM KHẢO

Phạm Văn Vân, Nguyễn Thanh Trà (2009). Đánh giá sử dụng hiệu quả đất nông nghiệp ở huyện Chương Mỹ - Hà Nội. Tạp chí Khoa học và Phát triển, 8(5): 850 - 855.

Hadle W., Simar L. (2003). Applied multivariate Statistical Analysic, 2nd Springer, p: 271-285

Tô Cẩm Tú, Nguyễn Huy Hoàng (2003). Phân tích số liệu nhiều chiều. Nhà xuất bản Khoa học và Kỹ thuật.

Nguyễn Đình Hiền (2008). Thống kê nhiều chiều. Bài giảng phân tích số liệu và bố trí thí nghiệm, Khoa Công nghệ thông tin, Đại học Nông nghiệp Hà Nội.