Mục lục

Lời nói đầu **2**

1 Một số kiến thức thống kê liên quan **4**
1.1 Các khái niệm cơ bản 4
1.1.1 Vectơ ngẫu nhiên 4
1.1.2 Khoảng cách giữa hai vectơ 6
1.1.3 Các loại khoảng cách thường dùng 6
1.2 Phân tích chùm 7
1.2.1 Phân tích chùm là gì? 7
1.2.2 Các bước của phân tích chùm 8
1.2.3 Kiểm tra độ phù hợp của sự phân nhóm. 10
1.3 Phân tích thành phần chính 11
1.3.1 Cấu trúc của các thành phần chính 12
1.3.2 Các thành phần chính của các biến đã chuẩn hóa 16
1.3.3 Phân tích các thành phần chính dựa trên một mẫu 18
1.3.4 Các kết luận thống kê dựa trên mẫu lớn 19

2 Ứng dụng trong phân tích thổ nhưỡng đất trồng trọt của huyện Thanh Ba - Phú Thọ **21**
2.1 Phần mềm trợ giúp việc tính toán 21
2.1.1 Phần mềm SPSS 21
2.2 Số liệu thổ nhưỡng đất 21
2.2.1 Thổ nhưỡng đất 21
2.2.2 Sơ lược về điều tra đất 22
2.2.3 Một số vấn đề về phẫu diện đất tại Thanh Ba - Phú Thọ . . 22
2.3 Kết quả áp dụng phương pháp phân tích chùm 23
2.4 Kết quả áp dụng phương pháp phân tích thành phần chính 25

Tài liệu tham khảo **30**

LỜI NÓI ĐẦU

Phân tích chùm (Cluster Analysis - CA) là một phương pháp thống kê nhằm phân loại các đối tượng (các biến) sao cho mỗi đối tượng (biến) là rất giống so với các đối tượng (biến) khác trong cùng một nhóm dựa vào một vài tiêu chí đã được xác định trước.
Phân tích thành phần chính (Principal Component Analysis - PCA) cũng là một phương pháp thống kê nhằm rút gọn số liệu, biểu diễn và giải thích tập các số liệu dựa trên việc biến đổi phân tích cấu trúc của một ma trận hiệp phương sai của vectơ ngẫu nhiên thông qua việc phân tích các tổ hợp tuyến tính của các thành phần của nó.
Trong khuôn khổ thời gian cho phép của luận văn Thạc sĩ, mục tiêu chính của luận văn là tìm hiểu, hệ thống lại các kiến thức cơ bản có liên quan đến Phân tích chùm, Phân tích thành phần chính dưới góc độ cơ sở toán học và ứng dụng từ đó phân tích trên một số liệu cụ thể. Luận văn được chia làm hai chương:
Chương một đề cập đến một số kiến thức thống kê liên quan. Các khái niệm cơ bản của lý thuyết xác suất thống kê liên quan đến Phân tích chùm và Phân tích thành phần chính như vectơ ngẫu nhiên, khoảng cách giữa hai vectơ. Sau đó là trình bày chi tiết về Phân tích chùm và Phân tích thành phần chính, là cơ sở toán học cho ứng dụng của luận văn.
Chương hai đầu tiên là giới thiệu sơ lược về phần mềm trợ giúp việc tính toán, về thổ nhưỡng đất. Từ đó, đưa ra các kết luận cho số liệu thổ nhưỡng đất trồng trọt của huyện Thanh Ba - Phú Thọ. Với kiến thức chuyên ngành chưa sâu sắc nên luận văn chỉ mới đưa ra được một số kết quả ban đầu. Tuy nhiên, các kết quả có được khá phù hợp với phân tích chuyên ngành và thực tế.

Lời cảm ơn

Trước tiên tôi xin bày tỏ lòng biết ơn sâu sắc tới thầy giáo hướng dẫn - Phó giáo sư, Tiến sĩ Hồ Đăng Phúc, người thầy đã động viên, giúp đỡ và hướng dẫn tôi tận tình trong quá trình hoàn thành luận văn.
Tôi cũng xin gửi lời cảm ơn chân thành tới các thầy cô giáo trong tổ Xác suất - Thống kê đã giúp đỡ tôi rất nhiều trong quá trình học tập cũng như làm luận văn.

Đặc biệt, tôi xin gửi lời cảm ơn đến thầy giáo Lê Đức Vĩnh - Nguyên trưởng bộ môn Toán Trường Đại học Nông Nghiệp Hà Nội đã nhiệt tình giúp đỡ, cung cấp dữ liệu chính xác và một số kiến thức giúp tôi hoàn thành luận văn này.
Cuối cùng là lời cảm ơn chân thành tới gia đình, bạn bè những người đã động viên, giúp đỡ tôi trong quá trình thực hiện luận văn.

Hà Nội, tháng 02 năm 2012

Chương 1

Một số kiến thức thống kê liên quan

1.1 Các khái niệm cơ bản

1.1.1 Vectơ ngẫu nhiên

Vectơ ngẫu nhiên n chiều là một ánh xạ từ không gian mẫu Ω vào $\mathbb{R}^n$. Hay nói cách khác, vectơ ngẫu nhiên $X = (X_1, \ldots, X_n)$ là một vectơ mà mỗi thành phần $X_1, \ldots, X_n$ của nó là một biến ngẫu nhiên.

Ma trận ngẫu nhiên

Nếu $X = (X_{ij})$ là ma trận cấp $n \times p$ mà các thành phần X_{ij} của nó là các biến ngẫu nhiên sẽ được gọi là *ma trận ngẫu nhiên*.

Vectơ trung bình và ma trận phương sai

Cho $X = (X_1, \ldots, X_n)^T$ là một ma trận ngẫu nhiên $n \times 1$. Vectơ $EX = (EX_1, \ldots, EX_n)^T = (\mu_1, \ldots, \mu_n)^T$ được gọi là *vectơ giá trị trung bình*. Đại lượng $\sigma_{ii} = E(X_i - \mu_i)^2, i = 1, \ldots, n$ được gọi là *phương sai* của X_i; $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_i)$ với $\mu_i = E(X_i), \mu_j = E(X_j)$ được gọi là *hiệp phương sai* của hai biến X_i và X_j. *Ma trận hiệp phương sai:* Kí hiệu

$$\text{cov}(X - \mu)(X - \mu)^T = [E(X_i - \mu_i)(X_j - \mu_j)]$$

và gọi đó là *ma trận hiệp phương sai* của vectơ X. Đặt $\sum = \text{cov(X)} = (\sigma_{ij})$ khi đó $\sum$ là ma trận đối xứng xác định không âm cấp n.

Hệ số tương quan và ma trận tương quan: Đại lượng $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$; $\sigma_{ii} = 1$ được gọi là hệ số tương quan của X_i và X_j, còn ma trận

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{n1} & \rho_{n2} & \ldots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ . & 1 & \ldots & \rho_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ . & . & \ldots & 1 \end{bmatrix}$$

được gọi là *ma trận tương quan* của vectơ X.

Vectơ trung bình mẫu và ma trận hiệp phương sai mẫu

Xét véc tơ ngẫu nhiên $X^T = (X_1, X_2, \ldots, X_p)$. Ta thực hiện n quan sát độc lập về X^T. Giả sử quan sát lần thứ nhất ta thu được $X_1 = (x_{11}, x_{12}, \ldots, x_{1p})$, quan sát lần thứ hai ta thu được $X_2 = (x_{21}, x_{22}, \ldots, x_{2p}), \ldots,$ và quan sát thứ n ta thu được $X_n = (x_{n1}, x_{n2}, \ldots, x_{np})$. Kí hiệu

$$X = \begin{bmatrix} X_1^T \\ X_2^T \\ \ldots \\ X_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

là ma trận được tạo bởi các quan sát. Đặt

$$\overline{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}, \ldots, \overline{x}_p = \frac{1}{n}\sum_{i=1}^{n} x_{ip}.$$

Vectơ $\overline{x}_T = (\overline{x}_1, \ldots, \overline{x}_p)$ được gọi là *vectơ trung bình mẫu*. Ma trận

$$S = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1p} \\ s_{21} & s_{22} & \ldots & s_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}$$

được gọi là *ma trận hiệp phương sai mẫu*, trong đó

$$s_{ij} = \frac{1}{n}\sum_{i=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ki}x_{kj} - \overline{x}_i\overline{x}_j$$

được gọi là *hiệp phương sai mẫu*.

Ma trận tương quan mẫu

Ma trận $R = (r_{ij})$ với $r_{ij} = \dfrac{s_{ij}}{(s_{ii}s_{jj})^{1/2}}$ được gọi là *ma trận hệ số tương quan mẫu.*

Như vậy

$$R = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \dfrac{s_{12}}{(s_{11}s_{22})^{1/2}} & \ldots & \dfrac{s_{1p}}{(s_{11}s_{pp})^{1/2}} \\ . & 1 & \ldots & \dfrac{s_{2p}}{(s_{22}s_{pp})^{1/2}} \\ \ldots & \ldots & \ldots & \ldots \\ . & . & \ldots & 1 \end{bmatrix}$$

1.1.2 Khoảng cách giữa hai vectơ

Khoảng cách giữa hai vectơ x và y của không gian vectơ V với tích vô hướng là số $d(x,y) = ||x-y||$. Từ định nghĩa khoảng cách ta có ngay các tính chất sau:

1. $d(x,y) \geq 0 \ \forall \ x,y \in V;\ d(x,y) = 0 \Leftrightarrow x = y.$
2. $d(x,y) = d(y,x) \ \forall \ x,y \in V.$
3. $d(x,z) \leq d(x,y) + d(y,z) \ \forall \ x,y,z \in V.$

1.1.3 Các loại khoảng cách thường dùng

Xét hai véc tơ $x = (x_1,\ldots,x_n)^T$ và $y = (y_1,\ldots,y_n)^T$. Sau đây là các khoảng cách thường dùng để đo sự "gần nhau" giữa hai đối tượng.
Khoảng cách Euclid

$$d_1^2(x,y) = \sum_{i=1}^n (x_i - y_i)^2 = (x-y)^T(x-y).$$

Khoảng cách tổng

$$d_2(x,y) = \sum_{i=1}^n |x_i - y_i|.$$

Khoảng cách max

$$d_3(x,y) = \max_i |x_i - y_i|.$$

Khoảng cách Minkowski

$$d_4(x,y) = (\sum_{i=1}^n |x_i - y_i|^m)^{\frac{1}{m}}, \quad m = 1,2,3,\ldots$$

Khoảng cách thống kê

$$d_5^2(x,y) = (x-y)^T A(x-y)$$

trong đó A là ma trận xác định dương.
Khoảng cách Mahalanobis

$$d_6(x,y) = \sqrt{(x-y)^T \sum^{-1}(x-y)}$$

Khoảng cách giữa các tập con rời nhau
Cho tập $A = \{a_1,\ldots,a_n\}$ với $a_i = (x_{i1},\ldots,x_{in})$ khi đó $C(A) = C_1, C_2, \ldots, C_m$ được gọi là phân hoạch bậc m của tập hợp A nếu thỏa mãn ba điều kiện sau:

1. $C_i \subset A, C_i \neq \varnothing \ \ \forall i = \overline{1,m},$
2. $C_i \cap C_j = \varnothing \ \ \forall i \neq j,$

3. $\bigcup_{i=1}^{m} C_i = A.$

Mỗi $C_i \in C(A)$ còn được gọi là một lớp của phân hoạch $C(A)$. Số phần tử n_i của lớp C_i được gọi là lực lượng của lớp C_i.
Gọi $\overline{c}_i$ là trọng tâm của lớp C_i, $\overline{c}_j$ là trọng tâm của lớp C_j.
Ta có các khoảng cách xác định trong $C(A)$ như sau:

1. $D_1(C_i, C_j) = \min d(a, b)$ với $a \in C_i, b \in C_j$.
2. $D_2(C_i, C_j) = \max d(a, b)$ với $a \in C_i, b \in C_j$.
3. $D_3(C_i, C_j) = d(\overline{c}_i, \overline{c}_j)$.
4. $D_4(C_i, C_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} d(\overline{c}_i, \overline{c}_j)$.
5. $D_5(C_i, C_j) = \sqrt{(\overline{c}_j - \overline{c}_i)^T \sum^{-1} (\overline{c}_j - \overline{c}_i)}$.

Chú ý:

1. Các D_i nói chung không phải là một metric mà chỉ là một siêu metric.
2. Trong định nghĩa khoảng cách giữa các tập con rời nhau thì d có thể là $d_1, d_2, d_3, d_4, d_5, d_6$.

1.2 Phân tích chùm

1.2.1 Phân tích chùm là gì?

Phân tích chùm là tên của những kỹ thuật nhiều biến mà mục đích chính của chúng là phân loại các thực thể tương tự từ những đặc trưng của chúng. Với một vài tiêu chí lựa chọn đã được xác định trước chúng ta xác định và phân loại các đối tượng (các biến) sao cho mỗi đối tượng (biến) là rất giống so với các đối tượng (biến) khác trong cùng một nhóm. Việc phân nhóm như vậy sẽ chỉ ra có tính thuần nhất cao trong mỗi nhóm, tính khác biệt cao giữa các nhóm. Như vậy, nếu phân loại là thành công, các đối tượng trong cùng một nhóm sẽ gần nhau hơn nếu được biểu diễn một cách hình học, trong khi các đối tượng trong các nhóm khác nhau sẽ xa nhau hơn.

1.2.2 Các bước của phân tích chùm

Trong thực hành, phân tích chùm có thể được chia làm ba giai đoạn chính: tiếp cận vấn đề, định danh các nhóm, và chứng minh tính đúng đắn và đưa ra thông tin hữu ích.

Bước 1: Tiếp cận vấn đề

Trong suốt bước này, bốn câu hỏi chính cần được xem xét kỹ lưỡng: Các biến được sử dụng trong tính toán sự giống nhau giữa các nhóm là gì? Sự giống nhau bên trong nhóm nên được đo như thế nào? Trong các nhóm, thuật toán nào nên được sử dụng để hoán đổi các đối tượng tương tự? Nên tạo ra bao nhiêu nhóm? Nhiều cách tiếp cận có thể được sử dụng để trả lời những câu hỏi này, nhưng không cách nào là tuyệt đối để đưa ra được một câu trả lời xác định cho mọi vấn đề. Hơn nữa, những cách tiếp cận trên có thể cho ra những câu trả lời khác nhau cho cùng một tập dữ liệu. Như vậy phân tích chùm, cùng với phân tích nhân tố giống nghệ thuật nhiều hơn là giống khoa học. Bởi lý do này, chúng ta chỉ thảo luận những vấn đề mang tính tổng quát nhất mà không tập trung vào những hạn chế lý thuyết cũng như thực hành của chúng.

Lựa chọn biến trong phân tích chùm phải được hoàn thành với sự xem xét cẩn thận cả yếu tố lý thuyết lẫn thực hành. Giúp đưa ra một cách phân nhóm phù hợp nhất đối với các đối tượng thông qua tất cả các biến. Nhà nghiên cứu phải nhận ra sự quan trọng của việc lựa chọn chỉ những biến mà thể hiện đặc trưng các đối tượng được phân nhóm, và gắn kết sự lựa chọn đó với các mục tiêu của phân tích chùm. Kỹ thuật phân tích chùm không có nghĩa là chỉ ra sự khác nhau của các biến có liên quan với các biến không liên quan. Nó chỉ phát triển từ các nhóm phù hợp nhất của các đối tượng thông qua tất cả các biến.

Thuật toán phân nhóm

Câu hỏi thứ hai cần được trả lời trong giai đoạn tiếp cận vấn đề này là phương thức nào nên được sử dụng để hoán đổi các đối tượng tương tự trong các nhóm? Nghĩa là thuật toán nhóm nào hay bộ các quy tắc nào là chính xác nhất? Đây là một vấn đề không đơn giản vì đã có hàng trăm chương trình máy tính đang sử dụng các thuật toán khác nhau và nhiều chương trình đang được phát triển.

Phương pháp phân nhóm có thứ bậc

Phương pháp phân nhóm có thứ bậc liên quan đến sự xây dựng một cấu trúc phân bậc hay hình cây. Có hai kiểu cơ bản của phương pháp phân nhóm có thứ bậc là cộng gộp và chia tách. Trong phương pháp cộng gộp, mỗi đối tượng khởi đầu với nhóm của chính nó. Trong các bước tiếp theo, hai nhóm (cá thể) gần nhau nhất được kết hợp vào trong một nhóm mới như vậy giảm số nhóm sau mỗi bước xuống một đơn vị. Trong một số trường hợp, một cá thể thứ ba tham gia với hai cá thể đầu tiên trong một nhóm. Trong một số trường hợp khác, nhóm khác của hai cá thể tham gia cùng nhau để tạo một nhóm mới. Cuối cùng là, tất cả các cá thể được ghép nhóm vào trong một nhóm lớn hơn.

Năm phương pháp cộng gộp phổ biến được sử dụng để phát triển nhóm là:

+) Phương pháp liên kết đơn được xây dựng dựa trên khoảng cách nhỏ nhất.

+) Phương pháp liên kết hoàn toàn là gần tương tự như liên kết đơn ngoại trừ rằng tiêu chí nhóm được xây dựng dựa trên khoảng cách cực đại.

+) Liên kết trung bình trong một nhóm khởi đầu giống liên kết đơn và liên kết hoàn thành nhưng tiêu chí xác định nhóm là khoảng cách trung bình trong tất cả các khoảng cách từ một các thể của nhóm này đến một cá thể của nhóm kia.

+) Trong phương pháp của Ward, khoảng cách của hai nhóm là tổng của các bình phương giữa hai nhóm được lấy tổng trên tất cả các biến.

+) Trong phương pháp điểm trung tâm khoảng cách giữa hai nhóm là khoảng cách (thường là khoảng cách Euclide) giữa các điểm trung tâm của chúng.

Bao nhiêu nhóm nên được kiến tạo?

Một vấn đế chính với tất cả kỹ thuật ghép nhóm là bao nhiêu nhóm nên được kiên tạo. Có nhiều tiêu chí và hướng dẫn cho sự tiếp cận vấn đề này. Tuy nhiên, không tiêu chuẩn, phương thức là chung cho tất cả các bài toán. Khoảng cách giữa các nhóm tại các bước kế tiếp có thể cung cấp hướng dẫn hữu ích để lựa chọn thời điểm dừng lại khi mà khoảng cách này đạt tới một giá trị cho trước hoặc khi khoảng cách kế tiếp giữa các bước tạo ra một bước nhảy vọt. Cũng vậy, những hiểu biết lý thuyết có thể gợi ý để lựa chọn số lượng các nhóm.

Bước 2: Định danh các nhóm

Bước này liên quan đến kiểm tra những khẳng định đã được sử dụng để phát triển các nhóm với mục đích đặt tên hoặc gán một nhãn mà diễn tả một cách chính xác sự tự nhiên của các nhóm.

Khi khởi động quá trình định danh các nhóm, một độ đo được sử dụng một cách thường xuyên là điểm trung tâm của nhóm. Nếu phương thức nhóm được thực hiện trên dữ liệu nguyên bản, điều này sẽ là một sự diễn tả hợp logic. Nếu dữ liệu đã được chuẩn hóa, hoặc nếu phân tích nhóm được thực hiện sử dụng các thành phần phân tích nhân tố, ta sẽ phải quay trở lại tới dữ liệu nguyên bản cho các giá trị gốc và tính các thông tin hữu ích trung bình sử dụng những giá trị này.

Bước 3: Đánh giá và đưa ra thông tin hữu ích

Đánh giá bao gồm các nỗ lực bởi những nhà phân tích để đảm bảo rằng các nhóm là đại diện cho đám đông, tổng quát tới các đối tượng khác và ổn định trong một thời gian. Sự tiếp cận trực tiếp nhất của hướng này là phân tách các mẫu, so sánh lời giải nhóm và đánh giá sự tương ứng của các kết quả. Tuy nhiên cách tiếp cận này thường là khó thực hành bởi vì thời gian, chi phí, hoặc tính sẵn có của đối tượng cho phân tích nhóm nhiều chiều. Trong những ví dụ này một sự tiếp cận chung là phân tách mẫu thành hai nhóm. Mỗi nhóm được phân tích nhóm một cách tách biệt, sau đó các kết quả được đem ra so sánh. Một dạng đã chỉnh sửa là để nhận các tâm nhóm từ một nhóm và sử dụng chúng với các nhóm còn lại để xác định các nhóm cần ghép, sau đó so sánh kết quả giữa hai nhóm trên.

Việc phân nhóm có thứ bậc kết nối một tập gồm N phần tử có các bước như sau:

1. Bắt đầu với N nhóm, mỗi nhóm chứa một phần tử, lập ma trận các khoảng cách cấp N là $D = (d_{ik})$.

2. Tìm một ma trận khoảng cách của các cặp các nhóm gần nhất. Giả sử khoảng cách giữa hai nhóm gần nhất U, V là d_{UV}.

3. Gộp nhóm U với nhóm V, kí hiệu nhóm mới là (UV). Lập các phần tử của ma trận khoảng cách mới bằng cách.

   +) Loại các hàng và cột tương ứng với nhóm U, V.

   +) Thêm vào một hàng và một cột gồm các khoảng cách từ nhóm (UV) đến các nhóm còn lại.

4. Lặp lại bước hai và bước ba $N - 1$ lần. Tất cả các phần tử sẽ tạo thành một nhóm duy nhất sau khi kết thúc quá trình.

5. Ghi lại sự nhận dạng của các nhóm đã được kết hợp và mức độ (khoảng cách hoặc sự tương tự) mà ở đó việc kết hợp các nhóm đã được thực hiện.

1.2.3 Kiểm tra độ phù hợp của sự phân nhóm.

Một trong các tính chất mà ta mong muốn khi tiến hành phân nhóm các đối tượng là thu được các nhóm càng tách biệt nhau càng tốt. Để áp dụng tiêu chuẩn thống kê khi khảo sát sự tách biệt giữa các nhóm ta làm các bước sau:
1. Thực hiện việc so sánh từng cặp nhóm.
Ta xét hai nhóm là N_1 và N_2 đã được tách biệt, mỗi nhóm chứa n_1 và n_2 phần tử tương ứng. Giả sử $(x_{1j}, x_{2j}, \ldots, x_{kj})$, $j = \overline{1, n_1}$ là các biến đặc trưng cho phần tử thứ j của nhóm N_1 và $(y_{1j}, y_{2j}, \ldots, y_{kj})$, $j = \overline{1, n_2}$ là các biến đặc trưng cho phần tử thứ j của nhóm N_2.
Đặt $\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k)$ là tâm của nhóm N_1, $\overline{Y} = (\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k)$ là tâm của nhóm N_2, trong đó

$$\overline{x}_i = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{ij}; \overline{y}_i = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{ij};\ i = \overline{1, k},$$

và đặt

$$S^{(1)} = (S_{ij}^{(1)})_{i,j=\overline{1,k}}; S^{(2)} = (S_{ij}^{(2)})_{i,j=\overline{1,k}}$$

là các ma trận hiệp phương sai mẫu của các dấu hiệu của các phần tử của lớp N_1

và N_2 tương ứng, trong đó

$$S_{ij}^{(1)} = \frac{1}{n_1 - 1}[\sum_{k=1}^{n_1}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)],$$
$$= \frac{1}{n_1 - 1}[\sum_{k=1}^{n_1} x_{ik}x_{jk} - n_1\overline{x}_i\overline{x}_j],$$
$$S_{ij}^{(2)} = \frac{1}{n_2 - 1}[\sum_{k=1}^{n_2}(y_{ik} - \overline{y}_i)(y_{jk} - \overline{y}_j)],$$
$$= \frac{1}{n_2 - 1}[\sum_{k=1}^{n_2} y_{ik}y_{jk} - n_1\overline{y}_i\overline{y}_j].$$

Xét thống kê thử nghiệm (khoảng cách thống kê Hotelling giữa hai tâm $\overline{X}, \overline{Y}$ của hai nhóm)

$$T^2 = (\overline{X} - \overline{Y})^T S^{-1}(\overline{X} - \overline{Y}),$$

trong đó

$$S = \frac{1}{n_1}S^{(1)} + \frac{1}{n_2}S^{(2)}.$$

Người ta đã chứng minh được rằng khi n_1, n_2 đủ lớn thì T^2 có phân phối xấp xỉ χ_k^2. Ký hiệu $\chi_k^2(\alpha)$ là phân vị mức α ($\alpha = 0.05; 0.01$) của phân bố χ_k^2. Khi đó, nếu khoảng cách Hotelling $T^2 > \chi_k^2(\alpha)$ thì hai nhóm được coi là tách biệt nhau một cách có ý nghĩa.

2. Nếu hai nhóm không tách biệt nhau một cách có ý nghĩa thì thông thường ta phải tiến hành tách nhóm lại thành một số ít nhóm hơn hoặc nhập hai nhóm đó thành một nhóm.

1.3 Phân tích thành phần chính

Giả sử chúng ta có các quan sát về p biến ngẫu nhiên. Chúng ta tìm cách đơn giản tình hình, xem khi nào có thể tìm được p biến mới $\xi_1, \xi_2, \ldots, \xi_p$ không tương quan với nhau, được biểu diễn tuyến tính qua các biến cũ và không làm mất thông tin về các biến ban đầu. Phân tích thành phần chính là nhằm mục đích như vậy. Nó dựa vào phân tích cấu trúc của một ma trận hiệp phương sai Σ của vectơ ngẫu nhiên X thông qua việc phân tích các tổ hợp tuyến tính của các thành phần của nó. Mục tiêu cơ bản của phân tích thành phần chính là

1. Rút gọn số liệu.

2. Biểu diễn và giải thích tập các số liệu.

Ta đã biết rằng, mỗi ma trận phương sai sinh ra một dạng toàn phương xác định không âm. Phân tích thành phần chính là đưa dạng toàn phương này về trục chính và sắp xếp các trục theo thứ tự giảm dần của các vectơ riêng.
Việc phân tích thành phần chính nhằm phát hiện ra các mối quan hệ ẩn trong các thành phần và cho phép giải thích các mối quan hệ đó.

1.3.1 Cấu trúc của các thành phần chính

Cho vectơ ngẫu nhiên k chiều $X = (X_1, \ldots, X_k) \in \mathbb{R}^k$ có ma trận hiệp phương sai $\text{cov}(X) = \Sigma$. Ta có thể coi $\mu = E(X) = 0$, nếu $\mu \neq 0$, ta có thể thay X bởi $X - \mu$.
Ta hãy xét k tổ hợp tuyến tính

$$\begin{aligned}
Y_1 &= a_1^T X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1k}X_k,\\
Y_2 &= a_2^T X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2k}X_k,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
Y_k &= a_k^T X = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kk}X_k.
\end{aligned}$$

Ta có

$$\begin{aligned}
D(Y_i) &= a_i^T \Sigma a_i,\\
\text{cov}(Y_i, Y_j) &= a_i^T \Sigma a_j.
\end{aligned}$$

Ta tìm các vectơ $a_1^*, \ldots, a_k^*$ sao cho
a_1^* làm cực đại $D(a_1^T X)$ trên tập hợp $\{a_1 : a_1^T a_1 = 1\}$,
a_2^* làm cực đại $D(a_2^T X)$ trên tập hợp $\{a_2 : a_2^T a_2 = 1; \text{cov}(Y_1, a_2^T X) = 0\}$,
.....
a_k^* làm cực đại $D(a_k^T X)$ trên tập hợp

$$\{a_k : a_k^T a_k = 1; \text{cov}(Y_i, a_k^T X) = 0\}, i = \overline{1, k-1}. \qquad (*)$$

Các điều kiện trên đưa đến việc tìm các vectơ $a_1^*, \ldots, a_k^*$ sao cho

$$\begin{aligned}
&D(Y_1) = D((a_1^*)^T X) = \max\{a_1^T \Sigma a_1 : a_1^T a_1 = 1\},,\\
&D(Y_2) = D((a_2^*)^T X) = \max\{a_2^T \Sigma a_2 : a_2^T a_2 = 1; (a_1^*)^T a_2 = 0\},, \qquad (**)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&D(Y_k) = D((a_k^*)^T X) = \max\{a_k^T \Sigma a_k : a_k^T a_k = 1; (a_i^*)^T a_k = 0\},; \forall i = \overline{1, k-1}.
\end{aligned}$$

Định nghĩa 1.3.1. Các đại lượng $Y_i = (a_i^*)^T X$ thỏa mãn điều kiện $(*)$ hoặc $(**)$ được gọi là thành phần chính thứ i của vectơ X với ma trận hiệp phương sai $\text{cov}(X) = \sum$.

Mệnh đề 1.3.2. Cho vectơ X có $\text{cov}(X) = \sum$. Giả sử $(\lambda_1, e_1), \ldots, (\lambda_k, e_k)$ là k cặp giá trị riêng và vectơ riêng của $\sum$ sao cho $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$. Khi đó thành phần chính thứ i của vectơ X được xác định bởi

$$Y_i = e_i X, i = \overline{1,k},$$

và với việc chọn như vậy ta có

$$(Y_i) = \lambda_i, \text{cov}(Y_i, Y_j) = 0, \quad \forall i, j = \overline{1,k}.$$

Để chứng minh Mệnh đề 1.3.2 trước hết ta có bổ đề sau.

Bổ đề 1.3.3. Giả sử $B > 0$ là ma trận cấp $n \times n$ với các giá trị riêng $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ và $e_1, e_2, \ldots, e_n$ là các vectơ riêng tương ứng sao cho nó tạo thành cơ sở trực chuẩn của $\mathbb{R}^n$. Khi đó,

$$\max_{x \neq 0} \frac{x^T B x}{x^T x} = \lambda_1 (\text{đạt được khi } \ x = e_1),$$
$$\min_{x \neq 0} \frac{x^T B x}{x^T x} = \lambda_n (\text{đạt được khi } \ x = e_n).$$

Hơn nữa,

$$\max_{x \perp e_1, \ldots, e_k} \frac{x^T B x}{x^T x} = \lambda_{k+1} (\text{đạt được khi } \ x = e_{k+1}), k = \overline{1, n-1},$$

trong đó $x \perp y$ nếu $x^T y = (x, y) = 0$.

Chứng minh. Đặt $P = (e_1, e_2, \ldots, e_n)$ là ma trận trực giao có các cột là các vectơ riêng, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n), B^{1/2} = P \Lambda^{1/2} P^T, y = P^T x \ (x \neq 0 \Leftrightarrow y \neq 0)$

$$\frac{x^T B x}{x^T x} = \frac{x^T B^{1/2} B^{1/2} x}{x^T P P^T x} = \frac{x^T P \Lambda^{1/2} P^T P \Lambda^{1/2} P^T x}{y^T y}$$
$$= \frac{y^T \Lambda y}{y^T y} = \frac{\sum_1^n \lambda_i y_i^2}{\sum_1^n y_i^2} \leq \lambda_1 \frac{\sum_1^n y_i^2}{\sum_1^n y_i^2} = \lambda_1.$$

Với $x = e_1$ thì $y = P^T e_1 = (1, 0, \ldots, 0)^T$, vì $e_j^T e_1 = 0 \ \forall \ j \neq 1$ nên $e_1^T B e_1 = \lambda_1$. Lý luận tương tự ta có hệ thức hai của bổ đề.
Tiếp theo $x = Py = y_1 e_1 + y_2 e_2 + \ldots + y_n e_n$ nếu $x \perp e_i, i = \overline{1,k}$ khi đó

$$0 = e_i^T x = y_1 e_i^T e_1 + y_2 e_i^T e_2 + \ldots + y_n e_i^T e_n = y_i, \ i \leq k.$$

Vì vậy khi $x \perp e_1, e_2, \ldots, e_k$ ta có

$$\frac{x^T B x}{x^T x} = \frac{\sum_{k+1}^{n} \lambda_i y_i^2}{\sum_{k+1}^{n} y_i^2} \leq \lambda_{k+1},$$

và đạt max khi $y_{k+1} = 1, y_{k+2} = \ldots = y_n = 0$ ứng với $x = e_{k+1}$. Đặt $x^0 = \dfrac{x}{|x|}$ thì x^0 thuộc hình cầu đơn vị, vì vậy

$$\frac{x^T B x}{x^T x} = x^{0^T} B x^0.$$

Do đó max hoặc min của biểu thức đó với $x \neq 0$ tương đương với việc tìm max hoặc min theo x^0 trên hình cầu đơn vị. ■

Bây giờ ta đi chứng minh Mệnh đề 1.3.2.

Chứng minh. Theo định nghĩa các thành phần chính Y_i được xác định bởi các điều kiện $(**)$ và theo Bổ đề 1.3.3, nếu chọn $a_i^* = e_i$ thì các điều kiện $(**)$ được thỏa mãn. Chú ý rằng nếu $\lambda_i = \lambda_{i+1}$ thì các thành phần chính tương ứng $Y_i = e_i^T X$ và $Y_{i+1} = e_{i+1}^T X$ được xác định nhưng không duy nhất. ■

Mệnh đề 1.3.4.

$$\sum_{i=1}^{k} D(X_i) = \sum_{i=1}^{k} D(Y_i) = \lambda_1 + \ldots + \lambda_k.$$

Thật vậy, $tr(\sum) = \sum\limits_{i=1}^{k} D(Y_i) = \lambda_1 + \ldots + \lambda_k$ và do đó ta có điều phải chứng minh.

Định nghĩa 1.3.5. Đại lượng $\dfrac{\lambda_i}{\lambda_1 + \ldots + \lambda_k}$ được gọi là tỷ lệ của biến sai tổng cộng của các thành phần của X do thành phần chính thứ i gây ra.

Mệnh đề 1.3.6. Hệ số tương quan và hiệp phương sai giữa thành phần chính Y_i và thành phần X_j của vectơ X là

$$\text{cov}(Y_i, X_j) = \lambda_i e_{ij},$$
$$\rho(X_i, X_j) = \frac{\sqrt{\lambda_i} e_{ij}}{\sqrt{\rho_{jj}}},$$

trong đó $e_i = (e_{i1}, \ldots, e_{ik})$ $\forall i = \overline{1,k}$.

Chứng minh. Ta có

$$Y_i = e_i^T X; X_j = u^{(j)} X,$$

trong đó $u^{(j)}$ là vectơ đơn vị của $\mathbb{R}^k$ có thành phần thứ j bằng 1. Như vậy

$$\begin{aligned}\text{cov}(Y_i, X_j) &= e_i^T \sum u^{(j)} = e_i^T (\lambda_1 e_1 e_1^T + \ldots + \lambda_k e_k e_k^T) u^{(j)} \\ &= \lambda_i e_i^T u^{(j)} = \lambda_i e_{ij}.\end{aligned}$$

Vì $D(Y_i) = \lambda_i$ nên

$$\rho(Y_i, X_j) = \frac{\text{cov}(Y_i, X_j)}{(D(Y_j) D(X_j))^{1/2}} = \frac{\lambda_i e_{ij}}{(\lambda_i \rho_{jj})^{1/2}} = \frac{\sqrt{\lambda_i} e_{ij}}{\sqrt{\rho_{jj}}}.$$ ■

Mệnh đề 1.3.7. Ký hiệu

$$\widetilde{X_j} = \beta_{1j} Y_1 + \beta_{2j} Y_2 + \ldots + \beta_{pj} Y_p, j = \overline{1,k}$$

là các dự báo tuyến tính tốt nhất của thành phần thứ j theo p thành phần chính đầu tiên $Y_1, Y_2, \ldots, Y_p$ với $1 \le p \le k$. Giả sử $\epsilon_j = \widetilde{X_j} - (\beta_{1j} Y_1 + \beta_{2j} Y_2 + \ldots + \beta_{pj} Y_p)$ là các sai số. Khi đó, tỷ lệ của tổng bình phương của các phương sai của các phần dư $\sum_{j=1}^{k} E(\epsilon_j)^2$ với $\sum_{j=1}^{k} D(X_j)$ được cho bởi

$$\frac{\sum_{j=1}^{k} E(\epsilon_j)^2}{\sum_{j=1}^{k} \sigma_{jj}} = \frac{\lambda_{p+1} + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_k},$$

trong đó $\epsilon_j = X_j - \widetilde{X_j}$ và tỷ số trên bằng 0 khi $p = k$.

Chứng minh. Ta có ma trận hiệp phương sai của vectơ vectơ X và $Y^{(p)}$ là

$$\mathrm{cov}([X|Y^{(p)}]) = \begin{bmatrix} \sum_{XX} & \vdots & \sum_{XY^{(p)}} \\ \dots & \vdots & \cdots \\ \sum_{Y^{(p)}X} & \vdots & \sum_{Y^{(p)}Y^{(p)}} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} & \vdots & \lambda_1 e_{11} & \cdots & \lambda_p e_{p1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \cdots & \sigma_{kk} & \vdots & \lambda_1 e_{1k} & \cdots & \lambda_p e_{pk} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \lambda_1 e_{11} & \cdots & \lambda_1 e_{1k} & \vdots & \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_p e_{p1} & \cdots & \lambda_p e_{pk} & \vdots & 0 & \cdots & \lambda_p \end{bmatrix}$$

trong đó $Y^{(p)} = (Y_1, \dots, Y_p)$. Đặt $\epsilon_j = X_j - \widetilde{X}_j$ là vectơ phần dư khi dự báo X_j theo $Y_1, \dots, Y_p$ và giả sử $\lambda_1 \geq \cdots \geq \lambda_p > 0$. Khi đó

$$\mathrm{var}(\epsilon) = \sum_{XX} - \sum_{XY^{(p)}} \sum_{Y^{(p)}Y^{(p)}}^{-1} \sum_{Y^{(p)}X}.$$

Vì vậy

$$\begin{aligned} \sum_{j=1}^{k} &= tr(\mathrm{var}(\epsilon)) = tr(\sum\nolimits_{XX}) - tr(\sum\nolimits_{XY^{(p)}} \sum\nolimits_{Y^{(p)}Y^{(p)}}^{-1} \sum\nolimits_{Y^{(p)}X}) \\ &= \sum_{j=1}^{k} \lambda_j - tr(\sum\nolimits_{Y^{(p)}Y^{(p)}}^{-1} \sum\nolimits_{Y^{(p)}X} \sum\nolimits_{XY^{(p)}}) \\ &= \sum_{j=1}^{k} \lambda_j - tr(\sum\nolimits_{Y^{(p)}Y^{(p)}}^{-1} \mathrm{diag}(\lambda_1^2, \dots, \lambda_p^2)) \\ &= \sum_{j=1}^{k} \lambda_j - \sum_{j=1}^{p} \lambda_j = \sum_{j=p+1}^{k} \lambda_j. \end{aligned}$$

ta có điều phải chứng minh. ■

1.3.2 Các thành phần chính của các biến đã chuẩn hóa

Vì tính toán với ma trận tương quan ρ sẽ ổn định hơn so với việc tính toán đối với ma trận hiệp phương sai $\sum$ nên người ta hay xét các thành phần chính của các biến

đã chuẩn hóa

$$\begin{aligned} Z_1 &= (X_1 - \mu_1)/\sigma_{11}^{1/2}, \\ Z_2 &= (X_2 - \mu_2)/\sigma_{22}^{1/2}, \\ &\cdots\cdots \\ Z_k &= (X_k - \mu_k)/\sigma_{kk}^{1/2}. \end{aligned}$$

Đặt $Z = (Z_1, Z_2, \ldots, Z_k)^T$ ta có

$$Z = V^{-1/2}(X - \mu),$$

trong đó

$$V = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{kk}).$$

Khi đó,

$$EZ = 0, \text{var}(Z) = V^{-1/2} \sum V^{-1/2} = \rho,$$

và các thành phần chính của Z được xác định nhờ các vectơ riêng của ρ, các vectơ riêng này nói chung khá với các vectơ riêng của $\sum$. Mệnh đề sau cho chúng ta công thức các định các thành phần chính của các biến đã chuẩn hóa.

Mệnh đề 1.3.8. Thành phần chính thứ i của các vectơ đã được chuẩn hóa $Z = (Z_1, Z_2, \ldots, Z_k)$ với ma trận tương quan ρ được xác định bởi

$$Y_i^0 = e_i^T Z = e_i^T V^{-1/2}(X - \mu), \ i = \overline{1,k}$$

Hơn nữa,

$$\sum_{i=1}^{k} D(Y_i^0) = \sum_{i=1}^{k} D(Z_i) = k,$$

và

$$\rho(Y_i^0, Z_j) = e_{ij}\sqrt{\lambda_i}, \quad i, j = \overline{1,k},$$

trong đó $(\lambda_1, e_1), \ldots, (\lambda_k, e_k)$ là các vectơ riêng của ρ với $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

1.3.3 Phân tích các thành phần chính dựa trên một mẫu

Bây giờ ta chuyển sang tìm hiểu về tính chất thống kê của phân tích thành phần chính khi sử dụng ma trận hiệp phương sai mẫu. Giả sử ta có n quan sát độc lập $X_1, X_2, \ldots, X_n$ về một vectơ ngẫu nhiên k chiều, X có $EX = \mu$ và $\text{var}(X) = \sum$.
Các số liệu đó sinh ra vectơ trung bình mẫu $\overline{X}$, ma trận hiệp phương sai mẫu S và ma trận tương quan mẫu R. Với $X_m = (x_{m1}, x_{m2}, \ldots, x_{mk})^T \in \mathbb{R}^k$ thì

$$\overline{X} = \frac{1}{n}\sum_{m=1}^{n} X_l, \overline{X} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k)^T,$$

$$S = (s_{ij})_{i,j=\overline{1,k}}, s_{ij} = \frac{1}{n}\sum_{m=1}^{n}(x_{mi} - \overline{x}_i)(x_{mj} - \overline{x}_j),$$

$$R = (r_{ij})_{i,j=\overline{1,n}}, r_{ij} = \frac{s_{ij}}{(s_{ii}s_{jj})^{1/2}}.$$

Ta sẽ xây dựng các tổ hợp tuyến tính của các thành phần của các số đo X_i sao cho các tổ hợp đó không tương quan với nhau và có phương sai mẫu lớn nhất.
Với mỗi $a \in \mathbb{R}^k$, tổ hợp tuyến tính

$$a_1^T X_m = a_{11}x_{m1} + a_{12}x_{m2} + \ldots + a_{1k}x_{mk},\ m = \overline{1,n}$$

là một quan sát của biến ngẫu nhiên $(a^T X)$. Vì vậy dãy $(a_1^T X_m)_{m=1}^n$ là một mẫu của $(a^T X)$ với trung bình mẫu $a^T\overline{X}$ và phương sai mẫu là $a^T S\, a$.
Hơn nữa, hai dãy $(a_1^T X_m)_{m=1}^n$ và $(a_2^T X_m)_{m=1}^n$ có hiệp phương sai mẫu là $a_1^T S\, a_2$.
Các thành phần chính mẫu được định nghĩa như sau:
Gọi $\widehat{Y}_{1m}, \widehat{Y}_{2m}, \ldots, \widehat{Y}_{km}$ là thành phần chính mẫu thứ $1, 2, \ldots, k$ của mẫu $X_1, X_2, \ldots, X_n$ nếu

1. Thành phần chính thứ nhất $(\widehat{Y}_{1m})_{m=1}^n$ với $\widehat{Y}_{1m} = a_1^T X_m$. Vectơ a_1 được chọn sao cho phương sai mẫu $a_1^T S\, a_1$ đạt giá trị lớn nhất với điều kiện $a_1^T a_1 = 1$.

2. Thành phần chính thứ hai $(\widehat{Y}_{2m})_{m=1}^n$ với $\widehat{Y}_{2m} = a_2^T X_m$. Vectơ a_2 được chọn sao cho phương sai mẫu $a_2^T S\, a_2$ đạt giá trị lớn nhất với điều kiện $a_1^T S\, a_2 = 0; a_2^T a_2 = 1. \ldots\ldots\ldots$

3. Thành phần chính thứ k $(\widehat{Y}_{km})_{m=1}^n$ với $\widehat{Y}_{km} = a_k^T X_m$. Vectơ a_k được chọn sao cho phương sai mẫu $a_k^T S\, a_k$ đạt giá trị lớn nhất với điều kiện $a_i^T S\, a_k = 0,\ \forall i = \overline{1, k-1}; a_k^T a_k = 1$.

Khi đó, ta xác định các thành phần chính mẫu thứ i bởi mệnh đề sau:

Mệnh đề 1.3.9. Giả sử $(\widehat{\lambda_1}, \widehat{e_1}), \ldots, (\widehat{\lambda_k}, \widehat{e_k})$ là k cặp giá trị riêng và vectơ riêng của ma trận hiệp phương sai mẫu S. Khi đó, thành phần chính mẫu thứ i

$$\widehat{Y_{im}} = \widehat{e_i}^T X_m = \widehat{e_{i1}} x_{m1} + \widehat{e_{i2}} x_{m2} + \ldots + \widehat{e_{ik}} x_{mk},\ i = \overline{1,k};\ m = \overline{1,n},$$

trong đó $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \ldots \geq \widehat{\lambda_k}$. Tổng quát, nếu $x = (x_1, \ldots, x_k)^T$ là một quan sát của vectơ ngẫu nhiên $X = (X_1, \ldots, X_k)^T$ thì

$$\widehat{y_i} = \widehat{e_i}^T X = \widehat{e_{i1}} x_1 + \widehat{e_{i2}} x_2 + \ldots + \widehat{e_{ik}} x_k,\ i = \overline{1,k}$$

là thành phần chính mẫu thứ i, ứng với quan sát x, dựa trên mẫu $X_1, \ldots, X_n$. Hơn nữa, phương sai mẫu của $\widehat{y_i}$ là $\widehat{\lambda_i}, i = \overline{1,k}$. Hiệp phương sai mẫu của $\widehat{y_i}$ và $\widehat{y_j}$ bằng $0,\ \forall i \neq j$.

Phương sai mẫu tổng cộng là $\sum_{i=1}^{k} s_{ii} = \widehat{\lambda_1} + \widehat{\lambda_2} + \ldots + \widehat{\lambda_k}$, và hệ số tương quan mẫu

$$r_{\widehat{y_i}, x_j} = \frac{\widehat{e_{ij}} \sqrt{\widehat{\lambda_i}}}{\sqrt{s_{jj}}};\ \ i, j = \overline{1,k}.$$

Nếu ta thay mẫu $X_1, \ldots, X_n$ bởi mẫu rút gọn $X_1^0, \ldots, X_n^0$ trong đó $X_m^0 = (x_{m1}^0, \ldots, x_{mk}^0)^T$ với $x_{mi}^0 = (x_{mi} - \overline{x}_i)/s_{ii}^{1/2}, i = \overline{1,k}, m = \overline{1,n}$ thì các ma trận hiệp phương sai mẫu của mẫu $X_1^0, \ldots, X_n^0$ chính là R và các thành phần chính mẫu $\widehat{Y_{im}^0}$ của mẫu $X_1^0, \ldots, X_n^0$ được thực hiện tương tự khi thay S bởi R.

Vấn đề đặt ra là cần giữ lại bao nhiêu thành phần chính và giải thích các vấn đề có liên quan đến thành phần chính. Thông thường thành phần chính ứng với $\widehat{\lambda_i}$ với $\widehat{\lambda_i} \approx 0$ không đóng vai trò gì trong phân tích số liệu và ta có thể loại đi. Trong thực hành, khi khảo sát trên mẫu số liệu cụ thể thì số thành phần chính được chọn còn phụ thuộc vào ý nghĩa thực tế.

Dựa vào cấu trúc của các thành phần chính người ta có thể giải thích vai trò của các biến X_i trong các thành phần chính đó và từ đó có một sự lý giải về vai trò của các nhân tố trong tự nhiên và xã hội.

1.3.4 Các kết luận thống kê dựa trên mẫu lớn

Việc giữ lại các thành phần chính được dựa trên các giá trị riêng λ_i của ma trận hiệp phương sai là lớn hay bé. Tuy nhiên, các λ_i là chưa biết. Hơn nữa, chất lượng của thành phần chính thứ i phụ thuộc vào λ_i và vectơ riêng e_i. Vì vậy, việc tìm khoảng tin cậy của các λ_i, vectơ riêng e_i và do đó thành phần chính Y_i dựa trên một mẫu ngẫu nhiên $X_1, \ldots, X_n$ cần được nghiên cứu. Tuy vậy, việc xác định phân bố

của $\widehat{\lambda_i}$ và $\widehat{e_i}$ rất khó. Tuy nhiên, người ta cũng đã thu được phân bố mẫu của $\widehat{\lambda_i}$ và $\widehat{e_i}$ như sau

Phân bố mẫu của $\widehat{\lambda_i}$ và $\widehat{e_i}$

Giả thiết rằng $X_1, \ldots, X_n$ là n quan sát độc lập của vectơ X có phân bố chuẩn $N_k(\mu, \Sigma)$, trong đó Σ là ma trận xác định dương với các giá trị riêng $\lambda_1 > \lambda_2 > \ldots > \lambda_k > 0$ đều chưa biết. Giả sử $(\widehat{\lambda_i}, \widehat{e_i})$ là cặp giá trị riêng và vectơ riêng của ma trận hiệp phương sai mẫu S sao cho $\widehat{\lambda_1} \geq \widehat{\lambda_2} \geq \ldots \geq \widehat{\lambda_k}$. Khi đó,

1. Đặt Λ là ma trận chéo với các giá trị riêng $\lambda_1, \lambda_2, \ldots, \lambda_k$ ở trên đường chéo chính. Khi đó $\sqrt{n}(\widehat{\lambda} - \lambda)$ có phân bố gần chuẩn $N_k(0, 2\Lambda^2)$, trong đó $\lambda^T = (\lambda_1, \lambda_2, \ldots, \lambda_k), \widehat{\lambda}^T = (\widehat{\lambda_1}, \widehat{\lambda_2}, \ldots, \widehat{\lambda_k})$.

2. Đặt

$$E_i = \lambda_i \sum_{j=1, j\neq i}^{k} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2} e_i e_j^T,$$

   trong đó e_j là vectơ riêng ứng với giá trị riêng λ_j của $\sum$. Khi đó $\sqrt{n}(\widehat{e_i} - e_i)$ có phân bố xấp xỉ chuẩn $N_k(0, E_i)$.

3. Mỗi $\widehat{\lambda_i}$ có phân bố độc lập với phần tử $\widehat{e_i}$. Từ kết luận 1 ta nhận được

$$P\{|\widehat{\lambda_i} - \lambda_i| \leq u(\frac{\alpha}{2})\lambda_i \sqrt{2/n}\} = 1 - \alpha$$

   do đó khoảng tin cậy với mức tin cậy $1 - \alpha$ của λ_i là

$$\frac{\widehat{\lambda_i}}{1 + u(\frac{\alpha}{2})\sqrt{2/n}} < \lambda_i < \frac{\widehat{\lambda_i}}{1 - u(\frac{\alpha}{2})\sqrt{2/n}}.$$

Chương 2

Ứng dụng trong phân tích thổ nhưỡng đất trồng trọt của huyện Thanh Ba - Phú Thọ

2.1 Phần mềm trợ giúp việc tính toán

2.1.1 Phần mềm SPSS

SPSS là tên viết tắt của một phần mềm thống kê nổi tiếng "Statistical Packages for Social Sciences" của công ty SPSS (Mỹ). Phần mềm này được phát triển từ năm 1960, lúc đầu chỉ hoạt động trên máy tính lớn. Khi máy tính cá nhân trở thành phổ biến, công ty SPSS đã thành công trong việc đưa ra các phiên bản SPSS chạy trên hệ điều hành Windows. SPSS trở thành một công cụ phân tích thống kê không thể thiếu được để thực hiện các phương pháp phân tích định lượng. Gần đây, công ty SPSS đã đổi tên phần mềm SPSS thành PASW (Predictive Analytics Software) Statistics nhằm thể hiện ý tưởng kết hợp công cụ thống kê toán học với việc phân tích dự báo.

2.2 Số liệu thổ nhưỡng đất

2.2.1 Thổ nhưỡng đất

Thổ nhưỡng là đất mặt tơi xốp của vỏ lục địa, có độ dầy khác nhau, có thể sản suất ra những sản phẩm của cây trồng. Nguồn gốc của đất là từ các đá mẹ nằm trong thiên nhiên lâu đời bị phá hủy dưới tác dụng của yếu tố lý học, hóa học và sinh học. Tiêu chuẩn cơ bản để phân biệt giữa đá mẹ và đất là độ phì nhiêu, nếu chưa có độ phì nhiêu, thực vật cao cấp chưa sống được thì chưa gọi là thổ nhưỡng. Các yếu tố hình thành đất là đá mẹ, các mẫu chất, sinh vật (động vật, thực vật và vi sinh vật), khí hậu, địa hình, thời gian và con người.

Các loại đá nằm trong thiên nhiên chịu tác dụng lý học, hóa học và sinh học dần dần bị phá hủy thành một sản phẩm gọi là mẫu chất. Trong mẫu chất mới chỉ có các nguyên tố hóa học chứa trong đá mẹ sinh ra nó, còn thiếu một số thành phần quan trọng như chất hữu cơ, đạm nước... vì thế thực vật cao cấp chưa sống được. Trải qua một thời gian dài nhờ tác dụng của sinh vật tích lũy được chất hữu cơ và đạm, thực vật cao cấp sống được, có nghĩa là đã hình thành thổ nhưỡng.
Dù là đất nông nghiệp, đất lâm nghiệp, đất đồng cỏ, thậm chí là đất hoang đều gồm có các thành phần cơ bản cụ thể là thổ nhưỡng gồm chất rắn (chất vô cơ, chất hữu cơ), khe hở giữa các hạt (không khí, nước) và các loài sinh vật.

2.2.2 Sơ lược về điều tra đất

Địa điểm đào phẫu diện phải thật đại diện cho khu vực điều tra. Sau đó khi đào phẫu diện thì đào đến khi nào gặp tầng cứng rắn, đá mẹ hoặc đến độ sâu tối thiểu là 125cm nếu chưa gặp tầng cứng rắn, chiều rộng 70 - 80cm, chiều dài 1.2 - 2.0m. Khi gặp loại đất giống đất ở phẫu diện chính gần đó thì đào phẫu diện phụ sâu 100cm.
Lấy mẫu đất đi phân tích theo trình tự sau: lấy mẫu đất ở đáy phẫu diện, sau đó lấy dần lên các tầng trên, lấy ở tất cả các tầng phát sinh, lấy đều theo độ dày tầng đất, tầng dày chưa đến 50cm lấy một mẫu, tầng dày 50 - 90cm lấy hai mẫu, tầng dày hơn 90cm lấy ba mẫu và mẫu đất phải lấy đủ trọng lượng 1kg. Lấy đất ở các tầng cho vào các ngăn của hộp tiêu bản bằng giấy, gỗ hoặc nhựa. Đất cho vào hộp phải giữ được dạng tự nhiên và đặc trưng cho tất cả các tầng đất. Sau đó mô tả phẫu diện đất.

2.2.3 Một số vấn đề về phẫu diện đất tại Thanh Ba - Phú Thọ

Dựa vào số liệu điều tra ta có thể biết được về phẫu diện đất. Cụ thể là ở số liệu trong luận văn là như sau.
Số liệu được điều tra tại các xã Hanh Cù, Vô Lao, Thanh Xá, Yên Nội, Đông Lĩnh, Chí Tiến, TT Thanh Ba, Thái Ninh, Đồng Xuân, Thanh Vân, Lương Lỗ, Đỗ Sơn, Đông Thành, Năng Yên, Quảng Nạp, Khải Xuân, Hoàng Cương, Thanh Hà, Đỗ Xuyên, Yểu Khê, Sơn Cương, Mạn Lan, Phương Lĩnh, Ninh Dân của huyện Thanh Ba tỉnh Phú Thọ. Trong đó, đã điều tra về địa hình, thành phần cơ giới, màu sắc, chất hữu cơ và tính kiềm của đất.
Cụ thể là số liệu đã được điều tra trên các loại địa hình khác nhau như đồi, gò, núi, bãi, dốc, nông trường và cánh đồng. Thành phần cơ giới gồm có sét, limon và cát. Và các nguyên tố hóa học chính trong đất như Si, Al, Fe, Ca, Mg, S, N, P, K và một số nguyên tố vi lượng khác.
Dung tích trao đổi cation của đất là tổng số cation (kể cả cation kiềm và không kiềm) được giữ ở trạng thái trao đổi trong 100g đất, tính bằng ly đương lượng gam,

ký hiệu bằng chữ CEC.
Dung tích trao đổi cation được xác định bằng cách phân tích trực tiếp hoặc tính theo công thức CEC=S+H. Trong đó S là tổng số cation kiềm, kiềm thổ hấp thụ (chủ yếu là Ca^{2+}, Mg^{2+}, K^+ và Na^+), H là tổng số ion H^+ và Al^{3+} hấp thụ (độ chua thủy phân). Tất cả đều tính bằng đơn vị 1đl/100g đất.
Dung tích trao đổi cation của đất phụ thuộc vào thành phần keo, thành phần cơ giới đất, tỷ lệ SiO_2/R_2O_3 và độ pH. Thành phần keo đất khác nhau thì CEC của đất cũng khác nhau, đất càng nhiều mùn, thành phần cơ giới đất càng nặng, tỉ lệ SiO_2/R_2O_3 càng lớn thì CEC càng lớn, độ pH tăng lên thì CEC cũng tăng lên.
Nói chung, CEC có giá trị càng cao thì đất càng tốt vì chứa nhiều keo. Tuy nhiên, dung tích trao đổi cation chỉ nói lên khả năng trao đổi cation mà chưa nói lên thành phần cation hấp thụ. Thực tế một số đất có CEC lớn nhưng do nhiều H^+ nên đất chua. Vì thế, cần có tỷ lệ CEC lớn nhưng tỷ lệ cation bazơ (cả cation kiềm và kiềm thổ) cũng lớn đất mới tốt. Bởi vậy người ta dùng chỉ tiêu độ no bazơ để đánh giá độ phì nhiêu của đất.
Độ no bazơ của đất là tỷ lệ phần trăm các cation kiềm, kiềm thổ chiếm trong tổng số cation kiềm hấp thụ, ký hiệu là BS, đơn vị % và được tính theo công thức BS(%)= (S x 100)/CEC = (S x 100)/(S+H). BS có giá trị càng lớn thì đất càng bão hòa bazơ. Cụ thể là BS < 50% là đất đói bazơ, BS từ 50% đến 75% là đất có độ bazơ trung bình còn BS > 75% là đất no bazơ.
Như vậy, cả độ pH, CEC và BS đều dùng để đo tính kiềm của đất.
Chất hữu cơ và mùn trong đất
Chất hữu cơ do xác sinh vật phân hủy chiếm dưới 5% trọng lượng hoặc 12% thể tích chất rắn. Dấu hiệu cơ bản làm đất khác đá mẹ là đất có chất hữu cơ và mùn. Số lượng và tính chất của chúng tác động mạnh mẽ đến quá trình hình thành đất, quyết định nhiều tính chất lý, hóa, sinh và độ phì nhiêu của đất. Về mặt số lượng chất hữu cơ, tiêu chí cơ bản nhất để đánh giá là tỷ lệ %OC (cacbon hữu cơ tổng số) hoặc tỷ lệ % mùn hoặc OM (chất hữu cơ tổng số = 1.72 x OC) so với đất khô kiệt.

2.3 Kết quả áp dụng phương pháp phân tích chùm

Mục tiêu chính của các nhà thổ nhưỡng là phân loại các mẫu đất ở các vùng khác nhau xem có thể xếp các mẫu đất đó vào những nhóm khác biệt như thế nào để quy hoạch việc trồng trọt các loại cây trên mẫu đất thích hợp. Vì vậy, ta cố gắng sắp xếp các đối tượng trong mẫu số liệu thành những mẫu đất có các đặt tính sinh hóa, tính chất, địa hình...khác nhau để phục vụ cho canh tác. Nên ta dùng phân tích chùm để chia các đối tượng ta thành các nhóm.
Bằng cách sử dụng khoảng cách Euclid bình phương với ngưỡng cắt thích hợp chia các đối tượng làm 3 nhóm cụ thể là nhóm 1 gồm 76 đối tượng, nhóm 2 gồm 12 đối tượng và nhóm 3 gồm 7 đối tượng (2.1). Từ đó ta thấy rằng, việc phân chia nhóm như thế này là chưa thực sự phù hợp với các nhà phân tích, chưa giúp được định

hướng cho việc khai thác giá trị trồng trọt trên các nhóm đất đó. Có thể là vì khi dùng các biến nguyên thủy thì số đo của các biến chưa thực sự giống nhau như có biến được đo với đơn vị là 1dl/100g, có biến lại đo với đơn vị là mg/100g và một số biến khác lại biểu diễn dưới dạng%. Để khắc phục điều đó ta đi phân tích trên các biến đã chuẩn hóa thì được kết quả (Hinh 2.2).

Cluster Membership

Case	3 Clusters
1:Hanh Cù	1
2:Thanh Vân	1
3:Hanh Cù	1
5:Hanh Cù	1
6:Hanh Cù	1
7:Võ Lao	1
8:Võ Lao	1
10:Võ Lao	1
11:Võ Lao	1
12:Võ Lao	1
13:Thanh Xá	1
14:Thanh Xá	1
15:Thanh Xá	1
16:Thanh Xá	1
17:Thanh Xá	1
19:Yên N?i	1
20:Yên N?i	1
21:Yên N?i	1
22:Yên N?i	1
23:Đồng Linh	1
24:Đồng Linh	1
25:Đồng Linh	1
26:Đồng Linh	1
27:Chí Ti?n	1
28:Chí Ti?n	1
29:Chí Ti?n	1
30:Chí Ti?n	1
31:Chí Ti?n	1
32:TT Thanh Ba	1
33:TT Thanh Ba	1
34:TT Thanh Ba	1
35:TT Thanh Ba	1
36:TT Thanh Ba	1
37:Thái Ninh	1
39:Thái Ninh	1
40:Thái Ninh	1
41:Thái Ninh	1
42:Thái Ninh	1
43:Thái Ninh	1
44:Đ?ng Xuân	1
46:Đ?ng Xuân	1
47:Đ?ng Xuân	1
48:Đ?ng Xuân	2
49:Thanh Vân	2
50:Thanh Vân	2
51:Thanh Vân	2
52:Thanh Vân	2
53:Lương L?	2
54:Lương L?	2
55:Lương L?	1
56:Đ? Sơn	1
57:Đ? Sơn	1
58:Đ? Sơn	1
59:Đồng Thành	1
60:Đồng Thành	1
61:Đồng Thành	1
62:Đồng Thành	1
64:Đồng Thành	1
65:Nang Yên	1
66:Nang Yên	1
67:Nang Yên	1
68:	1
69:	1
70:	1
72:Qu?ng N?p	1
73:Qu?ng N?p	3
74:Qu?ng N?p	1
75:Qu?ng N?p	1
76:Kh?i Xuân	3
77:Kh?i Xuân	2
78:Kh?i Xuân	1
79:Kh?i Xuân	3
81:Kh?i Xuân	2
82:Kh?i Xuân	2
83:	2
84:Hoàng Cương	1
85:Hoàng Cương	3
86:Thanh Hà	2
87:Thanh Hà	1
88:Thanh Hà	3
89:Đ? Xuyên	2
90:Đ? Xuyên	3
91:Đ? Xuyên	2
92:Y?u Khê	1
93:Y?u Khê	1
94:	1
95:	1
96:Sơn Cương	3
97:Sơn Cương	1
98:Sơn Cương	1
99:M?n Lan	1
102:M?n Lan	1

Hình 2.1: Biểu diễn các đối tượng được chia thành 3 nhóm.

Cluster Membership

Case	3 Clusters
1:Hanh Cù	1
2:Thanh Vân	1
3:Hanh Cù	1
5:Hanh Cù	1
6:Hanh Cù	1
7:Võ Lao	1
8:Võ Lao	1
10:Võ Lao	1
11:Võ Lao	1
12:Võ Lao	1
13:Thanh Xá	1
14:Thanh Xá	1
15:Thanh Xá	1
16:Thanh Xá	1
17:Thanh Xá	1
19:Yên N?i	1
20:Yên N?i	1
21:Yên N?i	1
22:Yên N?i	1
23:Đồng Linh	1
24:Đồng Linh	1
25:Đồng Linh	1
26:Đồng Linh	1
27:Chí Ti?n	1
28:Chí Ti?n	1
29:Chí Ti?n	1
30:Chí Ti?n	1
31:Chí Ti?n	1
32:TT Thanh Ba	1
33:TT Thanh Ba	1
34:TT Thanh Ba	1
35:TT Thanh Ba	1
36:TT Thanh Ba	1
37:Thái Ninh	1
39:Thái Ninh	1
40:Thái Ninh	1
41:Thái Ninh	1
42:Thái Ninh	1
43:Thái Ninh	1
44:Đ?ng Xuân	1
46:Đ?ng Xuân	1
47:Đ?ng Xuân	1
48:Đ?ng Xuân	2
49:Thanh Vân	2
50:Thanh Vân	2
51:Thanh Vân	2
52:Thanh Vân	2
53:Lương L?	2
54:Lương L?	2
55:Lương L?	1
56:Đ? Sơn	1
57:Đ? Sơn	1
58:Đ? Sơn	1
59:Đồng Thành	1
60:Đồng Thành	1
61:Đồng Thành	1
62:Đồng Thành	1
64:Đồng Thành	1
65:Nang Yên	1
66:Nang Yên	1
67:Nang Yên	1
68:	1
69:	1
70:	1
72:Qu?ng N?p	1
73:Qu?ng N?p	3
74:Qu?ng N?p	1
75:Qu?ng N?p	1
76:Kh?i Xuân	3
77:Kh?i Xuân	2
78:Kh?i Xuân	1
79:Kh?i Xuân	3
81:Kh?i Xuân	2
82:Kh?i Xuân	2
83:	2
84:Hoàng Cương	1
85:Hoàng Cương	3
86:Thanh Hà	2
87:Thanh Hà	1
88:Thanh Hà	3
89:Đ? Xuyên	2
90:Đ? Xuyên	3
91:Đ? Xuyên	2
92:Y?u Khê	1
93:Y?u Khê	1
94:	1
95:	1
96:Sơn Cương	3
97:Sơn Cương	1
98:Sơn Cương	1
99:M?n Lan	1
102:M?n Lan	1

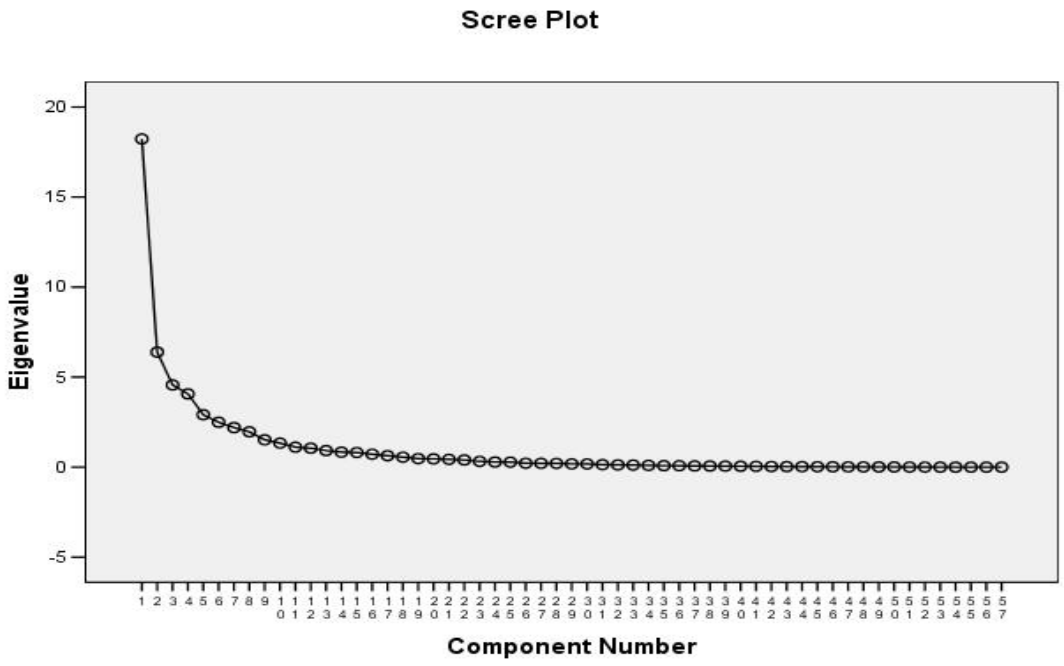
Hình 2.2: Biểu diễn các đối tượng khi các biến đã được chuẩn hóa.

Lại cắt tại ngưỡng chia làm 3 nhóm ta thấy có tiên bộ hơn nhưng vẫn chưa tạo ra được sự đồng đều tương đối về số liệu. Như vậy, phân tích chùm chưa đưa ra được

kết quả tương đối cho số liệu. Nên ta chuyển sang dùng phương pháp phân tích thành phần chính.

2.4 Kết quả áp dụng phương pháp phân tích thành phần chính

Khi phân tích chùm trên không gian các biến, chúng ta đã chia tập các biến ra làm 3 nhóm và cũng đã biết được mỗi nhóm có những biến nào, có những đặc trưng riêng của nhóm như thế nào. Tuy nhiên, số lượng các biến trong số liệu là lớn và theo lý thuyết về thổ nhưỡng thì ta cũng biết rằng không phải biến nào cũng quan trọng như nhau đối với đặc trưng riêng của nhóm hoặc đối với nhu cầu nghiên cứu của chuyên ngành thổ nhưỡng. Mặt khác, người ta cũng không thể phân tích chi tiết theo từng biến một, vì sẽ không tập trung được vào trọng tâm cần nghiên cứu. Để khắc phục nhược điểm đó, ta có thể dùng phương pháp phân tích thành phần chính trên không gian các biến để rút gọn số liệu, tổng hợp từ nhiều biến nguyên thủy ra một số biến quan trọng nhất, đại diện cho các nhóm biến và từ đó chúng ta có thể chia nhóm các đối tượng một cách rõ ràng và dễ diễn giải hơn.
Đưa 57 biến từ pH_{H_2O} *T*1 đến CatT3 vào để phân tích thành phần chính. Vì các biến được đo với đơn vị khác nhau nên ta dùng ma trận tương quan. Ta được kết quả như sau



Hình 2.3: Biểu diễn các thành phần chính.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	18.227	31.978	31.978	18.227	31.978	31.978
2	6.388	11.208	43.186	6.388	11.208	43.186
3	4.562	8.003	51.189	4.562	8.003	51.189
4	4.065	7.132	58.321	4.065	7.132	58.321
5	2.910	5.105	63.426			
6	2.498	4.374	67.801			
7	2.194	3.850	71.650			
8	1.959	3.438	75.088			
9	1.518	2.663	77.751			
10	1.333	2.339	80.090			
11	1.109	1.946	82.035			
12	1.053	1.848	83.883			
13	919	1.612	85.495			
14	834	1.463	86.958			
15	811	1.423	88.381			
16	717	1.258	89.639			
17	635	1.115	90.754			
18	563	.571	91.724			
19	472	.828	92.552			
20	457	.802	93.353			
21	431	.756	94.109			
22	399	.699	94.809			
23	320	.561	95.370			
24	285	.500	95.870			
25	279	.489	96.359			
26	217	.380	96.740			
27	.211	.370	97.110			
28	.205	.361	97.470			
29	.181	.317	97.787			
30	.174	.306	98.093			
31	.138	.241	98.335			
32	.123	.217	98.551			
33	.114	.200	98.751			
34	.098	.172	98.923			
35	.081	.143	99.066			
36	.078	.137	99.203			
37	.066	.115	99.319			
38	.060	.104	99.423			
39	.056	.098	99.521			
40	.049	.086	99.607			
41	.043	.075	99.683			
42	.033	.058	99.741			
43	.030	.052	99.793			
44	.026	.045	99.838			
45	.024	.042	99.880			
46	.018	.031	99.912			
47	.014	.024	99.936			
48	.014	.024	99.960			
49	.010	.017	99.977			
50	.009	.015	99.992			
51	.003	.005	99.998			
52	.001	.002	100.000			
53	1.85E-016	3.24E-015	100.000			
54	3.41E-017	5.98E-017	100.000			
55	-1.8E-016	-3.24E-015	100.000			
56	-8.6E-016	-1.50E-015	100.000			
57	-1.7E-015	-3.00E-015	100.000			

Hình 2.4: Tỉ lệ biến động của các thành phần chính.

Nhìn vào đồ thị 2.4, ta thấy rằng độ biến động chủ yếu do bốn thành phần chính đầu tiên cung cấp, bốn thành phần chính này chứa đến 58% độ biến động. Trong đó thành phần chính thứ nhất chứa 32% độ biến động, thành phần chính thứ hai chứa 11% độ biến động, thành phần chính thứ ba chứa 8% độ biến động và thành phần chính thứ tư chứa 7% độ biến động. Ta cũng thấy rằng, độ biến động không biến đổi đáng kể từ thành phần chính thứ năm trở về sau.

Mặc dù hai thành phần chính đầu tiên chứa < 50% nhưng số lượng thành phần chính là quá lớn và các thành phần chính phía sau có tỉ lệ biến động thay đổi chậm như nhau với giá trị rất nhỏ nên ta tập trung xét đến hai thành phần chính đầu tiên. Để thấy rõ ý nghĩa từng thành phần chính ta xem xét ma trận hệ số tương quan sau:

Dựa vào ma trận hệ số tương quan ta thấy rằng:

+) Thành phần chính thứ nhất chủ yếu liên quan đến pH_{H_2O}, pH_{KCl}, BS với hệ số tương quan tương ứng là 0.872, 0.757, 0.734 đối với tầng một, tương ứng 0.918, 0.897, 0.909 đối với tầng hai và tương ứng 0.891, 0.884, 0.893 đối với tầng ba. Ngoài ra còn liên quan đến Ca^{2+}, Mg^{2+}, K_2O, P_2O_5, Fe^{2+} và Fe^{3+}. Như vậy, có thể coi thành phần chính thứ nhất là liên quan chủ yếu đến độ kiềm của đất và ta gọi thành phần chính này là thành phần độ kiềm.

+) Thành phần chính thứ hai liên quan chủ yếu đến OC, N với hệ số tương quan tương ứng là 0.737, 0.722 đối với tầng một, tương ứng 0.349, 0.380 đối với tầng hai và tương ứng 0.368, 0.340 đối với tầng ba. Ngoài ra còn liên quan đến Ca^{2+},

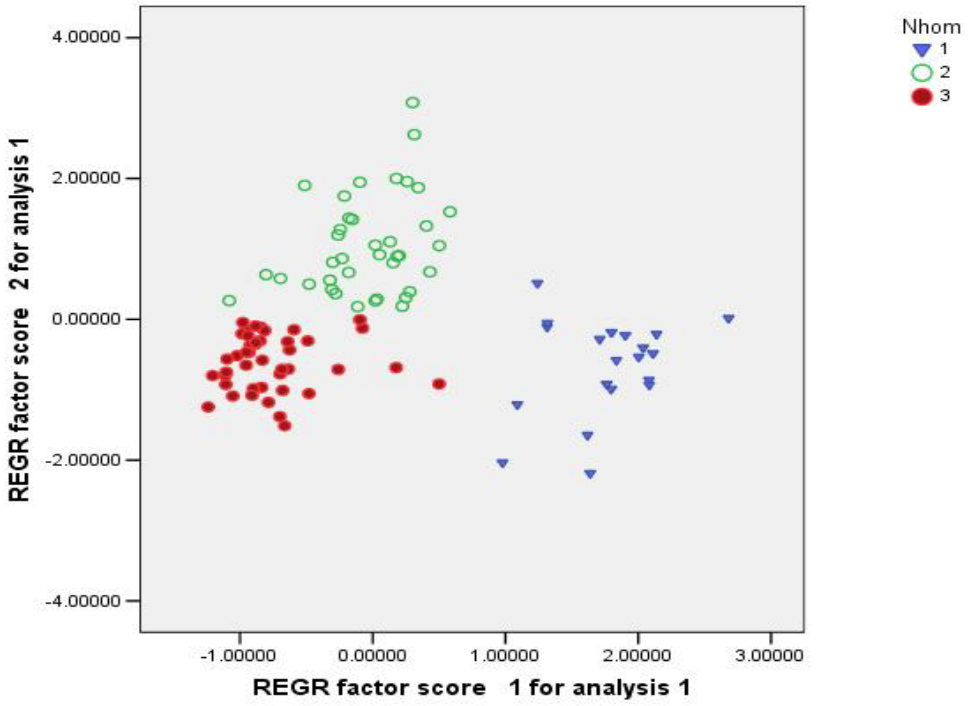
Correlations

		REGR factor score 1 for analysis 1	REGR factor score 2 for analysis 1	REGR factor score 3 for analysis 1	REGR factor score 4 for analysis 1	REGR factor score 5 for analysis 1	ph_h2oT1	ph__kclT1	Al3+(1dl/100g)	Ca2+(1dl/100g)	Mg2+(1dl/100g)	K+(1dl/100g)	Na+(1dl/100g)	CEC(1dl/100g)
REGR factor score 1 for analysis 1	Pearson Correlation	1	.000	.000	.000	.000	.782**	.757**	-.494**	.628**	.302**	.546**	-.175	-.231*
REGR factor score	Pearson Correlation	.000	1	.000	.000	.000	.147	.213*	-.605**	.524**	.627**	.075	.223*	.480**
REGR factor score	Pearson Correlation	.000	.000	1	.000	.000	-.441**	-.436**	.193	-.180	.048	-.184	.317**	.363**
REGR factor score	Pearson Correlation	.000	.000	.000	1	.000	.253*	.290**	-.024	.313**	-.077	.108	-.134	.272**
REGR factor score	Pearson Correlation	.000	.000	.000	.000	1	.072	.017	.008	-.053	.194	.567**	.339**	-.232*
ph_h2oT1	Pearson Correlation	.782**	.147	-.441**	.253*	.072	1	.969**	-.606**	.704**	.286**	.546**	-.253*	-.274**
ph__kclT1	Pearson Correlation	.757**	.213*	-.436**	.290**	.017	.969**	1	-.637**	.761**	.285**	.533**	-.231*	-.200*
Al3+(1dl/100g)	Pearson Correlation	-.494**	-.605**	.193	-.024	.008	-.606**	-.637**	1	-.624**	-.472**	-.331**	-.014	-.082
Ca2+(1dl/100g)	Pearson Correlation	.628**	.524**	-.180	.313**	-.053	.704**	.761**	-.624**	1	.563**	.455**	.043	.262**
Mg2+(1dl/100g)	Pearson Correlation	.302**	.627**	.048	-.077	.194	.286**	.285**	-.472**	.563**	1	.299**	.334**	.290**
K+(1dl/100g)	Pearson Correlation	.546**	.075	-.184	.108	.567**	.546**	.533**	-.331**	.455**	.299**	1	.059	-.112
Na+(1dl/100g)	Pearson Correlation	-.175	.223*	.317**	-.134	.339**	-.253*	-.231*	-.014	.043	.334**	.059	1	.259**
CEC(1dl/100g)	Pearson Correlation	-.231*	.480**	.363**	.272**	-.232*	-.274**	-.200*	-.082	.262**	.290**	-.112	.259**	1
bsT1	Pearson Correlation	.734**	.348**	-.394**	.245*	.124	.910**	.925**	-.660**	.847**	.472**	.576**	-.112	-.182
ocT1	Pearson Correlation	-.199*	.737**	.294**	.163	-.107	-.226*	-.153	-.137	.296**	.389**	-.175	.226*	.666**
nT1	Pearson Correlation	-.104	.722**	.293**	.164	-.040	-.138	-.081	-.184	.323**	.399**	-.125	.241*	.560**
p2o5T1	Pearson Correlation	.468**	.486**	-.139	.376**	.020	.559**	.598**	-.528**	.694**	.347**	.341**	-.153	.180
k2oT1	Pearson Correlation	.681**	-.190	-.371**	.379**	.071	.769**	.756**	-.217*	.541**	.063	.533**	-.287**	-.226*
K2O(mg/100g)	Pearson Correlation	.546**	.075	-.184	.108	.567**	.546**	.533**	-.331**	.455**	.299**	1.000**	.059	-.112
P2O5(mg/100g)	Pearson Correlation	.728**	-.045	-.440**	.361**	.014	.865**	.880**	-.421**	.613**	.121	.504**	-.299**	-.287**

Hình 2.5: Ma trận hệ số tương quan.

Mg^{2+} và cat. Như vậy, có thể coi thành phần chính thứ hai liên quan đến hàm lượng chất hữu cơ và nitơ trong đất và gọi là thành phần chất hữu cơ.

Ở trên ta chỉ mới biết được đặc trưng của hai thành phần chính. Để biết được mối quan hệ của chúng và từ đó có thể phân tích số liệu ta đi biểu diễn các đối tượng trên mặt phẳng chính như sau: Các đối tượng phân nhóm rõ ràng, cụm lại thành 3 nhóm chính. Từ mặt phẳng chính trên ta thấy rằng, có thể chia không gian các đối tượng ra làm ba nhóm khác nhau trong đó: Nhóm thứ nhất có thành phần độ kiềm lớn hơn 0.8 và thành phần chất hữu cơ nhỏ hơn 0. Nhóm thứ hai có thành phần độ kiềm nhỏ hơn hoặc bằng 0.8 và thành phần chất hữu cơ nhỏ hơn 0. Nhóm thứ ba có thành phần độ kiềm nhỏ hơn hoặc bằng 0.8 và thành phần chất hữu cơ lớn hơn hoặc bằng 0. Đối chiếu cách phân nhóm trên với bảng số liệu gốc ta thấy rằng các vùng đất trong một xã thường là nằm cùng một nhóm, một "mẫu" nằm vào nhóm khác so với các mẫu đất cùng xã chỉ khi mẫu đó được điều tra ở độ dốc khác so với các mẫu đất kia. Trong ba nhóm đất kể trên, nhóm 1 bao gồm các mẫu đất có phẫu diện với đặc điểm chung là được điều tra ở địa hình E2 trong trạng thái mặt đất ẩm ướt xói mòn ít, thực vật tự nhiên chủ yếu là cỏ, đồng thời lúa là cây trồng chủ yếu trong khu vực. Nhóm 2 bao gồm các mẫu đất có phẫu diện với đặc điểm chung là được điều tra ở địa hình vàn, E2 trong trạng thái mặt đất khô xói mòn mạnh, thực vật tự nhiên chủ yếu là mua, tè, còn cây trồng trọt chủ yếu trong khu vực là lúa và bạch đàn. Nhóm 3 bao gồm các mẫu đất có phẫu diện với đặc điểm chung là được điều tra ở địa hình E2 trong trạng thái mặt đất ẩm xói mòn trung bình, thực vật tự



Hình 2.6: Biểu diễn các đối tượng trên mặt phẳng chính.

nhiên chủ yếu là cây bụi, cỏ, còn cây trồng trọt chủ yếu là lúa, chè và bạch đàn.
Khi chia nhóm bằng hai phương pháp trên ta thấy rằng, nếu dùng phân tích chùm để chia nhóm thì hầu hết các đối tượng được xếp vào nhóm 1, nhóm 2 và nhóm 3 chỉ chứa rất ít các đối tượng do vậy việc phân nhóm bằng phương pháp phân tích chùm không cho kết quả hợp lý. Trong khi đó, nếu ta dùng phương pháp phân tích thành phần chính trên hai thành phần chính thứ nhất và thứ hai để chia nhóm đối tượng thì ta có thể chia các mẫu đất thành ba nhóm có số lượng tương đối đồng đều. Mặt khác, có thể dựa vào ý nghĩa các thành phần chính đó để diễn giải đặc tính riêng của từng nhóm.
Mặc dù các nhóm đất trên có các đặc điểm sinh hóa khác nhau về độ kiềm, về thành phần hữu cơ nhưng vẫn được sử dụng để trồng lúa là chủ yếu. Như vậy, cần phải xác định được chế độ canh tác riêng, phù hợp cho từng nhóm đất để có được thu hoạch lớn nhất. Chẳng hạn, đối với nhóm 1, là nhóm có độ kiềm tương đối cao để nâng cao năng suất lúa cần tập trung vào việc bổ sung thêm các thành phần hữu cơ cho đất. Đối với nhóm 2, là nhóm có độ kiềm tương đối nhỏ đồng thời thành phần hữu cơ ít nên phải đồng thời có những tác động để giảm độ chua của đất và dùng phân hữu cơ, phân chuồng để tăng cường chất hữu cơ cho đất. Đối với nhóm 3, là nhóm có thành phần hữu cơ cao nên phải cải tạo đất tập trung vào việc giảm độ chua.

KẾT LUẬN

Mặc dù luận văn chưa đưa ra được những kết quả chuyên sâu về phân tích thổ nhưỡng của huyện Thanh Ba - Phú Thọ, nhưng luận văn đã làm được một số vấn đề chính sau:

- Luận văn đã trình bày chi tiết về những khái niệm mở đầu của phân tích chùm và phân tích thành phần chính. Các ví dụ trong phần mở đầu cho thấy các kỹ thuật này có ứng dụng rất rỗng rãi trong nhiều lĩnh vực khác nhau để phân tích dữ diệu.
- Luận văn đã áp dụng những công cụ kể trên để phân tích thổ nhưỡng đất trồng trọt của huyện Thanh Ba - Phú Thọ, và đã đưa ra được một số kết quả có tính thực hành cao.
- Luận văn cũng chỉ ra rằng việc tính toán bằng tay là không khả thi khi thực hiện trên một tập dữ liệu lớn, do vậy việc sử dụng phần mềm SPSS hoặc tương đương là cần thiết. Vì vậy trong luận văn có trình bày ở mức độ cơ bản những thao tác cần thực hiện trên SPSS.
- Hướng nghiên cứu tiếp là sử dụng các phương pháp đã được trình bày ở trên để nghiên cứu thổ nhưỡng đối với các khu vực rộng hơn, nhiều tỉnh hơn. Có thể giúp cho công tác quy hoạch đất nông nghiệp.

Tài liệu tham khảo

[1] Bộ môn khoa học đất trường Đại học Nông Nghiệp Hà Nội (2006), *Giáo trình Thổ Nhưỡng học*, Nxb Nông Nghiệp, Hà Nội.

[2] Hội khoa học đất Việt Nam (1999), *Sổ tay điều tra phân loại đánh giá đất*, Nxb Nông Nghiệp, Hà Nội.

[3] Nguyễn Đình Hiền (2008), *Xử lý thống kê nhiều chiều và ứng dụng trong Nông Nghiệp*, Nxb Nông Nghiệp, Hà Nội.

[4] Nguyễn Văn Hữu, Nguyên Hữu Dư (2003), *Phân tích thống kê và dự báo*, Nxb Đại học Quốc Gia Hà Nội.

[5] Joseph F. Hair. Jr., Rolph E. Anderson, Ronald L. Tatham, William C. Black (1992), *Multivariate Data Analysis with Reading*, Macmillan Publishing Company New York.