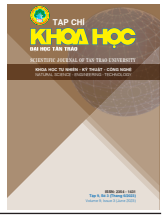
TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/



APPLYING PRINCIPAL COMPONENT ANALYSIS AND CLUSTERING TO ASSESS ACCREDITATION RESULTS INHIGHER EDUCATION INSTITUTIONS

Phuoc Thanh Le

Quang Nam University, Vietnam

Email address: lephuocthanhkt@gmail.com

DOI: 10.51453/2354-1431/2023/976

Article info

Received: 25/12/2012

Revised: 22/03/2023

Accepted: 16/5/2023

Keywords:

Principal Component Analysis, Clustering, K-MEANS clustering algorithm,Correlation Coefficient, Higher Education Quality Accreditation

Abstract:

Currently, the centers for education accreditation (CEA) have announced university accreditation results by the standard set under Circular 12/2017 / TT-BGDĐT. The accreditation results are standardized in the form of a multi-dimensional database based on these standards. This research is carried out in a combination of two main techniques- principal component analysis and clustering- to present, analyze and extract useful knowledge from the accreditation results. At the same time, the paper points out the educational institutions' strengths and weaknesses based on the standards, the relationship between different fields as well as compare the assessment levels among accreditation centers. This is the foundation to compare and improve the quality in educational institutions.

TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO
ISSN: 2354 - 1431
http://tckh.daihoctantrao.edu.vn/

ỨNG DỤNG PHƯƠNG PHÁP PHÂN TÍCH THÀNH PHẦN CHÍNH VÀ PHÂN CỤM DỮ LIỆU ĐÁNH GIÁ KẾT QUẢ KIỂM ĐỊNH CHẤT LƯỢNG CƠ SỞ GIÁO DỤC ĐẠI HỌC

Lê Phước Thành
Đại học Quảng Nam, Việt Nam
**Địa chỉ email: lephuocthanhkt@gmail.com*
DOI:10.51453/2354-1431/2023/976

Thông tin bài viết

Ngày nhận bài: 25/12/2012
Ngày sửa bài:22/03/2023
Ngày duyệt đăng: 16/5/2023

Từ khóa:

Phân tích thành phần chính, Phân cụm dữ liệu, Thuật toán K-Means, Hệ số tương quan, Kiểm định chất lượnggiáo dục đại học.

Tóm tắt

Hiện nay các trung tâm kiểm định chất lượng giáo dục đại học (KĐCLGDĐH) đã công bố kết quả kiểm định các trường đại học theo bộ tiêu chuẩn của thông tư 12/2017/TT-BGDĐT. Kết quả kiểm định được chuẩn hóa dưới dạng một cơ sở dữ liệu đa chiều theo các tiêu chuẩn. Sự kết hợp giữa hai kỹ thuật phân tích thành phần chính với phân cụm dữ liệu nhằm trình bày, phân tích và trích ra những tri thức hữu ích trong việc đánh giá. Theo đó, bài báo chỉ ra những điểm mạnh, yếu về hoạt động của các trường theo các tiêu chuẩn, mối quan hệ giữa các lĩnh vực cũng như so sánh mức độ đánh giá giữa các trung tâm kiểm định với nhau. Đây là cơ sở để thực hiện việc đối sánh và cải tiến chất lượng tại các cơ sở giáo dục.

1. Introduction

1.1. Introduction to quality assurance in higher education institutions in Vietnam

On May 19, 2017, the Ministry of Education and Training issued Circular No. 12/2017/TT-BGDĐT, which provides regulations on quality assurance in higher education institutions. According to this circular, the set of evaluation criteria consists of 25 standards and 111 criteria, divided into four domains:

(1) Quality assurance in terms of strategy: Standards 01 to 08 cover issues related to mission, vision, purpose, strategic objectives, and policies.

(2) Quality assurance in terms of systems: Standards 09 to 12 address issues regarding internal quality assurance systems, information systems, and more.

(3) Quality assurance in terms of function implementation: Standards 13 to 21 focus on issues related to educational activities, scientific research, and community services.

(4) Performance outcomes: Standards 22 to 25 encompass issues related to the outcomes of educational activities, scientific research, community services, and financial-market aspects.

Each standard is assessed on a 4-point scale. This set of standards follows the evaluation model for quality assurance in higher education known as the ASEAN University Network - Quality Assurance (AUN-QA).

As of September 2020, Vietnam has five quality assurance centers that have announced the assessment results for 28 universities and institutes based on these

standards. The centers and the number of institutions assessed are as follows:

(1) Center for Quality Assurance in Higher Education - Hanoi National University (CEA_HN), 6 institutions.

(2) Center for Quality Assurance in Higher Education - Ho Chi Minh City (CEA_TPHCM), 5 institutions.

(3) Center for Quality Assurance in Higher Education - University of Danang (CEA_DN), 4 institutions.

(4) Center for Quality Assurance in Higher Education - Vinh University (CEA_Vinh), 5 institutions.

(5) Center for Quality Assurance in Higher Education - Vietnam Association of Universities and Colleges (CEA_HiepHoi), 8 institutions.

1.2. Principal Component Analysis and Data Clustering Techniques

Principal Component Analysis (PCA) is a commonly used technique when working with datasets that have a high number of variables (attributes/ dimensions) represented in a multi-dimensional space but need to be visualized in 2 or 3 dimensions while preserving the variability of the original data. PCA also allows for the discovery of underlying relationships in the data that can be explored in the new space. The two main purposes of PCA are to find the relationship between objects and the dimensions of the new space and to examine the relationships between the original variables in the new space.

When objects are represented in a 2-dimensional space, with the horizontal axis being the first principal component (Component 1) and the vertical axis being the second principal component (Component 2), data clustering techniques can be applied to group objects that share similar characteristics based on certain criteria (e.g., distance), while objects from different clusters do not share the same characteristics.

Data clustering is a method used to identify groups or clusters of objects based on their similarity or dissimilarity. It helps to uncover patterns, structures, or relationships within the data. By applying clustering techniques to the transformed data from PCA, objects can be grouped together based on their proximity in the new space, enabling the identification of distinct clusters or subgroups within the dataset.

Overall, the combination of Principal Component Analysis and data clustering techniques provides a powerful approach to analyze and understand complex datasets, allowing for the visualization of data in reduced dimensions while discovering underlying patterns and grouping similar objects together based on certain criteria.

2. Research Methodology

2.1. Dataset for Analysis

The dataset used for analysis is obtained from the published results of quality assurance assessments on the websites of the 5 quality assurance centers [9], [10], [11], [12], [13] The dataset is collected from 28 universities (sample size: 28). The analysis attributes (dimensions) consist of 25 standards (T1->T25) with evaluation values on a 4-point scale, as shown in Table 1:

Table 1. Lookup Table of University Order Numbers in the Analysis

STT	TT KĐ	Trường đại học đã Kiểm định	T1	…	LV1	…
1.	DN	C. Nghệ TP HCM	4.60	.	4.44	…
2.	DN	Quốc tế Sài Gòn	4.00		3.93	
3.	DN	SPKT Vĩnh Long	4.20		4.07	
4.	DN	Văn Hiến	4.00		3.88	
5.	DN	Nội vụ Hà Nội	3.80		3.79	
6.	DN	Phan Thiết	4.20		3.84	
7.	DN	Phennikaa	4.40		4.01	
8.	DN	TDTT Hà Nội	4.00		3.84	
9.	DN	Thủy Lợi	4.60		4.37	
10.	DN	HV Ngoại giao	4.20		3.98	
11.	TPHCM	Đà Lạt	4.00		3.82	
12.	TPHCM	K.tế-TC TP HCM	3.80		3.81	
13.	TPHCM	Quốc tế Miền Đông	4.00		4.03	
14.	TPHCM	Trà Vinh	4.20		4.16	
15.	TPHCM	Văn hóa TP HCM	4.00		3.64	

STT	TT KĐ	Trường đại học đã Kiểm định	T1	…	LV1	…
16.	Vinh	C. nghệ M. Đông	3.80		3.72	
17.	Vinh	FPT	4.80		4.56	
18.	Vinh	Hoa Lư	3.80		3.79	
19.	Vinh	K. Tế C.N Long An	4.00		3.96	
20.	Vinh	Thủ Đô Hà Nội	4.00		4.03	
21.	HiepHoi	Bà Rịa-Vũng Tàu	4.00		4.19	
22.	HiepHoi	Đại Nam	4.00		3.85	
23.	HiepHoi	Dầu khí Việt Nam	4.20		4.25	
24.	HiepHoi	Đ.dưỡng Nam Định	4.40		4.04	
25.	HiepHoi	Hoa Sen	4.20		3.94	
26.	HiepHoi	Quốc tế Hồng Bàng	4.60		4.44	
27.	HiepHoi	Tân Trào	4.20		4.16	
28.	HiepHoi	Học viện Phụ nữ	3.80		3.83	

Below is a reference table (Table 2) that provides the names of the 25 standards for convenient tracking and evaluation of the strengths and weaknesses of the universities based on these standards. These standards are categorized into 4 domains (Table 3), and the scores for these 4 domains are the average scores of the standards within each domain.

Here is the reference table (Table 2) that lists the names of the 25 standards for easy monitoring and evaluation of the analysis results, assessing the strengths and weaknesses of the universities based on these standards. These standards are categorized into 4 domains (Table 3), and the scores for the 4 domains are calculated as the average scores of the standards within each domain.

Table 2: List of Standards

ordinal number	Standard name
1	Vision, mission, and culture
2	Management
3	Leadership and administration
4	Strategic management
5	Books on IT, research, and technology transfer
6	Human resource management
7	Financial and infrastructure management
8	Networks and international relations
9	Internal quality assurance system
10	Self-assessment and external evaluation
11	Internal information system
12	Quality improvement
13	Admissions and enrollment
14	Curriculum
15	Teaching and learning
16	Learner assessment
17	Learner support
18	Research management
19	Intellectual property asset management
20	Collaboration and research partnerships
21	Connection and community service
22	Training outcomes
23	Research outcomes
24	Community service outcomes
25	Financial and market outcomes

Table 3. Table of Inspection Fields Lookup

Area	Standards included
Quality Assurance in Strategy	1->8
Quality Assurance in Systems	9-12
Quality Assurance in Function Implementation	13->21
Operational Results	22->25

2.2. Algorithm

2.2.1. Principal Component Analysis [1], [2]

Bài toán: Provide the matrix X = $\{x_{i,j}\}$, Like this

(i) An object can be represented in space R^p, where each point x_{i1}, x_{i2},... x_{ip} has coordinates , i = $\overline{1,n}$ referred to as the space of objects.

(ii) A variable can be represented in space R^n, where each variable X_j has coordinates $X_j = x_{1j}$, x_{2j},... x_{nj}, j = $\overline{1,p}$ referred to as the space of variables.

The following steps aim to find the principal components in the space of objects (case (i)). For case (ii), a similar procedure is performed in the space of variables.

The steps to perform are as follows:

(1) Determine the center of the data cloud

Each object is always represented as a point in space, and the collection of these points is called a data cloud. Centering means translating the coordinate origin to the centroid of the data cloud. The center of the data cloud is achieved by transforming the data matrix into a matrix of deviations from the mean.
Each object i of the variable X_j subtracted by the mean value $\overline{x_j}$ of variable X_j. We obtain the centered matrix X = (x_{ij})np

(2) Find the principal axes.

a) Variance-Covariance Matrix

Variance-Covariance Matrix used to assess the variation (concentration or dispersion) of the data around the center of the data cloud. This matrix is calculated in the new coordinate system as follows:

X': The transpose matrix of matrix X.

If we represent the variation of the data geometrically, it means finding a line that passes through the center of the data cloud and is "close" to the data points, where the distance from the points to the line is minimized. In other words, it corresponds to finding the projection of the points onto the first axis (principal component 1) with the largest variation (variance).

b) Eigenvalues and Eigenvectors

To find the eigenvalues and eigenvectors for determining the lines that pass through the center and are closest to the data cloud, we need to calculate the eigenvalues. To find the eigenvalues, we perform the following steps: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ according to the equation:

|M0 – λI| = 0, I: unit matrix

In geometric terms, eigenvalues is the sum of squared distances of points to the lines such that this value is minimized. For each value λ_j (j = $\overline{1q}$, q<p) To determine the corresponding eigenvectors (unit vectors) for each value uj = (u_{1j}, u_{2j}, u_{pj},) by solving the equation:

Eigenvectors are a way to determine the variation between the projected points on the new axis with the new unit compared to the variation of the data on the old coordinate system with a unit variance of 1.

Based on the eigenvalue λ_j and eigenvector u_i to determine the first principal component (first principal axis), in PCA, the second principal axis passes through the center and is orthogonal to the first principal axis, the third principal axis passes through the center and is orthogonal to the plane formed by the previous two axes, and so on.

(3) Representing objects in the new coordinate system.

The projection of object i onto the j-th principal axis is zij = $x'_i u_j$

2.2.2. Data clustering using the K-Means algorithm [1], [2], [3]:

Problem:

Input: Given a database with n objects and k clusters.

Output: Assign each object to one of the k clusters.

Steps:

Step 1. Initialization: Randomly select k points as centroids.

Step 2. Calculate distances: For each object, calculate the distance to each centroid. Assign the object to the cluster with the closest centroid.

Step 3. Update centroids: Calculate the average distance between objects within each cluster and update the centroid (the centroid is the average distance between objects in the cluster).

Step 4. Termination condition: Repeat steps 2 and 3 until the centroids of the clusters no longer change.

2.2.3. The relationships between variables

When representing the original variables in a new coordinate system with two principal components, the relationships between variables are determined as follows:

If the angle between two small vectors (close to each other) is small, the variables have a strong correlation or interdependence.

If two vectors are nearly orthogonal, there is negligible dependence or no correlation between them.

If two vectors are opposite to each other by 180 degrees, it indicates a negative correlation.

Regarding objects, the relationship with variables is determined as follows: When objects are located on the positive side of the axis corresponding to a particular component, they have high values for variables close to that component, and vice versa.

3. Results and discussion

The following are the results of the analysis using charts and some evaluations by field: In field 1 (Figure 1): The schools with serial numbers 1, 17, 9, and 26 were rated the highest in all criteria in this field. In contrast, the schools with serial numbers 11 and 15 were rated the lowest in criteria 4, 5, and 7. Additionally, the schools with serial numbers 4, 18, and 22 were rated low in criteria 2, 3, and 8. The schools located around the origin were evaluated as average in terms of the criteria.

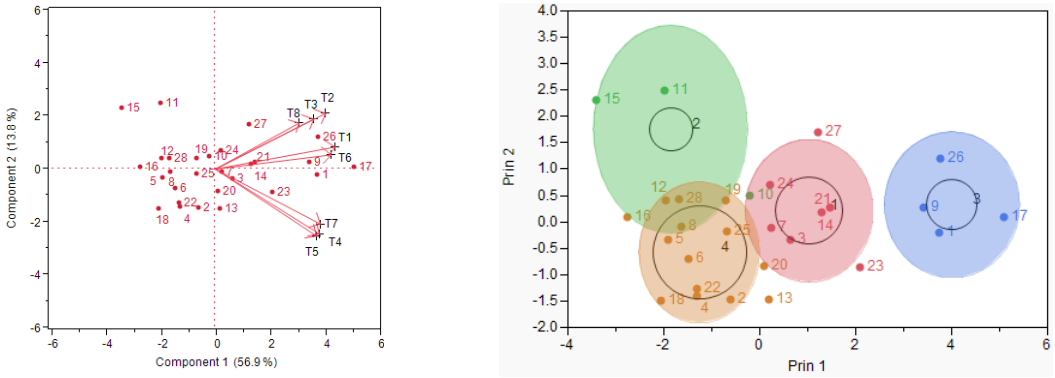


Figure 1 illustrates the distribution of objects according to their main components and the clustering approach in domain 1.

In domain 2 (Figure 2), the fields within group 1 and 26 are rated highest across all criteria within this domain. On the other hand, the fields within group 12 and 15 are rated lowest among the four specific domains, particularly in criterion 9. Field 11 is rated very low in criterion 12.

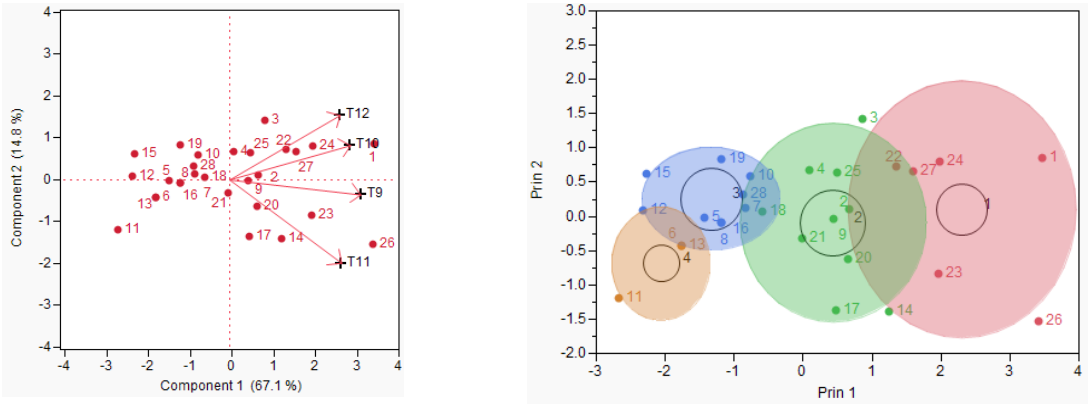


Figure 2 depicts the distribution of objects according to their main components and clustering approach in domain 2.

In domain 3 (Figure 3), the fields within group 1, 17, and 26 are rated highest across all criteria within this domain, with field 17 particularly excelling in criteria 13, 15, 16, and 17. On the other hand, fields 11 and 15 are rated lowest in criterion 21. Fields within group 6, 11, and 12 are rated low in criteria 14, 18, 19, and 20, but field 12 receives a very high rating in criterion 17.

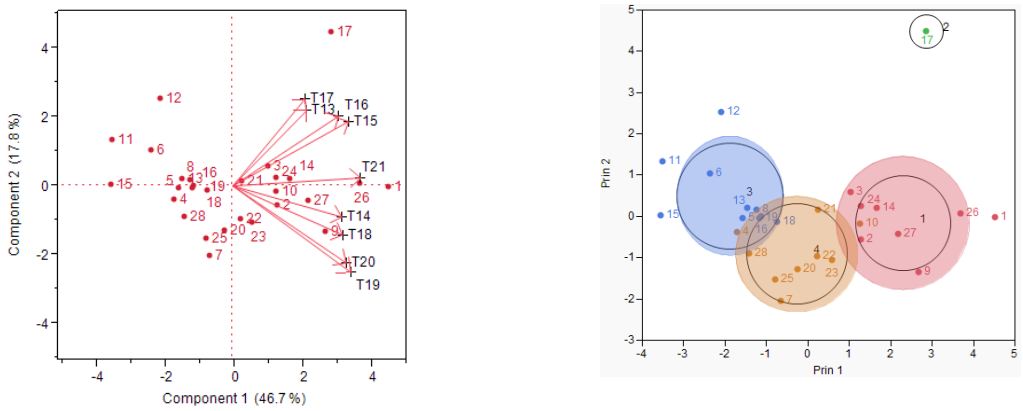


Figure 3 illustrates the distribution of objects according to their main components and the clustering approach in domain 3.

In domain 4 (Figure 4), the fields within group 1 and 17 are rated highest in criteria 22, 24, and 25. However, field 17 receives a low rating in criterion 23. The fields within group 7, 9, and 1 are highly rated in criterion 23, while fields within group 6, 13, and 15 receive low ratings in criterion 23. Field 11 is rated low in criteria 22, 24, and 25.

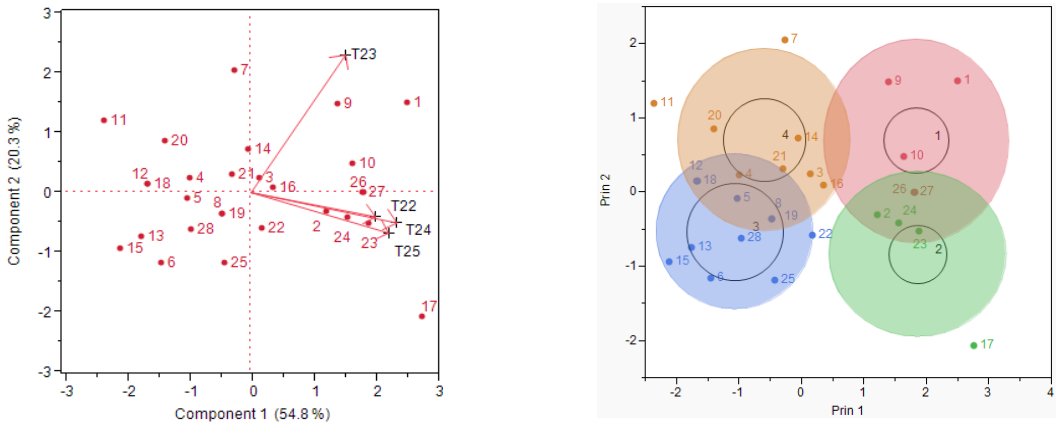


Figure 4 displays the distribution of objects according to their main components and the clustering approach in domain 4.

In summary, for the four domains (Figure 5), each domain consists of several criteria within that specific domain. The score of each domain is calculated as the average of its criteria.

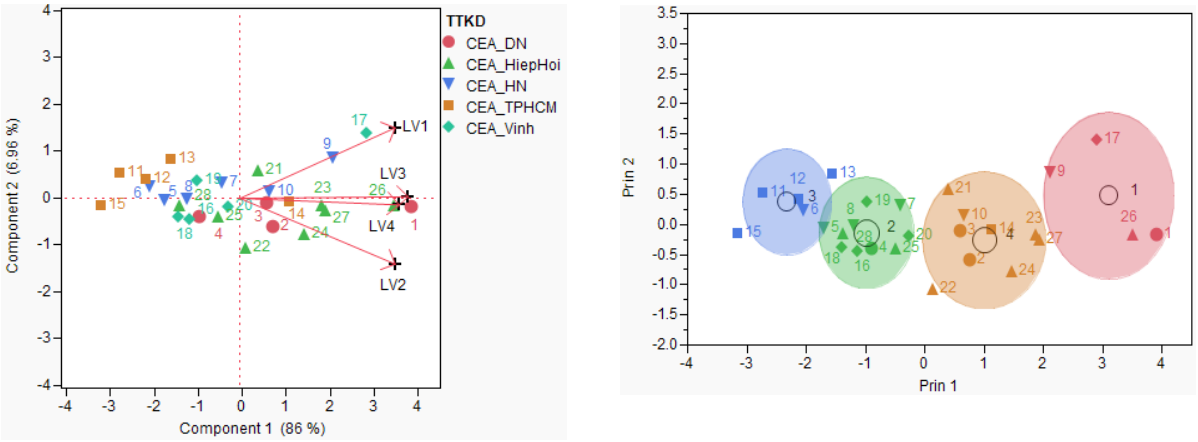


Figure 5: The distribution chart of objects based on their main components, clustering in the four domains, and the centers responsible for evaluating the fields.

Based on the eigenvalues (Table 4), the number of principal components can be determined. By selecting eigenvalues >= 1, there is only one principal component that includes all four domains extracted, explaining 85.9% of the data variation (retaining 85.9% of the initial information).

Table 4: Eigenvalues and the percentage of explained variance in the data.

TT	The eigenvalue	Percentage	Percentage chart.	Cumulative percentage.
1	3.4384	85.959		85.959
2	0.2785	6.962		92.921
3	0.2240	5.601		98.521
4	0.0591	1.479		100.000

Based on the eigenvectors, we can determine the relationship between the principal component and the variables. In other words, this represents the linear relationship between the principal component and the variables. The relationship is depicted in the loading matrix table (Table 5) as follows:

Table 5: Loading Matrix of Principal Components.

Field	Principal components			
	1	2	3	4
LV1	0.90147	0.38525	0.18539	0.06760
LV2	0.90174	-0.35876	0.23410	0.05791
LV3	0.97774	0.00926	-0.04143	-0.20550
LV4	0.92553	-0.03548	-0.36489	0.09483

Here are some evaluations based on the analysis of the 4 domains:

(i) The universities in group 1, 26, and 17 are highly rated in all 4 domains, especially university 27, which has the highest rating in domain 1. On the other hand, universities 15, 11, 6, and 12 receive low ratings in all 4 domains, with university 15 being the lowest rated in domains 3 and 4.

(ii) The validation results of the centers are relatively evenly distributed across high, medium, and low levels for the evaluated universities. However, the validation results for universities under the National University Center - Ho Chi Minh City are mostly low compared to other centers. This includes universities with order numbers 11, 12, 13, and 15.

(iii) The relationships between the domains: Constructing the correlation coefficients between domains (Figure 6), it is observed that domain 3 and 4 have the highest correlation coefficient, indicating that domain 3, concerning the function, system, policies of education, scientific research, and community service, directly impacts the performance in domain 4. On the other hand, the correlation coefficient between domain 2 and 3 is lower, suggesting a lack of strong connection between domain 1, which involves mission, vision, strategic objectives, and policies, and domain 2, which focuses on building an internal quality assurance system and information system.

Furthermore, when rotating the data with columns representing the evaluated universities and rows representing the domains, it can be observed that domain 3 is positioned near the origin, indicating that the scores in domain 3 do not have significant variations among universities (relatively consistent). On the other hand, domains 1, 2, and 4 show differences among universities. Specifically, universities with functions, systems, and policies regarding education, scientific research, and community service are relatively consistent.

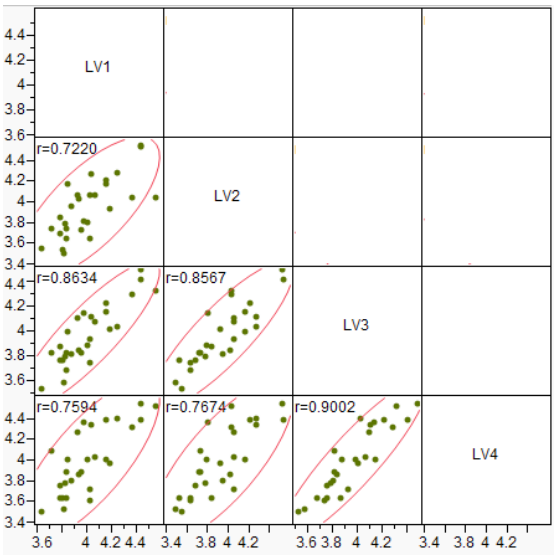


Figure 6. Correlation coefficients between domains

Another way to cluster the data is hierarchical clustering [2], [3], as shown in Figure 7. Here, the data is divided into 4 clusters, visually presenting the fields with high validation results such as Ho Chi Minh City University of Technology, Hong Bang International University, etc., and the fields with low validation results such as Phan Thiet University, Ho Chi Minh City University of Economics and Finance, etc. Both the K-Means clustering algorithm and hierarchical clustering approach result in equivalent clustering of the university groups.

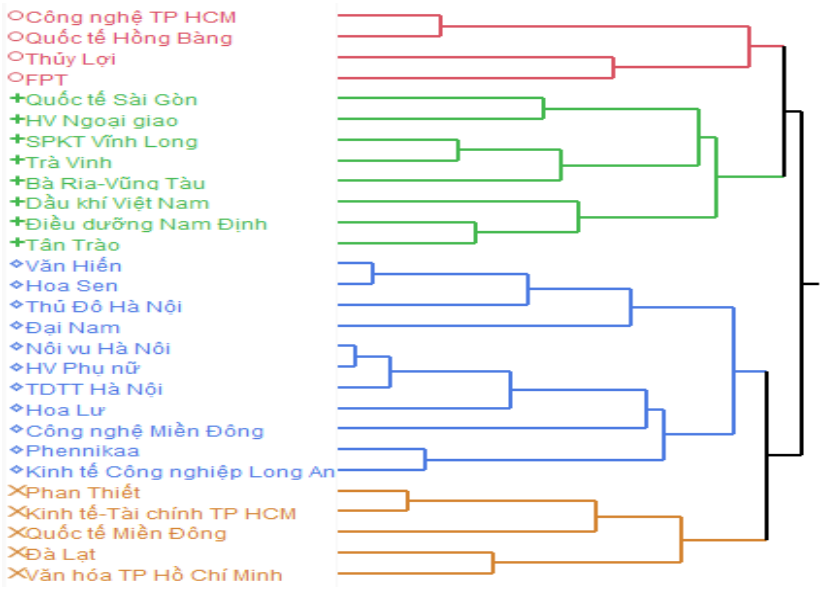


Figure 7. Hierarchical clustering dendrogram

4. Conclusion

The principal component analysis method based on mathematical models is a linear transformation from one space to another, where the dimensionality of the data is reduced while retaining most of the information. This method is advantageous for presenting, analyzing, and evaluating the quality of activities in educational institutions according to standards and fields. When the original problem space is projected onto a 2-dimensional plane with two principal components extracted, clustering techniques can be further applied based on the "similarity" between objects within each group. This helps group the institutions based on their strengths and weaknesses in different clusters for evaluation. The combination of these two techniques aims to visually represent the grouping of educational institutions according to the principal components (standards and fields).

This is just the result of evaluating 28 universities and colleges based on the standards set in Circular 12/2017/TT-BGDĐT. As the inspection centers provide more comprehensive evaluation results, analyzing the relationships between standards and fields will create opportunities for universities to have appropriate directions in establishing an internal quality assurance system within the institution.

References

[1] Đỗ Phúc (2008), *Giáo trình khai thác dữ liệu*. Nhà xuất bản Đại học Quốc gia TP Hồ Chí Minh.

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

[3] Hoàng Trọng & Chu Nguyễn Mộng Ngọc (2005), *Phân tích dữ liệu nghiên cứu với SPSS, tập 2*. Nhà xuất bản Thống kê.

[4] Hoàng Xuân Huấn (2015), *Giáo trình học máy,* Đại học Quốc gia Hà Nội.

[5] ZHOU Shuangxi (2015), *University Teachers' Performance Comprehensive Evaluation Based on Principal Component Analysis*, Higher Education of Social Science, CSCanada

[6] MengYi (2019), *Application of Principal Component Analysis in Teaching Evaluation*, Published by Francis Academic Press, UK

[7] *JMP 13 Multivariate Methods*, *Second Edition (2017)*. Cary, NC: SAS Institute Inc.

[8] Thông tư *12/2017/TT-BGDĐT Ban hành Quy định về kiểm định chất lượng cơ sở giáo dục đại học,* Bộ Giáo dục và Đào tạo.

[9] http://cea.udn.vn, truy cập ngày 24/09/2020

[10] http://cea.vnuhcm.edu.vn, truy cập ngày 24/09/2020

[11] http://cea.vnu.edu.vn, truy cập ngày 24/09/2020

[12] http://kdclgd.vinhuni.edu.vn, truy cập ngày 24/09/2020

[13] http://cea-avuc.edu.vn, truy cập ngày 24/09/2020