



Phân tích phân cụm các doanh nghiệp ở Việt Nam, Thái Lan và Indonesia trong giai đoạn 2018–2022:

Ứng dụng phương pháp phân tích dữ liệu đa hợp

ĐÀO THỊ THANH BÌNH^{*,a}, NGUYỄN ANH TUẤN^a, PHẠM VĂN HÙNG^a, LÊ THANH BÌNH^a

^a Trường Đại học Hà Nội

THÔNG TIN	TÓM TẮT
Ngày nhận: 25/12/2024 Ngày nhận lại: 29/04/2025 Duyệt đăng: 04/05/2025 Mã phân loại JEL: C58; F37; G17. Từ khóa: Phân tích dữ liệu đa hợp; Phân tích phân cụm; Chỉ số tài chính; Phân loại doanh nghiệp; COVID-19. Keywords: Compositional data analysis; Cluster analysis; Financial ratios;	Bài nghiên cứu tập trung vào việc sử dụng các chỉ số tài chính để phân tích phân cụm các doanh nghiệp, sử dụng phương pháp phân tích dữ liệu đa hợp. Với dữ liệu lấy từ báo cáo tài chính của 300 doanh nghiệp lớn tại Thái Lan, Indonesia và Việt Nam, bài nghiên cứu hướng đến việc phân cụm các doanh nghiệp và từ đó tìm hiểu được đặc tính của từng phân cụm cũng như sự dịch chuyển của doanh nghiệp giữa các cụm trong giai đoạn 2018–2022. Kết quả thu được cho thấy không có sự khác biệt rõ rệt khi sử dụng cả 2 phương pháp tỷ lệ logarit trung tâm và tỷ lệ logarit cộng tính, các chỉ số tài chính đều thể hiện tính hiệu quả trong việc phân loại doanh nghiệp, phân chia thành 3 phân cụm dựa theo hiệu quả kinh doanh và mức độ rủi ro tài chính. Phân tích trong giai đoạn trước và sau COVID-19 cũng cho thấy dấu hiệu phục hồi về kết quả kinh doanh thông qua sự chuyển dịch phân cụm của một số công ty. Đáng chú ý hơn, các doanh nghiệp Việt Nam thể hiện hiệu suất hoạt động tốt và duy trì sự hiện diện với số lượng lớn trong nhóm có kết quả hoạt động tốt nhất với mức độ an toàn thanh toán cao. Abstract The study focuses on using financial indicators to analyze the clustering of businesses, employing a compositional data analysis method. With data taken from the financial reports of 300 large enterprises in Thailand, Indonesia, and Vietnam, the study aims to cluster the

* Tác giả liên hệ.

Biên tập viên: Nguyễn Thị Hồng Cẩm.

Email: binhdt@hanu.edu.vn (Đào Thị Thanh Bình), tuanna@hanu.edu.vn (Nguyễn Anh Tuấn), hungpv@hanu.edu.vn (Phạm Văn Hùng), binhlt@hanu.edu.vn (Lê Thanh Bình).

Trích dẫn bài viết: Đào Thị Thanh Bình, Nguyễn Anh Tuấn, Phạm Văn Hùng, & Lê Thanh Bình. (2025). Phân tích phân cụm các doanh nghiệp ở Việt Nam, Thái Lan và Indonesia trong giai đoạn 2018–2022: Ứng dụng phương pháp phân tích dữ liệu đa hợp. *Tạp chí Nghiên cứu Kinh tế và Kinh doanh Châu Á*, 36(3), 61–78.

Business classification;
COVID-19.

enterprises and thereby understand the characteristics of each cluster as well as the movement of enterprises between clusters during the period 2018-2022. The results obtained show no significant difference when using both CLR and ALR methods, the financial indicators are effective in classifying companies, dividing them into three clusters based on business efficiency and financial risk levels. Analysis in the pre-COVID-19 and post-COVID-19 periods also shows signs of recovery in business results through the clustering shifts of some companies. Notably, Vietnamese businesses demonstrate good operational performance and maintain a significant presence in the group with the best performance results and high payment safety.

1. Giới thiệu

Sau đại dịch COVID-19, cả các quốc gia phát triển và đang phát triển đều gặp khó khăn trong phục hồi kinh tế. IMF dự báo tăng trưởng toàn cầu sẽ giảm còn 2,9% vào năm 2024. Tại Đông Nam Á, tăng trưởng các nền kinh tế mới nổi chỉ đạt 4,2% năm 2023 và 4,6% năm 2024, giảm 0,3% so với dự báo trước¹. Trong bối cảnh này, các doanh nghiệp lớn đóng vai trò quan trọng nhờ quy mô, tiềm lực tài chính và hiểu biết thị trường.

Các doanh nghiệp lớn là yếu tố kinh tế then chốt, góp phần nâng cao thu nhập và năng suất lao động (Idson & Oi, 1999), đồng thời ảnh hưởng đến doanh nghiệp nhỏ trong chuỗi giá trị thông qua sức mạnh thị trường (Kaplan, 1950). Ở các quốc gia đang phát triển, doanh nghiệp lớn niềm yết còn thúc đẩy tăng trưởng và giảm nghèo nhờ tiên phong trong đổi mới công nghệ và chuẩn quốc tế.

Phân loại doanh nghiệp là chủ đề nghiên cứu được quan tâm từ lâu, bắt đầu với những nghiên cứu của Beaver (Beaver, 1966) và Altman (Altman, 1968). Từ đó, các mô hình phân loại ngày càng phát triển với nhiều phương pháp phân tích dữ liệu, nhưng vẫn tập trung chủ yếu vào phân loại nhị phân, ví dụ như doanh nghiệp phá sản hoặc không phá sản.

Trong nghiên cứu này, phương pháp phân tích phân cụm (Cluster Analysis) được sử dụng để phân nhóm doanh nghiệp. Thêm vào đó, dữ liệu đa hợp (Compositional Data – CoDA) được sử dụng để phân loại các công ty thay cho việc sử dụng trực tiếp các tỷ số tài chính. Bài nghiên cứu kỳ vọng sẽ phân loại được doanh nghiệp với số lượng từ 03 phân cụm trở lên. Với việc phân cụm các doanh nghiệp, nhóm nghiên cứu đồng thời đặt mục tiêu phát hiện đặc điểm giữa các phân cụm khác nhau và tìm hiểu sự chuyển dịch của các doanh nghiệp giữa các phân cụm.

Tại Đông Nam Á, Indonesia và Thái Lan là các quốc gia có GDP lớn thứ nhất và thứ hai trong khối ASEAN, trong khi đó, Việt Nam có tốc độ tăng trưởng GDP nhanh nhất trong 10 năm (2013–2023)². Nghiên cứu này tập trung vào nhóm 100 công ty hàng đầu tại ba quốc gia nêu trên, với tiêu chí lựa chọn dựa trên tổng giá trị tài sản. Do đó, mục tiêu của nghiên cứu là phân tích hiệu quả hoạt động của các công ty lớn ở những quốc gia này trong giai đoạn trước, trong và sau đại dịch.

¹ Tham khảo từ IMF: <https://www.imf.org/en/Publications/WEO/Issues/2023/10/10/world-economic-outlook-october-2023>

² Tham khảo từ IMF: <https://www.imf.org/en/Publications/WEO/Issues/2023/10/10/world-economic-outlook-october-2023>

2. Tổng quan nghiên cứu

2.1. Phân loại công ty

Phần lớn các nền kinh tế phát triển đã bắt đầu phân loại hoặc phân cụm các doanh nghiệp trong thế kỷ 21. Các phương pháp phân loại được chia thành hai nhóm: (1) dựa trên các khái niệm phổ biến trước khi máy tính ra đời, và (2) dựa trên phân tích phân cụm định lượng (Bailey, 1994). Nhiều nhà khoa học từ nhiều lĩnh vực đã ghi nhận lợi ích của phân loại trong các nghiên cứu của họ trong suốt quá trình chuyển đổi giữa các phương pháp (Linares-Mustarós và cộng sự, 2018). Phân loại các công ty thành các nhóm riêng biệt có hỗ trợ rất nhiều trong việc phân tích chi tiết thuộc tính của doanh nghiệp (Rich, 1992). Bên cạnh đó, phân loại tổ chức cũng được coi là cơ sở để dự đoán các quyết định hoặc thay đổi của tổ chức và cho phép các nhà nghiên cứu dễ dàng xác định phạm vi mà mẫu tổ chức có thể được sử dụng (Carper & Snizek, 1980).

Nhiều nghiên cứu thực nghiệm đã đề cập đến những lợi ích của việc phân loại doanh nghiệp đối với sự thành công của họ, chẳng hạn như: việc thu hút nguồn nhân lực có chuyên môn đặc biệt (McCann & Folta, 2008), tiếp cận các yếu tố đầu vào sản xuất chuyên biệt. Do đó, cả giới học thuật lẫn quản trị doanh nghiệp đang ngày càng chú ý đến những lợi ích thực tế của việc phân cụm doanh nghiệp.

2.2. Các yếu tố tác động đến việc phân loại doanh nghiệp

Một trong những nghiên cứu đầu tiên về phân loại doanh nghiệp có thể kể đến là sự xuất hiện của chỉ số Altman's Z-Score năm 1968 (Altman, 1968). Nghiên cứu của Altman và cộng sự (2017) đã lựa chọn các chỉ số tài chính khác nhau để thể hiện tính thanh khoản, khả năng trả nợ, hiệu quả vận hành và tỷ suất sinh lời của doanh nghiệp.

2.2.1. Tính thanh khoản

Tính thanh khoản là một yếu tố phổ biến trong nhiều nghiên cứu liên quan đến mô hình phân loại. Các công ty trải qua thua lỗ trong hoạt động cũng sẽ dẫn bị thu hẹp tài sản ngắn hạn (Altman, 1968). Đặc điểm thanh khoản cũng thể hiện khả năng trả nợ của một công ty trong ngắn hạn (Chen, 2011). Nhìn chung, các tỷ lệ thanh khoản đã được sử dụng rộng rãi trong việc phát triển mô hình dự đoán tình trạng khó khăn tài chính (Chen, 2011; Mselmi và cộng sự, 2017).

2.2.2. Cấu trúc vốn và khả năng trả nợ

Nghiên cứu thực nghiệm từ các doanh nghiệp “sống thực vật” tại Nhật Bản đã chỉ ra rằng các ngân hàng vẫn có thể tăng cường cho vay đối với các công ty hoạt động kém hiệu quả, ngay cả khi các khoản vay này khó có khả năng sinh lời cho các ngân hàng (Peek & Rosengren, 2005). Khi công ty có nguy cơ phá sản cao, họ không còn khả năng đáp ứng các nghĩa vụ nợ nữa. Trong nghiên cứu của Laitinen và Laitinen (2000), tỷ lệ vốn chủ sở hữu trên tổng tài sản là biến quan trọng nhất trong việc dự đoán ba năm trước khi phá sản (Laitinen & Laitinen, 2000). Hơn nữa, các tỷ lệ thanh toán nợ vẫn là các biến dự đoán hàng đầu cho mô hình dự đoán đa lớp (Cortés và cộng sự, 2007; Jones & Wang, 2019).

2.2.3. Tỷ suất sinh lời

Khi nói đến phân loại công ty, yếu tố lợi nhuận thường được nhấn mạnh. Mức lợi nhuận thấp là một trong những chỉ số có thể dự báo sớm cho tình trạng của doanh nghiệp (Chen, 2011; Cortés và

cộng sự, 2007; Mselmi và cộng sự, 2017). Trong nghiên cứu của Mselmi và cộng sự (2017), các tỷ lệ lợi nhuận là một trong những yếu tố quyết định có thể phân biệt công ty khó khăn hai năm trước khi thất bại. Ngoài ra, khi công ty ngày càng tiến gần đến tình trạng thất bại, mức độ thua lỗ sẽ càng tăng lên.

2.3. Các hệ số tài chính dạng dữ liệu đa hợp

Dữ liệu đa hợp là một dạng dữ liệu mà các thành phần tỷ lệ với nhau và bị ràng buộc bởi một tổng cố định (bằng 1 hoặc 100%). Do dữ liệu đa hợp mô tả sự phân chia của một tổng thể, mỗi thành phần trong nó thể hiện phần đóng góp của một nhóm hay yếu tố, do đó dữ liệu đa hợp không nhận giá trị âm (Greenacre, 2018). Trong lĩnh vực tài chính, dữ liệu đa hợp ngày càng được sử dụng phổ biến, có thể kể đến các nghiên cứu về thị trường tài chính (Kokoszka và cộng sự, 2019) và danh mục đầu tư (Glassman & Riddick, 1996). Một số nghiên cứu đã sử dụng phương pháp này thay thế cho các chỉ số tài chính tiêu chuẩn để phân tích báo cáo tài chính (Coenders & Serrat, 2023; Linares-Mustarós và cộng sự, 2022) và từ đó giúp xác định đặc tính của các công ty.

Việc sử dụng các hệ số tài chính được bắt đầu áp dụng vào cuối thế kỷ 19 và được sử dụng để đánh giá hiệu quả quản lý, chấm điểm tín dụng và đánh giá chiến lược (Altman, 1968). Các hệ số tài chính tiêu chuẩn đã chứng minh được tác dụng trong việc đánh giá hiệu quả tài chính của doanh nghiệp. Tuy nhiên, độ tin cậy của các chỉ số này đã được tính toán và sử dụng không đúng cách, dẫn đến kết quả chẩn đoán sai tình trạng tài chính của doanh nghiệp (Lev & Sunder, 1979).

Việc sử dụng các hệ số tài chính tiêu chuẩn bộc lộ một số hạn chế. Trong phân tích dữ liệu, các số liệu kế toán thường bị ràng buộc bởi một tổng cố định (ví dụ tổng tài sản luôn bằng tổng vốn chủ sở hữu cộng với nợ phải trả). Các hệ số tài chính tiêu chuẩn không xử lý trực tiếp được ràng buộc này và có thể dẫn đến sai sót trong phân tích nếu các thành phần biến đổi đồng thời. Dữ liệu đa hợp khắc phục vấn đề thông qua sử dụng tỷ lệ log-ratio, cho phép phân tích sự phân bố giữa các thành phần.

Sử dụng hệ số tài chính đã được biến đổi từ dữ liệu đa hợp có khả năng giảm thiểu tác động của giá trị cực đoan và dữ liệu phân tán, giúp phân tích ổn định hơn (Linares-Mustarós và cộng sự, 2022). Dữ liệu đa hợp mang tính chất bất biến theo quy mô, tập trung vào mối quan hệ tỷ lệ giữa các thành phần mà không phụ thuộc vào quy mô tuyệt đối.

Việc sử dụng dữ liệu đa hợp yêu cầu các số liệu kế toán trong báo cáo tài chính không nhận giá trị âm và không chồng chéo (Coenders & Serrat, 2023). Trong báo cáo tài chính, số liệu kế toán có thể âm vì chúng là kết quả của việc trừ các số liệu kế toán lẫn nhau. Ví dụ như các chỉ tiêu kế toán như tỷ lệ tài sản, vốn chủ sở hữu, doanh thu, nợ vay không thể mang giá trị âm trong cấu trúc tỷ lệ, vì chúng đại diện cho phần đóng góp tích cực vào tổng thể. Không chồng chéo nghĩa là các thành phần của dữ liệu đa hợp không được lặp lại hoặc tính trùng nhau, để đảm bảo rằng chúng phản ánh đúng tỷ lệ của từng thành phần riêng lẻ trong tổng thể. Nếu ta đang phân tích tỷ lệ tài sản, nợ phải trả, và vốn chủ sở hữu, thì ba thành phần này cần phản ánh từng phần riêng biệt của tổng tài sản, không được tính trùng nhau. Các số liệu kế toán có thể được chuyển đổi dưới dạng tỷ lệ sử dụng biến đổi logarit trung tâm (Aitchison, 1983) hoặc biến đổi logarit theo cặp (Creixans-Tenas và cộng sự, 2019; Greenacre, 2019).

3. Phương pháp nghiên cứu

3.1. Số liệu nghiên cứu

Dữ liệu được sử dụng trong bài nghiên cứu này đến từ nguồn cơ sở dữ liệu ORBIS³. Các công ty được lựa chọn đến từ Việt Nam, Thái Lan và Indonesia trong khoảng thời gian 5 năm tính từ năm 2018 đến năm 2022. Cơ sở dữ liệu được thu thập bao gồm: Thông tin chung của các công ty như: ID, mã ISIN (nếu có), trạng thái, quốc gia, loại; và thông tin tài chính gồm: doanh thu, lợi nhuận ròng, lợi nhuận trước thuế, tổng tài sản, tài sản ngắn hạn, nợ dài hạn, nợ ngắn hạn, giá vốn hàng hóa, giá trị doanh nghiệp và lợi nhuận từ hoạt động kinh doanh.

Tiêu chí lựa chọn yêu cầu các công ty phải có sẵn tất cả các thông tin tài chính nêu trên trong các năm từ năm 2018 đến năm 2022, loại trừ các công ty có giá trị vốn chủ sở hữu âm. Mẫu dữ liệu được thu thập từ cơ sở dữ liệu ORBIS bao gồm 730 công ty từ Việt Nam, 1.292 công ty từ Thái Lan và 212 công ty từ Indonesia. Để đảm bảo tính cân bằng trong dữ liệu giữa ba quốc gia, mỗi quốc gia lọc ra 100 công ty có tổng giá trị tài sản cao nhất. Dữ liệu cuối cùng bao gồm 300 công ty từ ba quốc gia với tổng giá trị tài sản lớn nhất.

3.2. Lựa chọn biến số

Dựa theo mục tiêu nghiên cứu đã đề ra, các chỉ số (tỷ số) tài chính liên quan đến khả năng thanh toán và các tỷ lệ thành phần trong phân tích DuPont được lựa chọn. Như đã trình bày ở phần cơ sở lý thuyết, các dữ liệu thông tin tài chính được thu thập dựa theo các yếu tố có tác động đến việc phân loại doanh nghiệp, bao gồm tính thanh khoản, cấu trúc nợ và khả năng sinh lời. Các chỉ số tài chính được tính toán là các chỉ số thông dụng đã được sử dụng trong rất nhiều nghiên cứu để đại diện cho các nhóm đặc điểm này (du Jardin và cộng sự, 2019).

Các tỷ lệ này được tính từ dữ liệu thông tin kế toán thu thập được. Đầu tiên, các biến cho thông tin kế toán được gán thành các biến x_j , cụ thể như sau:

$$x_1 = \text{tài sản dài hạn} (= \text{Tổng tài sản} - \text{Tài sản ngắn hạn})$$

$$x_2 = \text{Tài sản ngắn hạn}$$

$$x_3 = \text{Nợ dài hạn}$$

$$x_4 = \text{Nợ ngắn hạn}$$

$$x_5 = \text{Doanh thu}$$

$$x_6 = \text{Chi phí} (= \text{Doanh thu} - \text{Lợi nhuận ròng})$$

Tiếp theo đó, các chỉ số tài chính tiêu chuẩn sẽ được tính từ các biến x được xác định ở trên. Bảng dưới đây cung cấp giải thích cho ý nghĩa của từng chỉ số.

³ Nguồn cơ sở dữ liệu ORBIS là cung cấp dữ liệu tài chính của gần 500 triệu tổ chức và doanh nghiệp toàn cầu, được quản lý bởi Bureau van Dijk, hiện là một công ty thuộc Moody's Analytics.

Bảng 1.

Giải thích các biến

Chỉ số tài chính	Phân loại	Công thức
Khả năng thanh toán hiện hành	Khả năng thanh toán	$\frac{x_2}{x_4}$
Tỷ số nợ dài hạn trên tổng nợ		$\frac{x_3}{x_3 + x_4}$
Cấu trúc vốn (Tổng nợ trên tổng tài sản)		$\frac{x_3 + x_4}{x_1 + x_2}$
Tỷ lệ vòng quay tổng tài sản	Phân tích của DuPont	$\frac{x_5}{x_1 + x_2}$
Biên lợi nhuận ròng		$\frac{x_5 - x_6}{x_5}$
Tỷ suất lợi nhuận trên tổng tài sản (ROA)		$\frac{x_5 - x_6}{x_1 + x_2}$
Hệ số đòn bẩy tài chính		$\frac{x_1 + x_2}{(x_1 + x_2) - (x_3 + x_4)}$
Tỷ suất lợi nhuận trên vốn chủ sở hữu (ROE)		$\frac{x_5 - x_6}{(x_1 + x_2) - (x_3 + x_4)}$

3.3. Việc sử dụng phương pháp phân tích dữ liệu đa hợp (CoDA)

Việc sử dụng phương pháp CoDA để chuyển đổi thông tin tài chính và các bước trong phân tích phân cụm của dữ liệu sẽ được mô tả tóm lược ở phần dưới đây:

3.3.1. Phân tích CoDA

Vector đa hợp được định nghĩa là một vector với D thành phần có giá trị dương, trong đó các phần thông tin được thể hiện bởi tỷ lệ tương đối giữa các thành phần này. Aitchison (1986) đã giới thiệu các nền tảng về độ đo Aitchison để xử lý các vector đa hợp.

$(x_1, x_2, \dots, x_D)$ với x_j là các chỉ số tài chính có giá trị dương, $j = 1, 2, \dots, D$.

Liên hệ với phân tích báo cáo tài chính, các thành phần nêu trên bao gồm các số liệu kế toán được trích xuất từ báo cáo tài chính. Đưa số liệu báo cáo tài chính vào các thành phần trong vector đòi hỏi phải loại trừ các số liệu âm (Creixans-Tenas và cộng sự, 2019). Sự tồn tại của các số liệu âm này có thể dẫn đến sự gián đoạn, các yếu tố ngoại lai hoặc đảo ngược việc nhận định phân tích, đặc biệt là khi số liệu kế toán âm được tìm thấy trong mẫu số (Lev & Sunder, 1979).

3.3.2. Chuyển đổi dữ liệu đa hợp sử dụng phép chuyển đổi logarit cộng tính và logarit trung tâm có trọng số

Do các thành phần trong vector đa hợp luôn có tổng bằng 1 (hoặc 100%), việc phân tích trực tiếp các thành phần có thể gây ra phụ thuộc giả. Để thuận tiện hơn trong việc xử lý và phân tích dữ liệu, CoDA sử dụng các phép biến đổi logarit để xử lý vector đa hợp, như phương pháp đã được giới thiệu

bởi Aitchison vào năm 1986. Trong nghiên cứu này, tỷ lệ logarit cộng tính (ALR) và tỷ lệ logarit trung tâm (CLR) được chọn để sử dụng. Sau khi thực hiện các phép chuyển đổi log-ratio, vector đa hợp ban đầu sẽ được đưa vào không gian Euclid mà ở đó các phương pháp thống kê thông thường (như hồi quy, phân cụm, phân tích thành phần chính – PCA) có thể được áp dụng một cách chính xác và không tạo ra các mối phụ thuộc giả.

Đối với một tập hợp các D loại tài khoản kế toán thu được từ báo cáo tài chính (các chỉ số tài chính được gán là x_j), phép biến đổi ALR được biểu diễn như sau:

$$alr(x) = (alr_1, alr_2, \dots, alr_{D-1}) = (\log \frac{x_1}{x_D}, \dots, \log \frac{x_{D-1}}{x_D}) \quad (1)$$

Phép biến đổi logarit cộng tính (ALR) chuyển đổi các biểu diễn trong đơn hình thành biểu diễn trong không gian vector thật. ALR chuyển đổi dữ liệu đa hợp bằng cách chọn một thành phần chuẩn và tính log-ratio giữa các thành phần còn lại với thành phần chuẩn này. Như đã thấy từ công thức biểu diễn, vector ALR sau phép biến đổi sẽ có D-1 phần tử.

Với một tập hợp gồm D loại tài khoản kế toán tương tự, phép biến đổi logarit trung tâm (CLR) có thể được biểu diễn như sau:

$$clr(x) = (clr_1, clr_2, \dots, clr_D) = \log \left(\frac{x_j}{\frac{1}{x_1^D} \frac{1}{x_2^D} \dots \frac{1}{x_D^D}} \right) = \log \left(\frac{x_j}{\sqrt[D]{x_1 \dots x_D}} \right) \quad (2)$$

Trong đó, $j = 1, \dots, D$

Tương tự như ALR, phép biến đổi CLR chuyển đổi biểu diễn đơn hình thành biểu diễn không gian vector. CLR chuyển đổi tất cả các thành phần trong vector đa hợp bằng cách so sánh với trung bình hình học của toàn bộ các thành phần. Vector CLR sau phép biến đổi sẽ có D phần tử. Sau khi biến đổi CLR, khoảng cách Euclid giữa 2 vector trong không gian log-ratio phản ánh đúng sự khác biệt tương đối giữa các thành phần của hai vector đa hợp (hay còn gọi là khoảng cách Aitchison).

Với thuộc tính bảo toàn khoảng cách, CLR thường được ưu tiên sử dụng hơn ALR trong các phân tích phân cụm. Tuy nhiên, trong các ứng dụng thông thường của phân tích dữ liệu đa hợp, một số nhóm dữ liệu tài chính nhất định có giá trị thường thấp có thể bộc lộ ra giá trị phương sai CLR rất lớn, gây ảnh hưởng đáng kể đến kết quả phân loại (Greenacre & Lewi, 2009; Jofre-Campuzano & Coenders, 2022). Điểm hạn chế này có thể được giải quyết bằng cách tính toán một biến thể có trọng số của phép biến đổi CLR, trong đó trọng số (w_j), với tổng tất cả trọng số bằng 1, tỷ lệ thuận với giá trị trung bình của x_j :

$$wclr_j = \sqrt{w_j} \log \log \left(\frac{x_j}{x_1^{w_1} x_2^{w_2} \dots x_D^{w_D}} \right) \quad (3)$$

3.3.3. Các bước phân tích phân cụm sử dụng phương pháp phân cụm WARD

Quy trình thực hiện phân tích phân cụm báo cáo tài chính có trọng số được thực hiện theo phương pháp tiếp cận của Jofre-Campuzano và Coenders (2022), bao gồm các bước như sau:

- *Bước 1:* Trước khi thực hiện biến đổi logarit (ALR hay CLR), cần phải thực hiện gán các giá trị bằng 0 (Martín-Fernández và cộng sự, 2011). Trong bước này, hàm `lrEM` trong gói `zCompositions` của R đã được sử dụng để thực hiện thuật toán EM (Palarea-Albaladejo & Martín-Fernández, 2008).

- *Bước 2:* Hàm ALR và CLR trong gói `easyCODA` của R đã được sử dụng để thực hiện phép biến đổi ALR và CLR có trọng số.

- *Bước 3:* Phương pháp phân cụm WARD đã được áp dụng bằng cách sử dụng hàm WARD.

Phân cụm WARD là một phương pháp tiêu chuẩn thường được áp dụng cho dữ liệu báo cáo tài chính (Jofre-Campuzano & Coenders, 2022; Linares-Mustarós và cộng sự, 2018; Martín-Oliver và cộng sự, 2017). Phương pháp WARD là một trong những phương pháp phân cụm phân cấp (Hierarchical Clustering), cụ thể hơn là nhóm phân cụm phân cấp kết hợp (Agglomerative Hierarchical Clustering). Đặc điểm nổi bật của phương pháp này nằm ở cách thức đo lường khoảng cách giữa các cụm và tiêu chí gộp các cụm sao cho tối ưu. Phương pháp WARD ưu tiên việc giảm thiểu sự khác biệt bên trong mỗi cụm, dẫn đến việc tạo ra các cụm đồng nhất, với các điểm dữ liệu trong cụm có xu hướng gần nhau về giá trị. Một trong những ưu điểm của phương pháp này là sự phù hợp với nhóm dữ liệu nhỏ hoặc trung bình.

- *Bước 4:* Phân tích kết quả phân cụm.

Để giải thích đặc điểm tài chính đặc trưng trong mỗi cụm, các giá trị đa hợp trung tâm trong mỗi cụm sẽ được tính toán, từ đó đưa ra liên hệ với các tỷ lệ tài chính tiêu chuẩn được sử dụng từ đầu (xem Bảng 1) (Saus-Sala và cộng sự, 2021). Ngoài ra, các biểu đồ khảm (Mosaic Plot) và biểu đồ khoảng tin cậy được sử dụng để kiểm tra tính liên quan của các cụm đến các biến phi tài chính (Goldstein & Healy, 1995). Nghiên cứu này cũng đồng thời đánh giá động lực giữa các cụm thông qua sử dụng ma trận chuyển tiếp.

4. Kết quả nghiên cứu và thảo luận

4.1. Thống kê mô tả

Bảng 2 cung cấp thông tin mô tả bộ dữ liệu gồm 300 doanh nghiệp từ Việt Nam, Thái Lan và Indonesia trong giai đoạn 2018–2022. Bảng chia thành hai phần. Phần đầu gồm 6 biến đầu vào (x_1 – x_6) với giá trị trung bình và độ lệch chuẩn, phản ánh sự biến động lớn do độ lệch chuẩn cao hơn nhiều so với trung bình. Phần hai gồm 8 biến tỷ số tài chính tính từ 6 biến đầu, cũng có giá trị trung bình và độ lệch chuẩn. Độ lệch chuẩn tương đương trung bình cho thấy mức độ phân tán tương đối. Các tỷ số thanh toán ổn định, thể hiện khả năng đáp ứng nghĩa vụ nợ dài hạn tốt. Đặc biệt, tỷ số tài chính ổn định giai đoạn 2018–2021 và năm 2022 đạt hiệu suất tốt nhất.

Bảng 2.

Thống kê mô tả dữ liệu

Năm	2018	2019	2020	2021	2022
Tài sản dài hạn (tỉ USD)	1,51 (3,43)	1,74 (4,20)	1,98 (4,58)	2,16 (5,05)	2,18 (5,14)
Tài sản ngắn hạn (tỉ USD)	0,80 (1,80)	0,85 (1,73)	0,86 (1,73)	0,99 (1,87)	1,09 (2,20)
Nợ dài hạn (tỉ USD)	0,68 (1,56)	0,80 (2,01)	0,99 (2,50)	1,05 (2,65)	1,07 (2,85)
Nợ ngắn hạn (tỉ USD)	0,58 (1,11)	0,65 (1,23)	0,68 (1,27)	0,76 (1,37)	0,78 (1,59)

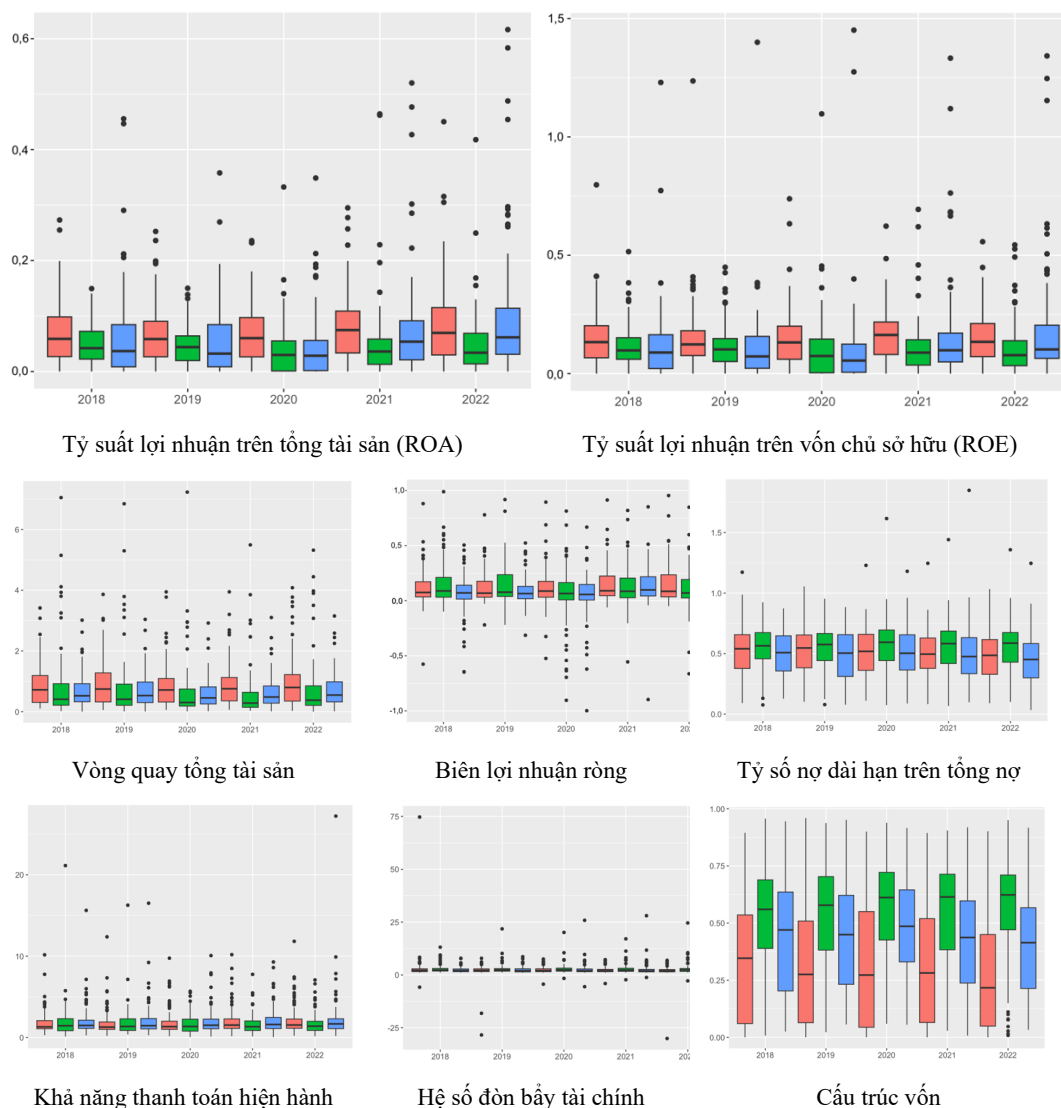
Năm	2018	2019	2020	2021	2022
Doanh thu (tỉ USD)	1,70 (4,89)	1,80 (5,02)	1,56 (3,93)	1,77 (4,65)	2,25 (6,54)
Chi phí (tỉ USD)	1,58 (4,57)	1,68 (4,77)	1,50 (3,81)	1,64 (4,35)	2,08 (6,30)
Tỷ số nợ dài hạn trên tổng nợ	0,43 (0,26)	0,43 (0,26)	0,45 (0,26)	0,43 (0,26)	0,42 (0,26)
Cơ cấu vốn	0,52 (0,20)	0,52 (0,20)	0,53 (0,21)	0,52 (0,22)	0,50 (0,21)
Hệ số đòn bẩy tài chính	2,79 (4,46)	2,41 (2,76)	2,63 (2,29)	2,60 (2,32)	2,37 (2,77)
Tỷ lệ vòng quay tổng tài sản	0,80 (0,83)	0,80 (0,82)	0,70 (0,75)	0,71 (0,71)	0,80 (0,82)
Biên lợi nhuận ròng	0,18 (0,50)	0,16 (0,31)	0,15 (0,36)	0,19 (0,47)	3,08 (50,42)
Khả năng thanh toán hiện hành	1,89 (1,88)	1,87 (1,84)	1,83 (1,47)	1,90 (1,52)	2,03 (2,13)
Tỷ suất lợi nhuận trên tổng tài sản (ROA)	0,06 (0,06)	0,06 (0,05)	0,05 (0,06)	0,07 (0,08)	0,08 (0,09)
Tỷ suất lợi nhuận trên vốn chủ sở hữu (ROE)	0,13 (0,13)	0,13 (0,13)	0,12 (0,16)	0,15 (0,16)	0,15 (0,16)

Ghi chú: Giá trị trung bình và độ lệch chuẩn (trong ngoặc) của các biến đầu vào và tỷ số tài chính.

Phân tích Biểu đồ 1 (Boxplot) cho 8 tỷ số tài chính chỉ ra tất cả các tỷ số đều có giá trị ngoại lệ, với các tỷ số như: Khả năng thanh khoản hiện hành, cơ cấu vốn, biên lợi nhuận ròng, tỷ lệ vòng quay tổng tài sản, ROA và ROE có giá trị ngoại lệ dương. Điều này phản ánh sự biến động lớn và chênh lệch trong hiệu suất tài chính giữa doanh nghiệp tại Việt Nam, Thái Lan và Indonesia.

Các tỷ số đều bất đối xứng, cho thấy sự khác biệt rõ rệt trong phân phối dữ liệu. Điều này nhấn mạnh nhu cầu cần một hệ thống phân loại để mang lại cái nhìn rõ hơn cho nhà đầu tư và người làm chính sách.

Phân loại doanh nghiệp dựa trên tính thanh khoản và lợi nhuận sẽ giúp các bên liên quan có cách tiếp cận chiến lược hơn, từ đó hỗ trợ ra quyết định và xây dựng chính sách hiệu quả hơn.

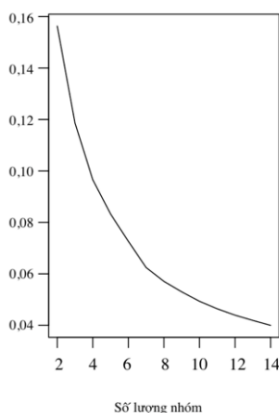
**Biểu đồ 1.** Boxplot các chỉ số tài chính

Bảng 3 mô tả sự thay đổi trong trọng số của nợ dài hạn và tài sản dài hạn, 2 biến có trọng số lớn nhất so với các biến còn lại. Sử dụng phương pháp ALR, khi lấy chi phí làm chuẩn, trọng số thu được cho thấy có sự thay đổi rõ rệt. Trong đó, nợ dài hạn và tài sản dài hạn vẫn giữ vai trò quan trọng nhất, với trọng số lớn nhất trong số các biến được phân tích. Biến tài sản dài hạn đã có sự cải thiện đáng kể về trọng số, phản ánh tầm quan trọng ngày càng tăng của nó trong mô hình đánh giá chi phí.

Bảng 3.

Trọng số các biến đóng góp trong phương sai tổng (%)

	Không trọng số	Có trọng số
Tài sản dài hạn/Chi phí	0,19	0,34
Tài sản ngắn hạn/Chi phí	0,10	0,11
Nợ dài hạn/Chi phí	0,62	0,46
Nợ ngắn hạn/Chi phí	0,08	0,06
Doanh thu/Chi phí	0,02	0,03

4.2. Phân loại có trọng số**Biểu đồ 2.** Đồ thị Scree Plot của các tổng bình phương trong cụm

Dựa trên thông tin thể hiện trong Biểu đồ 2, cây phân cấp hiển thị hai cụm lớn và một cụm thứ ba rất nhỏ. Ngoài ra, biểu đồ Scree cho thấy một điểm uốn ở 3, cho thấy có sự phân tách thành 3 cụm để tiếp tục phân tích chi tiết.

Bảng 4 tổng hợp các giá trị trung bình từ dữ liệu của 1.500 quan sát trong khoảng thời gian 5 năm, mỗi năm thu thập từ 300 công ty. Số liệu tổng hợp cho ta thấy một cái nhìn toàn diện về tình hình tài chính của các công ty trong suốt giai đoạn này. Phân tích qua nhiều năm và nhiều công ty khác nhau giúp đảm bảo tính đại diện và độ chính xác của các kết quả thu được, đồng thời phản ánh được xu hướng tài chính.

Phần lớn các quan sát rơi vào nhóm 1 và nhóm 2, với hơn 95% theo cả hai phương pháp ALR và CLR. Nhóm 1 chiếm hơn 50%, trong khi nhóm 2 đóng góp hơn 44%. 8 chỉ số tài chính được chia thành hai nhóm chính: (1) Khả năng thanh toán và thanh khoản, và (2) hiệu quả hoạt động. Nhóm 3 luôn chiếm ưu thế về hiệu quả hoạt động. Chỉ số thanh khoản và thanh toán của nhóm 3 cũng cho thấy mức rủi ro thấp hơn so với nhóm 1 và 2, trong khi hiệu quả hoạt động của nhóm này vượt trội đáng kể. Điều này được thể hiện qua tỷ lệ vòng quay tài sản cao hơn, dẫn đến ROE ở mức 21–25% cho nhóm 3, so với chỉ 10–11% ở nhóm 1. Kết quả này nhấn mạnh vai trò quan trọng của vòng quay tài sản trong việc tối ưu hóa lợi nhuận trên vốn chủ sở hữu.

Phương pháp nghiên cứu đạt hiệu quả trong khả năng phân loại các công ty theo mức độ rủi ro và hiệu quả hoạt động. Ba nhóm được phân tách rõ ràng, đặc biệt là sự khác biệt giữa nhóm tốt nhất và yếu nhất. Các công ty có rủi ro cao về thanh toán và thanh khoản thường có hiệu quả hoạt động kém, trong khi những công ty có rủi ro thấp lại đạt được hiệu quả hoạt động vượt trội. Điều này cho thấy mối quan hệ chặt chẽ giữa quản lý rủi ro và hiệu quả tài chính trong các công ty.

Bảng 4.

Đặc điểm các nhóm đã được phân tách

Các tỷ số tài chính	CLR			ALR		
	Nhóm 1	Nhóm 2	Nhóm 3	Nhóm 1	Nhóm 2	Nhóm 3
	n=796	n=680	n=24	n=750	n=664	n=86
Khả năng thanh toán hiện hành	1,77	2,04	1,86	1,94	1,88	1,83
Tỷ số nợ dài hạn trên tổng nợ	0,60	0,25	0,00	0,59	0,30	0,02
Cơ cấu vốn	0,58	0,46	0,44	0,56	0,48	0,49
Tỷ lệ vòng quay tổng tài sản	0,32	1,25	1,69	0,28	1,09	2,43
Biên lợi nhuận ròng	0,16	0,09	0,11	0,18	0,08	0,06
Tỷ suất lợi nhuận trên tổng tài sản (ROA)	0,04	0,09	0,14	0,04	0,08	0,10
Hệ số đòn bẩy tài chính	3,02	2,10	2,02	2,91	2,18	2,40
Tỷ suất lợi nhuận trên vốn chủ sở hữu (ROE)	0,10	0,16	0,25	0,11	0,15	0,21

Bảng 5.

Giá trị p-value của các kiểm định sự khác biệt về giá trị các biến trong các cụm được phân loại

Các cặp	Phương pháp CLR			Phương pháp ALR		
	1-2	1-3	2-3	1-2	1-3	2-3
Khả năng thanh toán hiện hành	0,0054	0,7346	0,3947	0,5505	0,3690	0,6607
Tỷ số nợ dài hạn trên tổng nợ	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Cơ cấu vốn	0,0000	0,0014	0,5751	0,0000	0,0055	0,5500
Tỷ lệ vòng quay tổng tài sản	0,0000	0,0000	0,0229	0,0000	0,0000	0,0000
Biên lợi nhuận ròng	0,0000	0,0089	0,4031	0,0000	0,0000	0,0022
Tỷ suất lợi nhuận trên tổng tài sản (ROA)	0,0000	0,0000	0,0124	0,0000	0,0000	0,0075
Hệ số đòn bẩy tài chính	0,0000	0,0001	0,6415	0,0000	0,0100	0,1805
Tỷ suất lợi nhuận trên vốn chủ sở hữu (ROE)	0,0000	0,0000	0,0058	0,0000	0,0000	0,0002

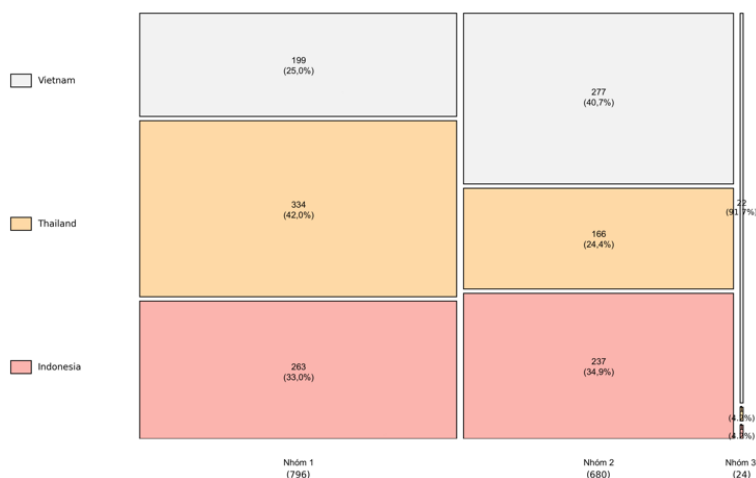
Ghi chú: Các giá trị in đậm thể hiện kết quả không có ý nghĩa thống kê ở mức 5%.

Bảng 5 trình bày kết quả kiểm định ý nghĩa thống kê của từng cặp nhóm (1-2, 1-3, 2-3) theo hai phương pháp ALR và CLR. Kết quả cho thấy, đối với cả hai phương pháp, sự khác biệt giữa nhóm 1 so với nhóm 2 và nhóm 3 đều đạt mức ý nghĩa thống kê. Điều này phản ánh tính độc lập và sự hợp lý trong việc phân loại nhóm 1, đồng thời khẳng định rằng nhóm này có hiệu quả thấp hơn hai nhóm còn lại trên cả hai phương diện: sử dụng đòn bẩy tài chính và kết quả kinh doanh.

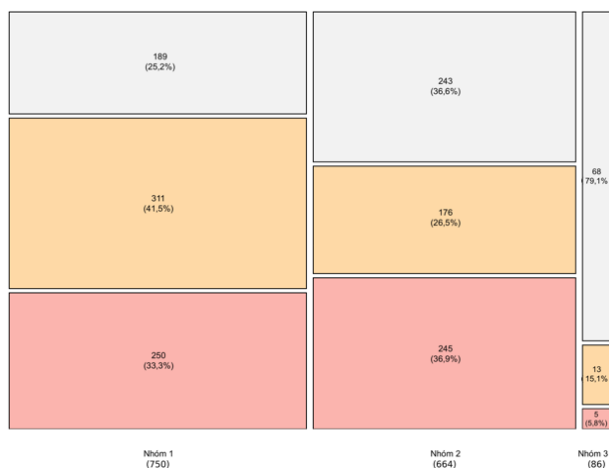
Một điểm đáng chú ý là trong cả hai phương pháp ALR và CLR, các biến chỉ số liên quan đến khả năng thanh toán, cơ cấu vốn và hệ số đòn bẩy tài chính đều không đạt mức ý nghĩa thống kê ($p\text{-value} > 5\%$). Điều này cho thấy sự phân cụm không thực sự có ý nghĩa nếu dựa trên các chỉ số này. Tuy nhiên, xét về mặt lý thuyết, đây đều là những chỉ số liên quan đến rủi ro tài chính, do đó có thể kết luận rằng phương pháp phân cụm theo yếu tố rủi ro tài chính (bao gồm các tỷ số thanh khoản và đòn bẩy tài chính) hiện chưa mang lại hiệu quả rõ rệt.

Ngược lại, các chỉ số phản ánh kết quả kinh doanh đều đạt mức ý nghĩa thống kê cao ($p\text{-value} < 1\%$), cho thấy khả năng phân tách giữa các nhóm là rõ ràng khi sử dụng tiêu chí này. Kết quả này cho thấy phương pháp phân cụm dựa trên hiệu quả kinh doanh có tính khả thi và mang lại sự phân biệt có ý nghĩa thống kê giữa các nhóm.

Biểu đồ Mosaic không chỉ cung cấp cái nhìn rõ nét hơn về tỷ lệ phân loại giữa các nhóm (với nhóm 1 và nhóm 2 chiếm đa số), mà còn cho thấy cấu trúc phân bố theo quốc gia trong từng nhóm. Trong số ba quốc gia được phân tích, Thái Lan chiếm tỷ trọng lớn nhất trong nhóm 1 (trên 40%), tiếp theo là Indonesia với 33,3%, và phần còn lại (khoảng 25%) đến từ Việt Nam. Khi áp dụng cả hai phương pháp CLR và ALR, kết quả cho thấy sự tương đồng. Nhóm 1 có kết quả kém nhất, các công ty Việt Nam chỉ chiếm 25%. Nhóm 3, nhóm được đánh giá là tốt nhất, chủ yếu bao gồm các công ty Việt Nam. Cụ thể, với phương pháp CLR, toàn bộ 24 quan sát đều đến từ Việt Nam; với phương pháp ALR, các quan sát từ Việt Nam chiếm 79%, tương đương 68 trên tổng số 86 quan sát.



Biểu đồ 5. Biểu đồ khảm (Mosaic) cho 3 nhóm giai đoạn 2018–2023 (CLR)



Biểu đồ 6. Biểu đồ khảm (Mosaic) cho 3 nhóm giai đoạn 2018–2023 (ALR)

Bảng 6.

Ma trận chuyển đổi nhóm qua các năm 2018–2023

	2018	2019	2020	2021	2022
Nhóm 1	155 (146)	152 (145)	169 (157)	165 (157)	155 (145)
Nhóm 2	140 (138)	145 (139)	125 (125)	129 (127)	141 (135)
Nhóm 3	5 (16)	3 (16)	6 (18)	6 (16)	4 (20)

Ghi chú: Trong ngoặc đơn () là kết quả của phương pháp ALR.

Ma trận chuyển đổi thể hiện sự biến động đáng kể trong giai đoạn 2020–2021, tuy nhiên, sự biến động này chủ yếu xuất hiện ở nhóm 1 và nhóm 2, với xu hướng chuyển dịch phần lớn từ nhóm 2 xuống nhóm 1. Trong khi đó, nhóm 3 cho thấy sự ổn định qua suốt cả 5 năm. Kết quả năm 2022 cho thấy một sự phục hồi ổn định, trở lại tình trạng tương tự như giai đoạn 2018–2019. Sự suy giảm trong hai năm 2020 và năm 2021 có thể được giải thích bởi tác động của đại dịch COVID-19, khi các doanh nghiệp trong giai đoạn này gặp nhiều khó khăn. Tuy nhiên, các công ty thuộc nhóm 3 (nhóm có các chỉ số tài chính tốt) dường như ít bị ảnh hưởng bởi những biến động này.

Kết quả nghiên cứu cũng cho thấy, với dữ liệu được thu thập và phân tích đến năm 2023, tác động của đại dịch COVID-19 đối với các nhóm doanh nghiệp trong mô hình phân cụm là không rõ rệt. Cả hai phương pháp ALR và CLR đều chỉ ra rằng mức độ biến động trong nhóm 3 – nhóm doanh nghiệp có hiệu quả hoạt động cao nhất – là rất nhỏ, thể hiện sự ổn định tương đối của nhóm này bất chấp ảnh hưởng của đại dịch.

Ngược lại, sự biến động chủ yếu tập trung ở nhóm 2, tức nhóm doanh nghiệp có hiệu quả hoạt động ở mức trung bình. Kết quả phân tích cho thấy một số doanh nghiệp thuộc nhóm này đã gặp khó khăn đáng kể trong giai đoạn hậu COVID-19, dẫn đến suy giảm hiệu quả hoạt động và bị phân loại xuống nhóm 1 – nhóm có hiệu quả thấp nhất. Điều này phản ánh sự phân hóa trong khả năng thích

ứng của doanh nghiệp trước các cú sốc kinh tế, khi những doanh nghiệp có nền tảng tài chính và khả năng quản trị rủi ro tốt (nhóm 3) ít bị ảnh hưởng hơn so với các doanh nghiệp trung bình (nhóm 2), vốn có sức chống chịu kém hơn trước những biến động lớn của nền kinh tế.

Các công ty duy trì vị trí trong nhóm 3 liên tục trong suốt 5 năm trong tổng số hơn 16 công ty qua các năm, chỉ có 50% công ty giữ vững vị trí này trong toàn bộ thời kỳ, 50% còn lại có sự chuyển dịch qua lại giữa nhóm 2 và nhóm 3. Đáng chú ý, trong số 12 công ty duy trì vị trí ổn định, 8 công ty đến từ Việt Nam. Thêm một bằng chứng thống kê cho thấy, Việt Nam chiếm ưu thế rõ rệt trong nhóm 3, nhóm được đánh giá là tốt nhất.

5. Kết luận

Nghiên cứu phân tích dữ liệu của 300 doanh nghiệp tại Việt Nam, Thái Lan và Indonesia giai đoạn 2018–2022, dựa trên 6 biến đầu vào và 8 tỷ số tài chính, gồm 3 tỷ số thanh khoản và 5 tỷ số hiệu quả hoạt động. Dù diễn ra trong bối cảnh COVID-19, đại dịch không ảnh hưởng rõ rệt đến kết quả phân cụm.

Phân tích bằng phương pháp WARD xác định 3 nhóm doanh nghiệp. Nhóm 3 có hiệu quả hoạt động cao nhất, với thanh khoản an toàn, vòng quay tài sản, ROA và ROE vượt trội. Ngược lại, nhóm 1 có hiệu suất và mức độ an toàn tài chính thấp nhất.

Phần lớn doanh nghiệp thuộc nhóm hiệu quả cao đến từ Việt Nam. Ba doanh nghiệp giữ vị trí ổn định trong 5 năm theo phương pháp CLR, và con số này tăng lên 8 với phương pháp ALR, cho thấy khả năng duy trì hiệu suất cao xuyên suốt đại dịch.

Phân cụm WARD chứng minh hiệu quả khi các nhóm có đặc điểm tài chính nhất quán, phản ánh cả tính ổn định và sự thay đổi trong cấu trúc doanh nghiệp do COVID-19.

Hạn chế chính của nghiên cứu là chỉ tập trung vào doanh nghiệp có tổng tài sản lớn do thiếu dữ liệu về vốn hóa. Trong tương lai, nghiên cứu có thể mở rộng về địa lý và tích hợp thêm tiêu chí để phân tích sâu hơn các yếu tố tài chính và kinh tế vĩ mô, từ đó nâng cao tính khái quát và hiểu rõ hơn động lực tài chính trong khu vực.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Bộ Giáo dục và Đào tạo Việt Nam, trong khuôn khổ đề tài nghiên cứu khoa học cấp Bộ, Mã số B2024-NHF-04, theo Quyết định số 1393/QĐ-BGDĐT ngày 16 tháng 5 năm 2023.

Tài liệu tham khảo

- Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70(1), 57-65. <https://doi.org/10.1093/biomet/70.1.57>
- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data (Monographs on Statistics and Applied Probability)* (1st ed.). Springer.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609. <https://doi.org/10.1111/J.1540-6261.1968.TB00843.X>

- Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-Score model. *Journal of International Financial Management & Accounting*, 28(2), 131-171. <https://doi.org/10.1111/JIFM.12053>
- Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71-111.
- Carper, W. B., & Snizek, W. E. (1980). The nature and types of organizational taxonomies: An overview. *The Academy of Management Review*, 5(1), 65-75. <https://doi.org/10.2307/257805>
- Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, 38(9), 11261-11272. <https://doi.org/10.1016/j.eswa.2011.02.173>
- Coenders, G., & Serrat, N. A. (2023). *Accounting statement analysis at industry level. A gentle introduction to the compositional approach*. <https://arxiv.org/abs/2305.16842v5>
- Cortés, A. E., Martínez, M. G., & Rubio, N. G. (2007). Multiclass corporate failure prediction by Adaboost.M1. *International Advances in Economic Research*, 13(3), 301-312. <https://doi.org/10.1007/s11294-007-9090-2>
- Creixans-Tenas, J., Coenders, G., & Arimany-Serrat, N. (2019). Corporate social responsibility and financial profile of Spanish private hospitals. *Heliyon*, 5(10), e02623. <https://doi.org/10.1016/j.heliyon.2019.e02623>
- Bailey, K. (1994). *Typologies and Taxonomies*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412986397>
- du Jardin, P., Veganzones, D., & Séverin, E. (2019). Forecasting corporate bankruptcy using accrual-based models. *Computational Economics*, 54(1), 7-43. <https://doi.org/10.1007/s10614-017-9681-9>
- Glassman, D. A., & Riddick, L. A. (1996). Why empirical international portfolio models fail: evidence that model misspecification creates home asset bias. *Journal of International Money and Finance*, 15(2), 275-312. [https://doi.org/10.1016/0261-5606\(95\)00046-1](https://doi.org/10.1016/0261-5606(95)00046-1)
- Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(1), 175-177. <https://doi.org/10.2307/2983411>
- Greenacre, M. (2018). *Compositional Data Analysis in Practice* (1st ed.). New York: Chapman and Hall/CRC. <https://doi.org/10.1201/9780429455537>
- Greenacre, M. (2019). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*, 51(5), 649-682. <https://doi.org/10.1007/s11004-018-9754-x>
- Greenacre, M., & Lewi, P. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *Journal of Classification*, 26(1), 29-54. <https://doi.org/10.1007/s00357-009-9027-y>
- Idson, T. L., & Oi, W. Y. (1999). Workers are more productive in large firms. *American Economic Review*, 89(2), 104-108. <https://doi.org/10.1257/AER.89.2.104>

- Jofre-Campuzano, P., & Coenders, G. (2022). Compositional classification of financial statement profiles: The weighted case. *Journal of Risk and Financial Management*, 15(12), 546. <https://doi.org/10.3390/JRFM15120546>
- Jones, S., & Wang, T. (2019). Predicting private company failure: A multi-class analysis. *Journal of International Financial Markets, Institutions and Money*, 61, 161-188. <https://doi.org/10.1016/j.intfin.2019.03.004>
- Kaplan, A. D. H. (1950). The influence of size of firms on the functioning of the economy. *The American Economic Review*, 40(2), 74-84. <http://www.jstor.org/stable/1818026>
- Kokoszka, P., Miao, H., Petersen, A., & Shang, H. L. (2019). Forecasting of density functions with an application to cross-sectional and intraday returns. *International Journal of Forecasting*, 35(4), 1304-1317. <https://doi.org/10.1016/J.IJFORECAST.2019.05.007>
- Laitinen, E. K., & Laitinen, T. (2000). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International Review of Financial Analysis*, 9(4), 327-349. [https://doi.org/10.1016/S1057-5219\(00\)00039-9](https://doi.org/10.1016/S1057-5219(00)00039-9)
- Lev, B., & Sunder, S. (1979). Methodological issues in the use of financial ratios. *Journal of Accounting and Economics*, 1(3), 187-210. [https://doi.org/10.1016/0165-4101\(79\)90007-7](https://doi.org/10.1016/0165-4101(79)90007-7)
- Linares-Mustarós, S., Coenders, G., & Vives-Mestres, M. (2018). Financial performance and distress profiles. From classification according to financial ratios to compositional classification. *Advances in Accounting*, 40, 1-10. <https://doi.org/10.1016/J.ADIAC.2017.10.003>
- Linares-Mustarós, S., Farreras-Noguer, M. À., Arimany-Serrat, N., & Coenders, G. (2022). New financial ratios based on the compositional data methodology. *Axioms*, 11(12), 694. <https://doi.org/10.3390/AXIOMS11120694>
- Martín-Fernández, J. A., Palarea-Albaladejo, J., & Olea, R. A. (2011). Dealing with Zeros. In V. Pawlowsky-Glahn, & A. Buccianti (eds), *Compositional Data Analysis: Theory and Applications* (pp. 43-58). <https://doi.org/10.1002/9781119976462.CH4>
- Martín-Oliver, A., Ruano, S., & Salas-Fumás, V. (2017). The fall of Spanish cajas: Lessons of ownership and governance for banks. *Journal of Financial Stability*, 33, 244-260. <https://doi.org/10.1016/J.JFS.2017.02.004>
- McCann, B. T., & Folta, T. B. (2008). Location matters: Where we have been and where we might go in agglomeration research. *Journal of Management*, 34(3), 532-565. <https://doi.org/10.1177/0149206308316057>
- Mselmi, N., Lahiani, A., & Hamza, T. (2017). Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis*, 50, 67-80. <https://doi.org/10.1016/j.irfa.2017.02.004>
- Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2008). A modified EM algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34(8), 902-917. <https://doi.org/10.1016/J.CAGEO.2007.09.015>
- Peek, J., & Rosengren, E. S. (2005). Unnatural selection: Perverse incentives and the misallocation of credit in Japan. *American Economic Review*, 95(4), 1144-1166. <https://doi.org/10.1257/0002828054825691>

- Rich, P. (1992). The organizational taxonomy: Definition and design. *The Academy of Management Review*, 17(4), 758-781. <https://doi.org/10.2307/258807>
- Saus-Sala, E., Farreras-Noguer, À., Arimany-Serrat, N., & Coenders, G. (2021). Compositional DuPont analysis. A visual tool for strategic financial performance assessment. In Filzmoser, P., Hron, K., Martín-Fernández, J.A., Palarea-Albaladejo, J. (eds), *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn* (pp. 189-206). Springer. https://doi.org/10.1007/978-3-030-71175-7_10