ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN

ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC

Đề tài:

NGHIÊN CỨU VÀ ỨNG DỤNG PHÂN TÍCH DỮ LIỆU ĐA HỢP

Sinh viên thực hiện: Phan Ngọc Hưng

Lớp hệ thống thông tin 01 - K64

Giảng viên hướng dẫn: ThS. Nguyễn Văn Hạnh

Hà Nội - 2024

LỜI CẢM ƠN

Trong quá trình nghiên cứu và hoàn thành đồ án, em xin gửi lời cảm ơn chân thành tới thầy Nguyễn Văn Hạnh đã tạo điều kiện, hướng dẫn và chỉ bảo tận tình giúp em có định hướng, cũng như trang bị những kiến thức cần thiết về đề tài này.

Em xin cảm ơn các thầy cô trường Đại học Bách Khoa Hà Nội đã dạy cho em những kiến thức quan trọng, xây dựng cho em nền tảng vững chắc trên con đường học tập và nghiên cứu ở hiện tại cũng như trong tương lai.

Tuy nhiên, trong quá trình thực hiện bài báo cáo, do khả năng và kiến thức của em vẫn còn hạn chế nên không thể tránh khỏi một số thiếu sót khi trình bày và đánh giá vấn đề. Em rất mong nhận được sự góp ý, đánh giá của giảng viên để đề tài của em thêm hoàn thiện hơn.

Em xin trân thành cảm ơn!

TÓM TẮT NỘI DUNG ĐỒ ÁN

Em lựa chọn đề tài "Nghiên cứu và ứng dụng phân tích dữ liệu đa hợp" do tầm quan trọng và sự phổ biến của các dữ liệu đa hợp trong các nghiên cứu ngày nay, sự ảnh hưởng bởi các tính chất dạng tổng của nó có thể tạo ra những tính toán và phân tích sai lệch cho các nghiên cứu. Vì vậy đồ án này sẽ là tài liệu hỗ trợ cơ sở lý thuyết và ứng dụng trong việc phân tích các dữ liệu đa hợp

Ngoài các phần Lời cảm ơn, Tóm tắt và Kết luận, báo cáo còn bao gồm các nội dung chính sau:

1. Cơ sở lý thuyết

    1.1 Các khái niệm cơ bản về dữ liệu đa hợp

    1.2 Các phép biến đổi trong phân tích dữ liệu đa hợp

    1.3 Các vấn đề cần giải quyết khi phân tích dữ liệu đa hợp

    1.4 Kết hợp mô hình hồi quy cho dữ liệu đa hợp

2. Ứng dụng với bộ dữ liệu thực tế

    2.1 Tìm hiểu về bộ dữ liệu

    2.2 Ứng dụng phân tích dữ liệu đa hợp vào bộ dữ liệu

    2.3 Sử dụng mô hình hồi quy để đánh giá

LỜI CAM ĐOAN

Tôi tên là Phan Ngọc Hưng, mã số sinh viên 20195964, sinh viên lớp Hệ thống thông tin quản lý 01 - K64, khóa K64. Người hướng dẫn là ThS. Nguyễn Văn Hạnh. Tôi xin cam đoan toàn bộ nội dung được trình bày và thông tin trích dẫn đều tuân thủ các quy định về sở hữu trí tuệ; các tài liệu tham khảo được liệt kê rõ ràng. Tôi xin chịu hoàn toàn trách nhiệm với những nội dung được viết trong đồ án này.

Hà Nội, ngày tháng năm 2024

Người cam đoan

Phan Ngọc Hưng

MỤC LỤC

DANH MỤC KÝ HIỆU VÀ CHỮ VIẾT TẮT

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

LỜI NÓI ĐẦU

Trong những năm gần đây, phân tích dữ liệu đa hợp đã nhận được sự quan tâm ngày càng tăng trong nhiều lĩnh vực nghiên cứu như phân tích kinh doanh, địa chất học, sinh thái học, nghiên cứu địa lý dân cư và nghiên cứu vi sinh vật. Sự phổ biến của chúng xuất hiện khi các dữ liệu thu thập được thường có các biến số liên kết với nhau thông qua một tổng. Aitchison đầu tiên đã mô tả nền tảng lý thuyết để xử lý dữ liệu như vậy dựa trên các biến đổi tỷ lệ log (Aitchison, 1981; Aitchison, 1982). Tuy nhiên, phương pháp này vẫn chưa được sử dụng rộng rãi mặc dù sự ràng buộc tổng mà bộ dữ liệu mang lại có thể khiến các phân tích thông thường tạo ra những đánh giá sai lệch về bộ dữ liệu. Trong thực tế, các phương pháp dữ liệu tổ hợp chủ yếu được áp dụng trong các ngành địa chất học nhưng ngay cả trong lĩnh vực này cũng không phải là một quy trình được chuẩn hóa (Buccianti và Grunsky, 2014).

Chính vì vậy, trong đồ án này, em xin chọn đề tài "Nghiên cứu và ứng dụng phân tích dữ liệu đa hợp" để nêu ra một hướng giải quyết cho vấn dề các dữ liệu có dạng đa hợp như vậy. Nội dung của đồ án này gồm có 2 chương chính như sau:

Chương 1: Cơ sở lý thuyết. Trình bày tổng quan về cơ sở lý thuyết về phân tích dữ liệu đa hợp, các phép biến đổi dữ liệu phổ biến được sử dụng để sứ lý các dữ liệu và một vài trường hợp của bộ dữ liệu cần lưu ý khi phân tích đa hợp. Bên cạnh đó chương này cũng sẽ giới thiệu về việc kết hợp mô hình hồi quy sau khi sử lý dữ liệu đa hợp.

Chương 2: Ứng dụng với dữ liệu. Tìm hiểu bộ dữ liệu, ứng dụng các phép biến đổi đã được giới thiệu trong Chương 1 và tiến hành sử dụng mô hình hồi quy để dự đoán và đánh giá bộ dữ liệu.

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT

1.1 Lý thuyết về phân tích dữ liệu đa hợp

Compositional data analysis (CoDA) là phân tích dữ liệu đa hợp trong đó dữ liệu đa hợp đề cập đến dữ liệu mô tả các phần của một tổng thể nào đó và thường được trình bày dưới dạng các vector tỷ lệ, phần trăm, nồng độ hoặc tần suất. Tổng của các vector này bị ràng buộc bởi một hằng số cố định, thường là 1 hoặc 100%. Aitchison lần đầu tiên mô tả nền tảng lý thuyết để xử lý loại dữ liệu này dựa trên các biến đổi tỷ số logarit (Aitchison, 1981; Aitchison, 1982). Thực tế, các phương pháp dữ liệu cấu phần chủ yếu được áp dụng trong khoa học địa chất, ví dụ, CoDa đã được áp dụng trong các khảo sát đất và địa hóa học (Zhang và cộng sự, 2013; Grunsky, Mueller và Corrigan, 2014; Tepanosyan và cộng sự, 2020; Zheng và cộng sự, 2021), trong các nghiên cứu về nước và nước ngầm (Buccianti, 2018). Ngoài khoa học địa chất, kỹ thuật này bắt đầu được sử dụng trong các nghiên cứu khác nhau ở nhiều lĩnh vực, bao gồm đánh giá phân phối nước đô thị (Ebrahimi và cộng sự, 2021), nghiên cứu sức khỏe (McKinley và cộng sự, 2020; Dumuid và cộng sự, 2021; Verswijveren và cộng sự, 2021), nghiên cứu dinh dưỡng (Leite, 2019), và dự báo cấu trúc tiêu thụ năng lượng (Wei và cộng sự, 2021).

Một vector u là một vector của D thành phần của một tổng thể mang thông tin với ràng buộc tổng đơn vị được biểu diễn trong một không gian đơn hình S^D định nghĩa bởi Aichison (1982) như sau:

$$S^D = \{u = (u_1, ..., u_D) \in R^D : u_m > 0, m = 1, ..., D; \sum_{m=1}^{D} u_m = 1\}, \qquad (1.1)$$

Phân tích dữ liệu hợp thành sử dụng các biến đổi tỷ lệ logarit (logarit transformations) để ánh xạ không gian đa diện S^D sang R^q (thường là $q = D-1$) do giá trị của bất kỳ phần nào cũng có thể biết được dựa trên giá trị của các phần còn lại. Các biến đổi cổ điển bao gồm biến đổi tỷ lệ logarithm cộng tính (alr), tỷ lệ logarithm trung tâm

(clr) và tỷ lệ logarithm đẳng cự (ilr). Cụ thể chúng ta sẽ đi tìm hiểu vào phần 1.2

1.2 Sử dụng logarit trong chuyển đổi, tính toán và phân tích vector đa hợp

1.2.1 Chuyển đổi loga

Ví dụ đơn giản nhất về logarit là tỷ lệ chuyển đổi log của hai phần của một tác phẩm hoặc logarit theo cặp, được LR biểu thị trong suốt quá trình xem xét này. Cho một tập J thành phần $x_1, x_2, ...mx_j$ tức là có $\frac{1}{2}J(J-1)$ các LR duy nhất có dạng:

$$\mathrm{LR}(j, j^{'}) = \log\left(\frac{x_j}{x_{j'}}\right), \quad j, j' = 1, \ldots, J, \quad j < j' \tag{1.2}$$

Bất kỳ tập con nào của $J-1$ các LR độc lập tuyến tính mà bao gồm tất cả các phần đa hợp tạo thành một cơ sở có thể tạo ra tất cả các LRs khác bằng tổ hợp tuyến tính. Tập con đơn giản nhất như vậy là các tỉ số logarit cộng (ALRs), trong đó một phần tham chiếu cụ thể được so sánh với tất cả các phần khác (ở đây, phần cuối cùng tức $J-1$ được chọn là phần tham chiếu trong mẫu số):

$$ALR\left(\frac{j}{J}\right) = log\left(\frac{x_j}{x_J}\right), j = 1, \ldots, J-1 \tag{1.3}$$

Có D phép biến đổi ALR có thể có, phụ thuộc vào phần tham chiếu được chọn, vì vậy một bước quan trọng sẽ là chọn phần tham chiếu để đáp ứng mục tiêu thống kê hoặc nghiên cứu.

Có một số biến đổi tỉ số log dựa trên trung bình hình học để kết hợp các phần, trong đó quan trọng nhất là biến đổi tỉ số log tập trung (CLR) (Aitchison 1986). Một CLR là tỉ số log giữa một phần và trung bình hình học của tất cả J phần trong tổ hợp(1 tổ hợp cấu thành dữ liệu đa hợp). Do đó, có J các CLRs, được định nghĩa là...:

$$\mathrm{CLR}(j) = \log\left(\frac{x_j}{\left(\prod_{j'} x_{j'}\right)^{1/J}}\right) = \log(x_j) - \frac{1}{J}\sum_{j'}\log(x_{j'}), \quad j = 1, \ldots, J \tag{1.4}$$

Do đó, một CLR là logarith của một phần, được tập trung so với trung bình của các logarith của tất cả các phần trong không gian mẫu. sự khác biệt giữa hai CLR

là giống như tỉ số log của hai phần tử trong tử số, và các CLR có thể được phân tích trong một phân tích thành phần giảm chiều để đại diện cho phân tích tương đương của tất cả các LRs (Aitchison và Greenacre 2002).

Khác với việc lấy một phần (được coi là trung tâm) rồi so sánh tương quan với trung bình hình học của tất cả các phần khác, thì Biến đổi tỉ số log đẳng cự (ILR) (Egozcue và cộng sự 2003) được định nghĩa so sánh tương quan trực tiếp giữa 2 phần với nhau,là sự kết hợp tuyến tính của các phần được chuyển đổi logarit, với tổng các hệ số của tổ hợp bằng 0 (với a_j là các hệ số bất kỳ): $\sum_j a_j log(x_j)$ trong đó $\sum_j a_j = 0$, khi này ta sẽ chọn sao cho nếu hệ số dương $a_j > 0$ xác định chỉ số tập con J_1 và các hệ số âm $a_j < 0$ chỉ số tập con J_2, độ tương quan giữa 2 logarit sẽ được tính:

$$\sum_{j\in I_1} |a_j| \log(x_j) - \sum_{j\in I_2} |a_j| \log(x_j) = \log\left(\frac{\prod_{j\in I_1} x_j^{|a_j|}}{\prod_{j\in I_2} x_j^{|a_j|}}\right), \tag{1.5}$$

Giả sử ta lấy $|J_1|$ và $|J_2|$ lần lượt đại diện cho số hệ số dương và số hệ số âm, vậy ta sẽ chọn các hệ số dương là $a_j = \frac{1}{|J_1|}$, và các hệ số âm là $a_j = \frac{-1}{|J_2|}$ khi đó ILR được định nghĩa như sau:

$$\begin{aligned} \text{ILR}(J_1, J_2) &= \sqrt{\frac{|J_1||J_2|}{|J_1|+|J_2|}} \log\left(\frac{\left(\prod_{j\in J_1} x_j\right)^{1/|J_1|}}{\left(\prod_{j\in J_2} x_j\right)^{1/|J_2|}}\right) \\ &= \sqrt{\frac{|J_1||J_2|}{|J_1|+|J_2|}} \left(\frac{1}{|J_1|}\sum_{j\in J_1} \log(x_j) - \frac{1}{|J_2|}\sum_{j\in J_2} \log(x_j)\right). \end{aligned} \tag{1.6}$$

Chú ý rằng độ tương phản logarit của phương trình tương quan giữa 2 logait chưa hề xét tới việc số lượng các phần tử của 2 tập bằng nhau nên $\sqrt{\frac{|J_1||J_2|}{|J_1|+|J_2|}}$ (tức trung bình của J_1 và J_2) giúp điều chỉnh cho sự khác biệt về số lượng phần tử.

Yếu tố điều chỉnh lại tập hợp các hệ số $\frac{1}{|J_1|}$ và $\frac{-1}{|J_2|}$ mà xác định cân bằng ILR để có độ dài đơn vị. Khi đó $J-1$ tập các hệ số trực giao với nhau, kết quả $J-1$ ILR xác định cơ sở chực chuẩn. Công thức tính ILR đã nêu là tổng trung bình của $|J_1|$x$|J_2|$ các LR.

Từ việc biến đổi các

Từ việc biến đổi các vector dữ liệu thông qua các phép logait sau ta thu được một bảng kết quả so sánh:

Tên	**Mapping**	**Bảo tồn khoảng cách**	**Ánh xạ ngược**
CLR	$\mathbb{S}^D \to \mathbb{R}^D$	Có	Không
ILR	$\mathbb{S}^D \to \mathbb{R}^{D-1}$	Có	Có
ALR	$\mathbb{S}^D \to \mathbb{R}^{D-1}$	Không	Có

Bảng 1.1 Bảng thông tin về các phép biến đổi logait

Như vậy, mỗi phép biến đổi đều sẽ có ưu nhược điểm khác nhau và phù hợp với từng đặc điểm của dữ liệu và mục đích nghiên cứu.

1.2.2 Phương sai của logarit

Vì LR là khái niệm trung tâm trong CoDA nên định nghĩa về độ biến thiên tổng thể trong tập dữ liệu thành phần bao gồm một số thành phần được quan sát được thực hiện theo chúng.

Việc định lượng tổng phương sai trong một tập dữ liệu thành phần là tất yếu, vì đây là phương sai mà chúng ta muốn giải thích, cả trong học không giám sát cũng như trong học có giám sát khi các thành phần được coi là biến phản hồi. Aitchison đã định nghĩa $D \times D$ ma trận phương sai $T = [\tau_{jk}]$, trong đó $\tau_{jk} = \mathrm{Var}(\log(x_j/x_k))$, là phương sai của LR (j, k)-th, LR(j, k), trong công thức tính LR. Đường chéo của ma trận đối xứng T bao gồm các số không và mỗi tam giác trên và dưới chứa $\frac{1}{2}D(D-1)$ các phương sai LR.

Một kết quả cũng được chỉ ra trong [2] và do ràng buộc tổng đơn vị, là các hiệp phương sai giữa các LR có thể được tính từ các phương sai của chúng:

$$\mathrm{Cov}(\mathrm{LR}(j,k),\mathrm{LR}(u,v)) = \frac{1}{2}(\tau_{jk} + \tau_{uv} - \tau_{ju} - \tau_{kv}) \tag{1.7}$$

Để đo lường tổng sự phân tán trong một tập dữ liệu thành phần, Aitchison đã

định nghĩa tổng biến thiên logarit là tổng của $\frac{1}{2}D(D-1)$ phương sai LR chia cho D,

$$\frac{1}{D}\sum_{j<k}\tau_{jk}. \tag{1.8}$$

Lý do cho việc chia cho D là để làm cho tổng biến thiên tương đương với cách tính hiệu quả hơn của cùng một phép đo này, bằng cách đơn giản là tổng hợp D phương sai của các CLR trong (4):

$$\sum_{j}\mathrm{Var}(\mathrm{CLR}(j)). \tag{1.9}$$

Một phép đo phương sai tổng hơi khác, được ký hiệu ở đây là TotVar, đã được giới thiệu trong [42], chia các phép đo của Aitchison cho D, điều này gán trọng số $1/D^2$ cho mỗi phương sai LR, và $1/D$ cho mỗi phương sai CLR, do đó tính trung bình các phương sai CLR thay vì tổng hợp chúng:

$$\mathrm{TotVar} = \frac{1}{D^2}(\sum\sum)_{j<k=2}^{D}\mathrm{Var}(\mathrm{LR}(j,k)) = \frac{1}{D}\sum_{j=1}^{D}\mathrm{Var}(\mathrm{CLR}(j)) \tag{1.10}$$

Như đã đề cập trước đó, định nghĩa này có ưu điểm là so sánh được với các phép đo tổng phương sai logarit trên các tập dữ liệu có số lượng thành phần khác nhau, trong khi $\sum_j \mathrm{Var}(\mathrm{CLR}(j))$ tăng lên với số lượng thành phần phần. Ngoài ra, việc nghĩ rằng mỗi phần nhận được một trọng số bằng $1/D$ gợi lên ý tưởng chung về việc thay đổi trọng số của các phần.

Ý tưởng về việc gán trọng số cho các phần trong CoDA bắt nguồn từ các công trình của Paul Lewi [66] — xem thêm [67, 68, 50].

Đóng góp của Lewi cần được cộng đồng CoDA công nhận rộng rãi hơn, vì ông đã tiên đoán nhiều khái niệm được định nghĩa sau này bởi Aitchison và các tác giả khác, như được mô tả trong [89]. Các định nghĩa tổng quát hơn của công thức tính TotVar, với các trọng số phần thay đổi c_j, $j \in \{1,\ldots,J\}$ thỏa mãn $\sum_j c_j = 1$ là:

$$\mathrm{TotVar} = \sum_{j<k}^{D} c_j c_k \mathrm{Var}(\mathrm{LR}(j,k)) = \sum_{j=1}^{D} c_j \mathrm{Var}(\mathrm{CLR}(j)) \tag{1.11}$$

trong đó các định nghĩa không trọng số (tức là, trọng số bằng nhau) là trường hợp đặc biệt $c_j = 1/D$ cho tất cả các phần D.

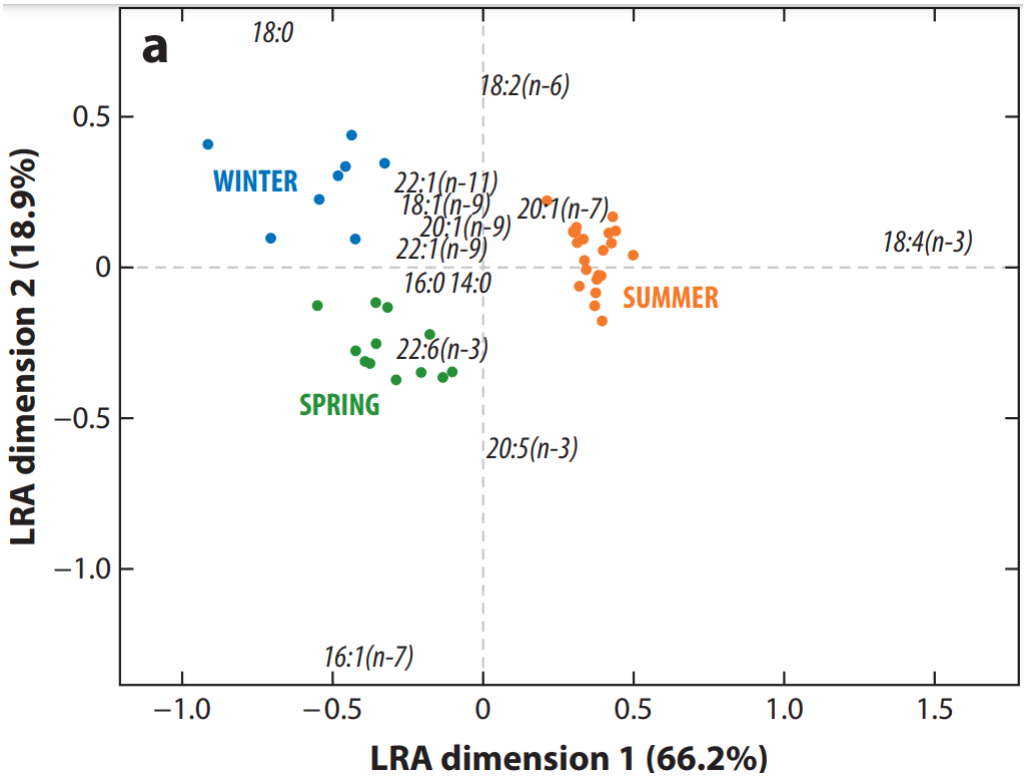
Trọng số khác biệt có thể hữu ích khi các phần có tỷ lệ trung bình thấp tạo ra tỷ lệ cao hơn nhiều so với các phần có tỷ lệ trung bình cao hơn, do đó đóng góp quá mức vào tổng phương sai logarit. Các phương sai logarit cao hơn trong các phần hiếm hơn thường do lỗi đo lường tương đối cao. Trong khái niệm CoDA của Lewi, trọng số phần mặc định là tỷ lệ trung bình của các phần, do đó giảm trọng số các phần có trung bình thấp so với các phần có trung bình cao hơn. Hệ thống trọng số này giống hệt với hệ thống được sử dụng trong phân tích tương quan, nhưng bất kỳ hệ thống nào khác phù hợp với dữ liệu và mục tiêu nghiên cứu đều có thể được chọn.

1.2.3 Khoảng cách loga: phân tích thành phần và cụm

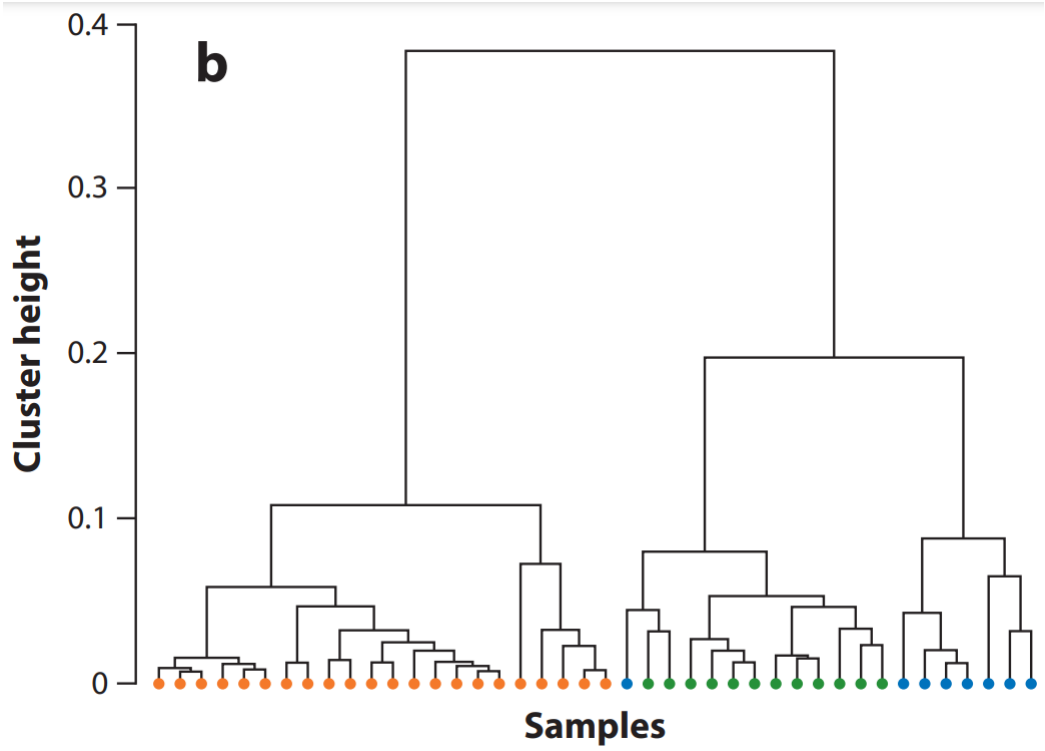
Khi dữ liệu thành phần đã được chuyển đổi logarit, về cơ bản, phân tích đa hợp có thể được tiến hành bình thường như trước đây đối với dữ liệu ở quy mô theo khoảng thời gian, với sự điều chỉnh thích hợp của cách giải thích với thực tế là các logarits có độ phức tạp khác nhau. Phân tích thành phần chính (PCA) của tất cả $\frac{1}{2}J(J-1)$ LR được gọi là phân tích logarit (LRA) (Greenacre và Lewi, 2009; Greenacre, 2018; Greenacre, 2019). LRA tương đương với PCA của J các CLR hiệu quả hơn nhiều về mặt tính toán.

Season	14:0	16:0	16:1(n-7)	18:0	18:1(n-9)	18:2(n-6)	18:4(n-3)	20:1(n-9)	20:1(n-7)	20:5(n-3)	22:1(n-11)	22:1(n-9)	22:6(n-3)
Winter	15.78	13.54	7.28	7.25	8.29	2.26	1.88	17.8	0.96	4.92	11.28	1.44	7.34
Winter	13.44	13.63	8.19	9.24	8.73	2.10	1.29	16.83	0.78	5.18	11.92	1.51	7.16
Winter	7.11	13.14	7.29	15.64	5.91	2.89	1.51	15.34	1.16	5.96	13.43	1.74	8.89
Winter	13.49	12.55	8.34	8.08	8.61	1.94	1.03	19.44	0.96	5.40	10.87	1.46	7.84
Winter	16.54	12.28	8.74	3.25	8.89	1.47	0.84	14.12	0.97	8.08	11.02	1.42	12.37
Winter	8.42	12.53	7.50	20.15	7.08	1.75	0.25	17.37	0.71	4.84	10.71	1.39	7.31
Winter	5.01	11.17	8.15	5.13	9.66	2.16	0.57	20.87	0.85	8.33	14.79	1.80	11.52
Winter	8.19	11.81	9.16	3.71	8.11	1.57	0.26	19.64	0.81	9.39	13.05	2.08	12.21
Spring	9.93	10.51	20.71	2.28	6.56	0.99	2.26	11.30	0.76	13.07	7.92	1.06	12.65
Spring	8.79	10.39	27.45	2.06	4.46	0.71	2.99	10.80	0.67	15.58	6.17	0.93	9.01
Spring	8.56	10.09	30.21	1.50	5.52	0.94	1.65	10.81	0.79	11.90	6.79	1.01	10.21
Spring	9.47	10.30	31.49	1.51	4.54	0.80	1.19	10.36	0.82	11.15	7.24	1.03	10.10
Spring	10.32	10.92	24.97	4.69	5.83	0.96	0.72	13.30	0.78	7.99	8.74	1.22	9.56
Spring	9.00	12.68	21.73	4.52	5.47	0.92	1.88	14.23	0.68	11.06	7.56	1.29	8.99
Spring	7.36	9.89	30.49	1.19	5.02	1.18	1.85	11.32	0.83	13.98	6.82	1.01	9.07
Spring	9.78	10.2	32.26	1.80	4.98	1.23	0.91	9.28	0.68	10.30	7.10	1.10	10.37
Spring	7.45	11.53	23.83	2.98	4.42	0.79	0.93	10.21	0.70	13.33	7.20	1.14	15.50
Spring	9.31	10.72	21.17	2.37	6.28	1.19	1.09	18.02	0.93	8.40	11.44	1.74	7.35
Spring	7.02	10.93	28.00	3.52	5.10	0.80	1.37	10.58	0.76	12.05	7.40	1.09	11.39
Spring	6.60	11.16	29.73	3.27	4.67	0.80	1.10	8.04	0.79	13.53	5.56	1.06	13.68
Summer	7.34	8.74	6.73	1.89	5.23	2.54	16.11	13.34	2.16	13.37	9.15	1.01	12.38
Summer	7.51	8.76	7.07	2.00	5.08	2.64	17.31	12.66	2.38	7.47	9.67	0.98	16.47
Summer	7.05	8.67	8.01	1.74	4.73	2.49	15.12	13.94	2.28	12.59	9.57	1.09	12.72
Summer	5.83	13.27	7.13	5.50	4.69	4.33	9.66	5.67	3.08	14.01	5.52	0.88	20.41
Summer	10.15	9.91	7.04	1.68	9.48	1.89	10.34	17.20	1.47	11.69	10.19	1.44	7.50
Summer	10.18	8.89	7.53	1.60	8.44	1.90	10.90	16.77	1.69	13.06	10.83	1.45	6.75
Summer	10.57	9.49	6.70	1.88	9.04	1.69	11.14	17.09	1.31	12.17	9.89	1.22	7.82
Summer	10.81	8.68	12.5	1.38	6.58	1.56	10.92	16.25	1.34	13.48	9.26	1.32	5.94
Summer	8.22	7.58	9.29	1.61	5.55	2.03	11.56	18.14	1.18	14.55	10.45	1.50	8.34
Summer	9.20	8.16	10.75	1.39	5.91	1.94	11.75	18.35	1.35	11.34	9.35	1.36	9.13
Summer	8.53	9.34	6.54	1.88	8.79	1.67	10.71	18.44	1.28	12.87	10.52	1.38	8.07
Summer	9.38	8.82	7.45	1.71	8.08	1.79	11.03	17.68	1.27	12.67	10.03	1.40	8.69
Summer	9.34	9.12	5.73	1.85	8.19	1.63	10.88	17.09	1.22	13.48	11.48	1.45	8.52
Summer	6.13	6.06	9.46	1.50	4.00	3.09	18.96	14.27	1.58	13.74	8.90	1.45	10.86
Summer	7.61	7.36	14.46	1.30	3.94	1.95	12.10	14.49	1.42	15.78	8.56	1.22	9.82
Summer	6.21	7.17	15.44	0.95	4.01	1.73	11.67	15.78	1.39	15.70	9.42	1.32	9.20
Summer	9.39	8.22	8.89	0.99	7.03	1.65	11.18	16.19	1.22	14.69	10.97	1.48	8.09
Summer	9.91	8.10	6.77	1.05	7.36	2.04	11.67	17.59	1.32	13.32	11.23	1.50	8.16
Summer	8.78	7.43	6.52	1.18	7.41	2.99	12.94	16.82	1.23	12.39	11.30	1.52	9.48
Summer	9.51	8.56	11.61	0.92	7.53	1.75	11.06	16.96	1.26	13.39	9.46	1.26	6.73
Summer	9.47	8.57	9.38	1.08	7.78	1.67	12.66	16.22	1.26	13.13	9.68	1.32	7.79
Summer	9.07	8.02	9.99	1.15	7.39	1.83	12.19	16.24	1.48	13.11	9.45	1.37	8.70

Hình 1.1 Tỷ lệ axit béo trong thành phần con gồm 13 thành phần của tập dữ liệu gồm 40 phép đo



Hình 1.2 Phân tích logarit (LRA) dạng biểu đồ kép (Greenacre, 2010) của thành phần phụ axit béo 13 phần (FA), trong đó khoảng cách giữa các mẫu gần đúng với khoảng cách logarit và hướng giữa các cặp FA biểu thị trục hai biểu đồ logarit.



Hình 1.3 Phân cụm các khoảng cách logarit giữa 42 mẫu

Hình 1.2 hiển thị LRA của tập dữ liệu FA, sử dụng hàm *LRA*() sau đó dễ dàng *CoDA* để tính toán, có thêm mã màu của các mẫu theo mùa. Mặc dù việc diễn giải vị trí của FA dưới dạng các biến trong PCA thông thường là rất tốt nhưng chúng chỉ có ý nghĩa ở vị trí theo cặp của chúng. Ví dụ: logarit $(16:1(n-7)/18:0)$ sẽ theo hướng kết nối giữa hai FA này, hướng gần như thẳng đứng xuống dưới, vì $16:1(n-7)$ nằm trong tử số. Do đó, cần luôn nhớ rằng PCA của CLR chỉ là một cách viết tắt để phân tích tất cả các LR theo cặp, cái sau là các biến được quan tâm chính.

Thường có hai loại thuật toán phân cụm: phân cụm phân cấp, thường được áp dụng cho ít hơn khoảng 100–150 đối tượng để hiển thị kết quả trong một cây phân cấp, và phân cụm không phân cấp, thường được áp dụng cho các tập hợp đối tượng lớn đến rất lớn nơi mục tiêu chỉ là chia các đối tượng thành các nhóm đồng nhất nội bộ, gọi là các cụm. Bởi vậy, khi số lượng các thành phần tăng lên, các biểu đồ kép như Hình 1.2 sớm trở nên đông đúc vì có quá nhiều điểm gần hoặc trồng lên nhau.

Cho nên việc phân tích cụm phân cấp trong trường hợp này là tối ưu hơn.

Phân tích cụm nhóm các quan sát tương tự theo các tiêu chí cụ thể . Do đó, các quan sát trong cùng một cụm là tương tự nhau và chúng khác với các quan sát trong các nhóm khác. Phương pháp phân cụm phân cấp của Ward (tức là phương pháp kết tụ từ dưới lên) dựa trên tổng tiêu chí bình phương đã được áp dụng trong hình 1.3. Phương pháp này nhằm mục đích chia tập hợp các phần của một chế phẩm thành hai nhóm phần. Các nhóm được tạo ra phải được chia lại thành hai nhóm phần, v.v., cho đến khi chúng ta có được các nhóm chỉ có một phần. Số lượng các bộ phận kết quả sẽ là D-1.

Bản chất của phân cụm Ward có thể nói rằng việc phân từng quan sát thành các cụm, cụ thể hơn, giả sử với 5 quan sát a_1 a_2 a_3 a_4 a_5 ta sẽ tính khoảng cách giữa các cụm (ban đầu mỗi cụm chính là 1 quan sát) sau đó giả sử khoảng cách giữa a_3 và a_5 là ngắn nhất (hay về mặt loga hóa là thay đổi ít nhất) sẽ được nhóm thành một cụm (a_3, a_5) rồi sau đó tiếp tục tính khoảng cách giữa các cụm là a_1 a_2 a_4 và (a_3, a_5), tiếp tục làm thế cho tới khi chỉ còn 1 cụm duy nhất bao gồm tất cả quan sát. Sau đây sẽ trình bày kỹ hơn về khoảng cách logarit giữa các quan sát (hay còn gọi là các mẫu)

Để thực hiện phân cụm, cần có một hàm khoảng cách giữa các đối tượng, và có một số lựa chọn:

Chuyển đổi dữ liệu hợp thành thành CLRs và sau đó sử dụng khoảng cách Euclidean. Đây là phân cụm trên các khoảng cách tỷ lệ logarithm (logarit) (không được trọng số).

Giống như trên, chuyển đổi dữ liệu hợp thành thành CLRs nhưng sau đó sử dụng phân cụm k-means không phân cấp, cũng sử dụng khoảng cách Euclidean.

Chuyển đổi dữ liệu hợp thành thành ALRs, là phương án tối ưu về mặt thống kê, hoặc dựa trên lý do địa hóa học, và sau đó sử dụng khoảng cách Euclidean; ở đây, tính tối ưu thống kê được định nghĩa là tạo ra hình học gần nhất với hình học của tất cả các tỷ lệ logarithm (Greenacre và cộng sự, 2021)

Để phân cụm, các thuật toán phân cụm phân cấp hoặc không phân cấp thông thường có thể được thực hiện trên các mẫu, sử dụng khoảng cách Euclide được xác định trên tất cả các LR, nhưng lại được tính toán hiệu quả hơn bằng cách sử dụng CLR trên không gian đơn hình. Nếu ma trận $Y = [y_{ij}]$ biểu thị tập dữ liệu được chuyển đổi CLR và $Z = [z_{i,jj'}]$ biểu thị ma trận của loga LR $(x_{ij}/x_{ij'})$,sau đó $d_{ii'}$-khoảng cách logarit giữa các mẫu i và i', có thể được định nghĩa ở hai dạng tương đương, tương tự như các định nghĩa trong Công thức 9 và 10, là

$$d_{ii'} = \sqrt{\frac{1}{J^2}\sum\sum_{j<j'}(z_{i,jj'} - z_{i',jj'})^2} = \sqrt{\frac{1}{J}\sum_{j}(y_{ij} - y_{ij'})^2} \tag{1.12}$$

Lưu ý rằng $z_{i,jj'} - z_{i',jj'} = \log\left(\frac{x_{ij}}{x_{ij'}}\right) - \log\left(\frac{x_{i'j}}{x_{i'j'}}\right) = \log\left(\frac{x_{ij}}{x_{i'j}}\right) - \log\left(\frac{x_{ij'}}{x_{i'j'}}\right)$ sự khác biệt giữa hai logarits theo hàng giống hệt với sự khác biệt được tính toán trên cùng bốn giá trị theo cột. Cả hai đều bằng $\log(\frac{x_{ij}x_{i'j'}}{x_{ij'}x_{i'j}})$ được biểu thị bằng $S_{ii'}, jj'$ tức là logarit của tỷ lệ sản phẩm chéo $\frac{(x_{ij}x_{i'j'})}{(x_{ij'}x_{i'j})}$. Định nghĩa đầu tiên về khoảng cách giữa các mẫu trong phương trình của $d_{ii'}$ do đó có thể được viết là $d_{ii'} = \sqrt{\frac{1}{J^2}\sum\sum_{j<j'}(S_{ii',jj'})^2}$. Theo kiểu đối xứng, khoảng cách logarit tương ứng giữa các phần j và j' sẽ bao gồm tổng của các giá trị bình phương giống nhau $(S_{ii',jj'})^2$, trên tất cả các cặp hàng duy nhất: $d_{jj'} = \sqrt{\frac{1}{I^2}\sum\sum_{i<i'}(S_{ii'jj',})^2}$

Để có được sự giống nhau giữa một phần khoảng cách sử dụng CLR, CLR trước tiên phải được tính toán lại theo cột —nghĩa là, mỗi cột của ma trận dữ liệu thành phần được chuyển đổi log phải được căn giữa theo giá trị trung bình của cột và sau đó là một công thức tương tự như về phải của phương trình tính ($d_{ii'}$ được áp dụng, tính trung bình trên I hàng.

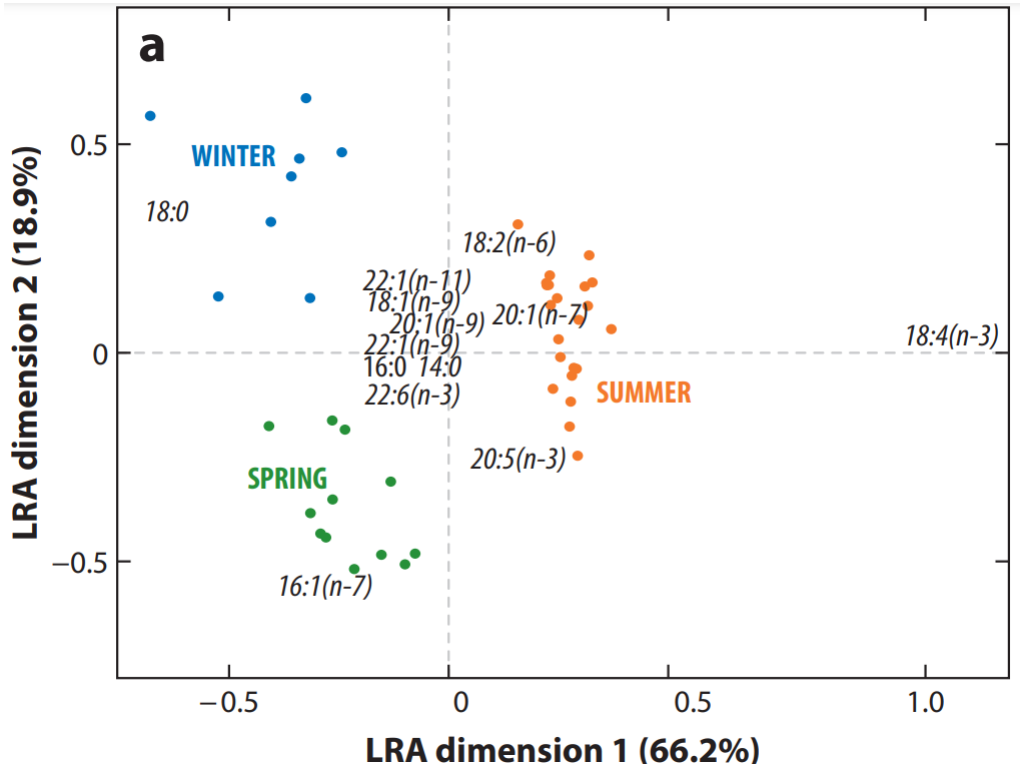
Tổng phương sai logarit có thể được tính tương đương từ tổng khoảng cách logarit bình phương giữa các mẫu hoặc giữa các phần:

$$TotVar = \frac{1}{I^2}\sum\sum_{i<i'} d_{ii'}^2 = \frac{1}{J^2}\sum\sum_{j<j'} d_{jj'}^2 \tag{1.13}$$

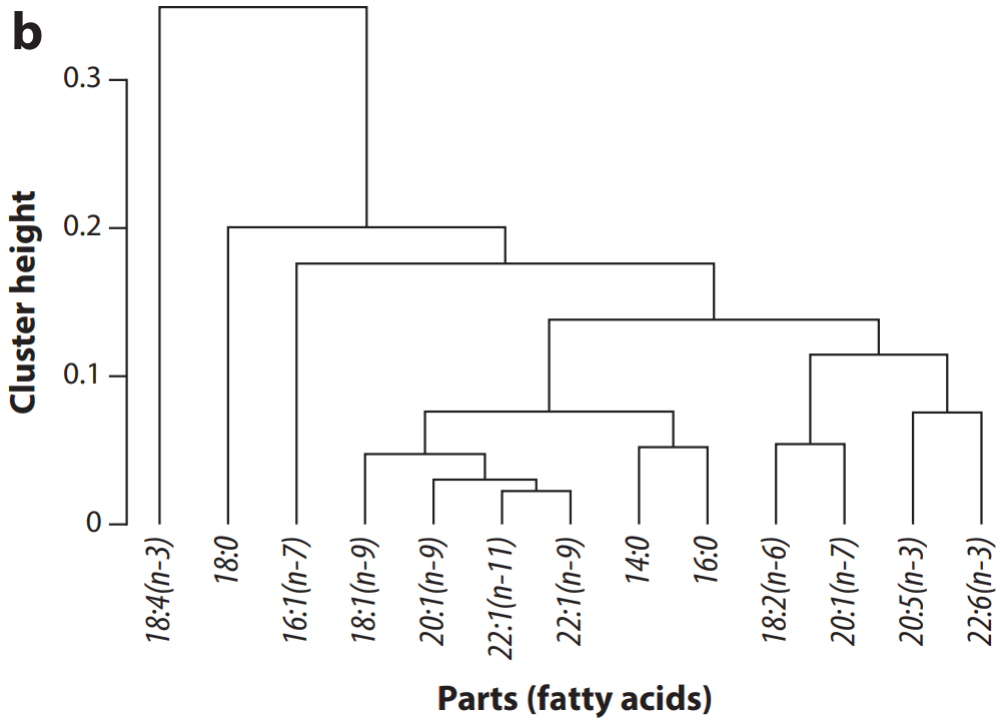
Tuy nhiên, một cách khác để thu được tổng phương sai logarit và khoảng cách logarit, giữa các hàng hoặc giữa các cột, là tính ma trận tâm kép của ma trận dữ liệu được chuyển đổi $\log(X)$:

$$H = (I - \frac{1}{I} 1_I 1_I^T) \log(X) (I - \frac{1}{J} 1_J 1_J^T) \tag{1.14}$$

Khi 1_I và 1_J là các vectơ của I và J tương ứng. Khi đó tổng phương sai là trung bình các phần tử bình phương của H : $TotVar = \frac{1}{IJ} \sum_i \sum_j b_{ij}^2$ khoảng cách giữa các bình phương là mức trung bình của chênh lệch bình phương giữa các hàng (mẫu): $d_{ii'}^2 = \frac{1}{J} \sum_j (h_{ij} - h_{i'j})^2$ và khoảng cách bình phương giữa các cột là giá trị trung bình của chênh lệch bình phương giữa các cột (phần): $d_{jj'}^2 = \frac{1}{I} \sum_i (h_{ij} - hij')^2$ Hơn nữa, LRA (Hình 1.2 Và 1.4) tương đương với việc thực hiện phân rã giá trị số ít (SVD) của H. Điều này là do phép chuyển đổi CLR loại bỏ phương tiện hàng của dữ liệu được chuyển đổi $\log(X)$, sau đó PCA loại bỏ cột không có nghĩa; do đó,$\log(X)$ là tâm kép trong LRA.



Hình 1.4 khoảng cách logarit giữa các axit béo (FA) hiện được hiển thị gần đúng



Hình 1.5 Phân cụm phường của khoảng cách logarit giữa các FA

Ngoài ra để phù hợp với những dữ liệu lớn hơn (khi 1 biểu đồ phân cấp sẽ trở nên phức tạp nếu số lượng thành phần của dữ liệu đa hợp lớn hơn 100), chúng ta sẽ trình bày với việc phân cụm không phân cấp bằng thuật toán phân cụm K-mean.

Trong thuật toán k-Means mỗi cụm dữ liệu được đặc trưng bởi một tâm (centroid). tâm là điểm đại diện nhất cho một cụm và có giá trị bằng trung bình của toàn bộ các quan sát nằm trong cụm. Chúng ta sẽ dựa vào khoảng cách từ mỗi quan sát tới các tâm để xác định nhãn cho chúng trùng thuộc về tâm gần nhất. Ban đầu thuật toán sẽ khởi tạo ngẫu nhiên một số lượng xác định trước tâm cụm. Sau đó tiến hành xác định nhãn cho từng điểm dữ liệu và tiếp tục cập nhật lại tâm cụm. Thuật toán sẽ dừng cho tới khi toàn bộ các điểm dữ liệu được phân về đúng cụm hoặc số lượt cập nhật tâm chạm ngưỡng.

Về cơ bản thuật toán được thục hiện như sau:

1.- Khởi tạo ngẫu nhiên C_k tâm cụm $\mu_1, \mu_2, ..., \mu_k$

trong đó C_k là một thành phần của dữ liệu đa hợp

2.- Lặp lại quá trình cập nhật tâm cụm cho tới khi dừng:

a. Xác định nhãn cho từng điểm dữ liệu c_i dựa vào khoảng cách tới từng tâm cụm:

$$c_1 = \arg\min ||x_i - \mu_j||_2^2 \tag{1.15}$$

b. Tính toán lại tâm cho từng cụm theo trung bình của toàn bộ các điểm dữ liệu trong một cụm:

$$\mu_j := \frac{\sum_{i=1}^{n} 1(c_i = j)x_i}{\sum_{i=1}^{n} 1(c_i = j)} \tag{1.16}$$

Trong công thức 2.a thì $\left\|x_2^2\right\|$ là bình phương của norm chuẩn bậc 2, kí hiệu là L_2, norm chuẩn bậc 2 là một độ đo khoảng cách thường được sử dụng trong machine learning.

Trong công thức 2.b chúng ta sử dụng hàm $1(.)$, hàm này có giá trị trả về là 1 nếu nhãn của điểm dữ liệu c_i được dự báo thuộc về cụm j, trái lại thì trả về giá trị 0. Như vậy tử số của vế phải trong công thức 2.b chính là tổng khoảng cách của toàn bộ các điểm dữ liệu nằm trong cụm j trong khi mẫu số chính là số lượng các điểm dữ liệu thuộc cụm j. μ_j chính là vị trí của tâm cụm j mà ta dự báo tại thời điểm hiện tại. Trong thuật toán trên thì tham số mà chúng ta cần lựa chọn chính là số lượng cụm k. Thời điểm ban đầu ta sẽ khởi tạo k điểm dữ liệu một cách ngẫu nhiên và sau đó gán các tâm bằng giá trị của k điểm dữ liệu này. Các bước trong vòng lặp ở bước 2 thực chất là:

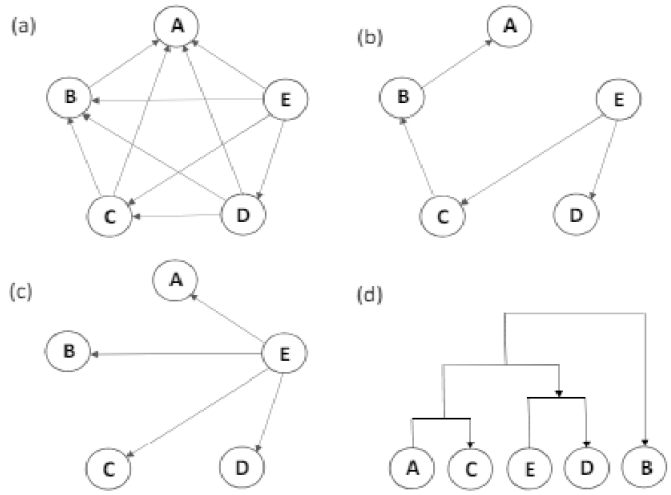
a. Gán nhãn cho mỗi điểm dữ liệu bằng với nhãn của tâm cụm gần nhất.

b. Dịch chuyển dần dần tâm cụm μ_j tới trung bình của những điểm dữ liệu mà được phân về j.

(lấy một ví dụ minh họa khi đã chuyển hóa loga rồi sử dụng k-mean, có thể thử dùng dữ liệu trong hình 1.1)

1.2.4 Lựa chọn biến loga

Vấn đề lựa chọn biến số rất quan trọng trong tất cả các khía cạnh của phân tích dữ liệu thành phần (CoDA), bởi vì có rất nhiều tỷ lệ logarit (LR) có sẵn và nhiều phép biến đổi logarit cộng tính (ALR) có thể thực hiện. Không gian logarit có chiều d = D - 1, có nghĩa là cần một tập hợp con chỉ gồm D - 1 LR độc lập tuyến tính để tạo ra tất cả các LR khác trong số $\frac{1}{2}D(D-1)-(D-1)=\frac{1}{2}(D-1)(D-2)$ các LR



Hình 1.6 Ảnh thể hện việc lựa chọn và phân cấp các LR

Đối với một thành phần gồm 5 phần, ví dụ, Hình (a) hiển thị đồ thị của tất cả 10 tỷ lệ logarit, trong đó các mũi tên chỉ vào mẫu số của các tỷ lệ tương ứng. Hình (b) và (c) hiển thị hai tập hợp con khác nhau gồm 4 LR độc lập tuyến tính mà tạo ra tất cả 6 LR khác. Hình (b) biểu đồ các tỷ lệ A/B, B/C, C/E và D/E. Sau đó, để có tỷ lệ A/E, có thể theo các mũi tên để thu được: $C/EB/CA/B = A/E$ (bất kỳ hướng mũi tên nào cũng có thể được đảo ngược để cho ra tỷ lệ ngược). Tuyến tính trong các logarit, sự bình đẳng này là $log(C/E)+log(B/C)+log(A/B)=log(A/E)$. Năm phần tạo thành một đồ thị có hướng không chu kỳ (DAG). Hình (c) cho thấy một biểu đồ có hướng

(DAG) khác với bốn mũi tên, đó là phép biến đổi ALR với E là phần tham chiếu. Cuối cùng, Hình (d) cho thấy một cách khác để định nghĩa bốn logarit, trong đó các phần được kết hợp tại mỗi nút của cây phân cấp. Nếu chúng được kết hợp bằng cách sử dụng trung bình hình học, điều này tương đương với việc định nghĩa một tập hợp ILR. Nhưng chúng cũng có thể được kết hợp đơn giản hơn dưới dạng các tổ hợp trong trường hợp đó, phương sai giải thích lý thuyết không đạt 100% nhưng trên thực tế rất gần với 100%

Thay vì giảm chiều, nơi tất cả các biến thành phần tham gia, ý tưởng giảm số lượng biến số là một phương án thay thế . Vấn đề lựa chọn một tập hợp nhỏ các LR đáp ứng mục tiêu thực tế đã được đề cập trong . Ý tưởng rất đơn giản và liên quan đến việc chọn LR theo từng bước, với mục tiêu giải thích được lượng phương sai tối đa trong tập dữ liệu thành phần. Một tập hợp đầy đủ D - 1 LR độc lập tạo thành DAG giải thích 100% phương sai logarit (giả sử D nhỏ hơn số quan sát N, nếu không đối với ma trận rộng cần N - 1 LR). Nhưng có lẽ ít LR hơn cũng đủ để đại diện cho toàn bộ các LR, vì tất cả dữ liệu đều có lỗi nội tại và do đó giải thích 100% phương sai là không cần thiết.

Thuật toán từng bước bắt đầu bằng cách chọn LR giải thích nhiều phương sai logarit nhất. Sau đó, nó được giữ lại và logarit tiếp theo, cùng với logarit đầu tiên, giải thích được phương sai tối đa, được giữ lại làm LR thứ hai, và cứ tiếp tục như vậy cho đến khi một tỷ lệ lớn phương sai, chẳng hạn khoảng 95%, được giải thích. Ngoài ra, một quy tắc dừng có thể áp đặt hình phạt mạnh lên số lượng logarit được chọn .

Tương tự, nếu một tập hợp ALR gần đẳng cự đã được xác định, chúng giải thích 100% phương sai logarit, nhưng một tập hợp nhỏ hơn có thể được chọn, lần này bằng cách loại bỏ ngược từng bước. Tại mỗi bước, ALR bị loại bỏ làm giảm phương sai giải thích và/hoặc hệ số Procrustes ít nhất.

Quá trình loại bỏ ngược này của ALR thực chất là chọn một thành phần con của các phần tử, do đó để tính hệ số Procrustes, một phương án thay thế là sử dụng hình

học logarit của thành phần con, bao gồm tất cả các LR trong thành phần con, không chỉ ALR. Sử dụng thành phần con gồm 25 phần thu được từ phân tích ALR, sau khi loại bỏ 27 nguyên tố, cung cấp cùng mức phương sai giải thích (95,3%), nhưng hệ số Procrustes được cải thiện nhẹ lên 0.972.

Các phương pháp trên phù hợp với bối cảnh không giám sát nơi cần giảm dữ liệu, nhưng cũng có thể được thực hiện trong bối cảnh giám sát, khi có một biến phản hồi quan sát, trong trường hợp đó, các LR được chọn từng bước để giải thích hoặc dự đoán tối đa biến phản hồi, chẳng hạn như trong bối cảnh các mô hình tuyến tính tổng quát , hoặc cây phân loại và hồi quy. Trong trường hợp này, việc sử dụng thước đo tương quan Procrustes của đẳng cự là không cần thiết, vì không có lý do gì ở đây mà các LR phải gần với đẳng cự.

Trong tất cả các chiến lược lựa chọn logarit cặp đôi, các tiêu chí tối ưu thống kê có thể được đặt cạnh kiến thức chuyên môn để chọn một tập hợp logarit thỏa mãn cả tiêu chí thống kê và sự liên quan thực chất đến câu hỏi nghiên cứu. Cách tiếp cận này đã được thực hiện thành công trong ba nghiên cứu, hai trong số đó là hóa sinh và một là khảo cổ học , trong đó ở mỗi bước, danh sách top 20 logarit, chẳng hạn, theo tiêu chí thống kê được tư vấn bởi nhà nghiên cứu, người sau đó chọn một logarit theo kiến thức chuyên môn. Thường thì các logarit hàng đầu rất gần nhau về tối ưu thống kê, vì vậy rất ít điều bị hy sinh khi chọn một logarit kém tối ưu hơn một chút nhưng có cách giải thích rõ ràng hơn..

Áp dụng phương pháp không giám sát, dữ liệu có tổng phương sai logarit, TotVar, bằng 0,2462, vì vậy câu hỏi đặt ra là: LR nào giải thích được phần lớn nhất của phương sai này? Bất kỳ LR nào cũng giải thích phần phương sai chứa đựng của nó, nhưng nó cũng giải thích các phần phương sai chứa trong nhiều LR khác mà nó có tương quan. Nó chỉ ra rằng, một mặt, $LR\log(16:0/18:4(n-3))$ giải thích được 65,6% TotVar,nhiều hơn bất kỳ LR nào khác. Mặt khác, phương sai chứa đựng của LR này chỉ là 5,0% của TotVar,nhưng chính phương sai được giải thích mới có liên

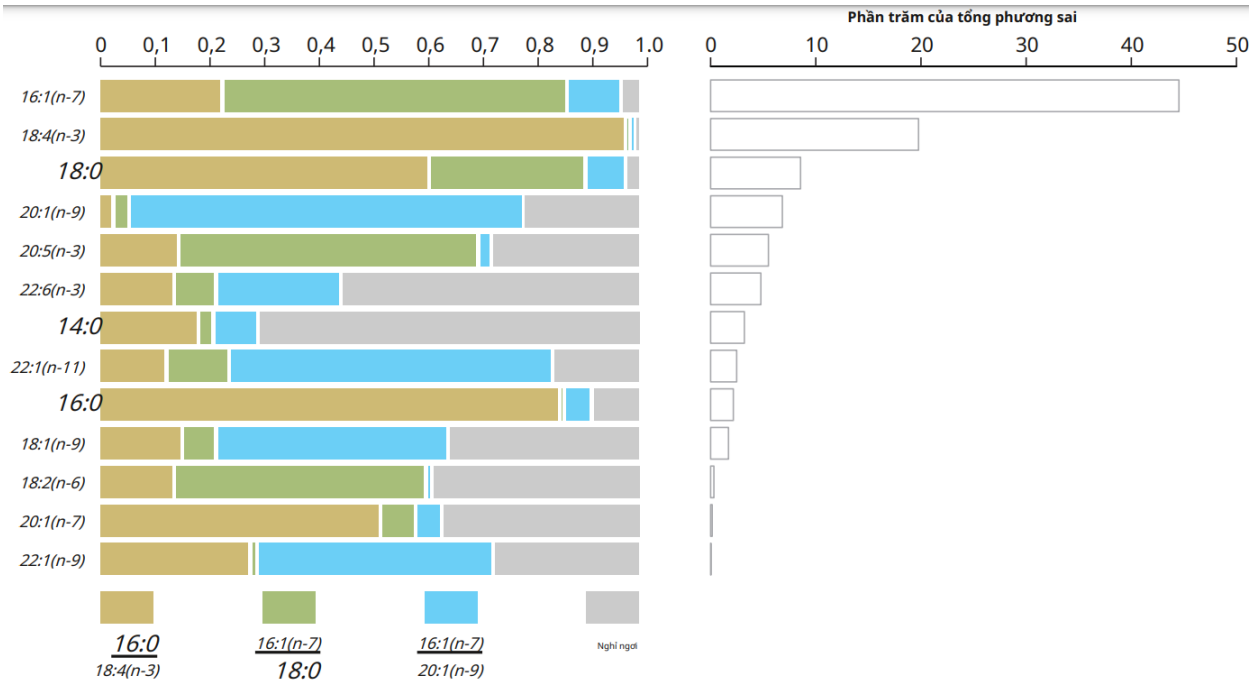
quan đến việc lựa chọn biến. Phân tích được sử dụng để xác định phương sai được giải thích được gọi là phân tích dự phòng (RDA), một khái quát hóa phân tích hồi quy cho dữ liệu đa phản hồi (Van Den Wollenberg, 1977; Gittins, 1985; Zuur và cộng sự, 2007)—chi tiết bổ sung được cung cấp trongPhụ lục bổ sung. Một lần nữa, số lượng CLR nhỏ hơn sẽ tạo thành tập hợp các biến phản hồi một cách thuận tiện, tương đương với việc sử dụng tất cả các LR.

LR đầu tiên nêu trên được giữ lại, sau đó bước tiếp theo là tìm LR nào khác bổ sung phương sai được giải thích bổ sung tối đa — quy trình từng bước được Greenacre (2019) giải thích chi tiết. LR tiếp theo được xác định là $(log(16:1(n-7)/18:0))$, giải thích thêm 18,2%, đưa phương sai được giải thích bởi hai LR này lên 83,8%. Việc thêm nhiều LR theo cách này sẽ mang lại phương sai được giải thích lên 100% khi $J1 = 12$ LR đã được nhập—trên thực tế, Hình 2c hiển thị 12 tỷ lệ được chọn dưới dạng biểu đồ được kết nối theo chu kỳ. Tuy nhiên, chỉ ba LR được chọn đầu tiên mang lại phương sai được giải thích lên tới hơn 90%, điều này có thể được coi là thỏa đáng, thay thế hiệu quả toàn bộ tập dữ liệu 13 phần chỉ bằng ba LR. Ba LR này là một phần của biểu đồ trong Hình 2c, kết nối năm FA được dán nhãn màu đỏ.

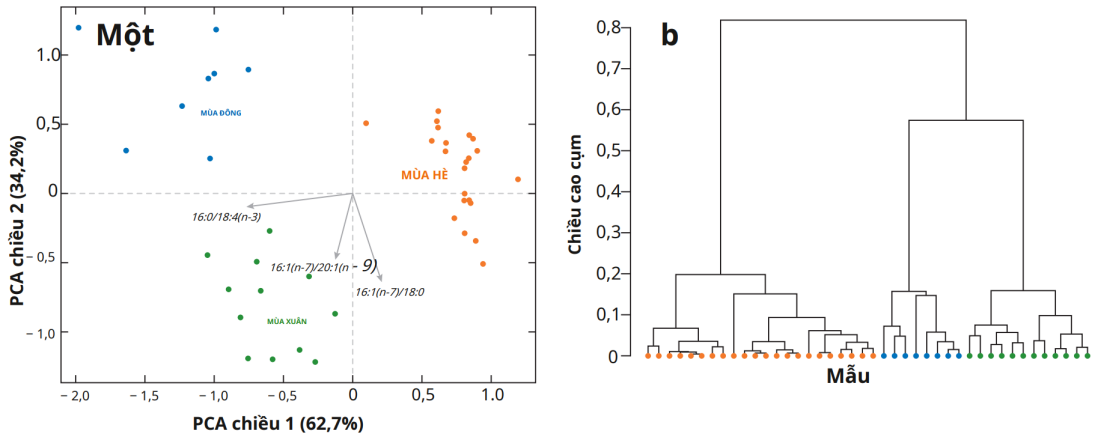
Ba LR được đưa ra trong Hình 5, điều này cũng cho thấy mức độ phương sai logarit của từng FA riêng lẻ (cụ thể là phương sai của CLR của chúng) mà mỗi trong số ba LR giải thích. Các FA được sắp xếp theo thứ tự giảm dần của phương sai logarit (biểu đồ thanh ở bảng bên phải hiển thị chúng dưới dạng phần trăm của tổng phương sai). Các thanh trong bảng điều khiển bên trái cho thấy phương sai logarit cao có hầu hết phương sai được giải thích bởi LR, với các phần phương sai không giải thích được tập trung chủ yếu ở FA có phương sai thấp

Chỉ riêng PCA của ba LR này, chỉ bao gồm năm phần (FA) và phân tích cụm mẫu chỉ dựa trên ba phần này được hiển thị trong Hình 6. Cả PCA và phân tích cụm đều cho thấy sự phân chia các mùa được cải thiện, mặc dù kiến thức về nhóm này đã chưa được tính đến trong phân tích. Phân tích cụm bây giờ hoàn toàn trùng khớp với

ba nhóm theo mùa.



Hình 1.7 Giải thích phương sai logarit của ba LR .



Hình 1.8 (Một) PCA của ba logarit theo cặp giải thích 90,9% tổng phương sai

Phương sai được giải thích bởi ba LR này là 90,9% tổng phương sai logarit. Chúng bao gồm năm phần, $16:0, 16:1(n-7), 18:0, 18:4(n-3), V20:1(n-9)$, có thể được sử dụng làm thành phần phụ gồm 5 phần. Việc sử dụng năm CLR của thành phần phụ này sẽ giải thích được 92,7% tổng phương sai, thêm 1,8 điểm phần trăm.

Điều này là do ba LR không kết nối tất cả năm phần của thành phần phụ, được hiển thị bằng màu đỏ trong Hình 2c—cần thêm một LR nữa để tạo một đồ thị liên thông không tuần hoàn của năm phần. Bất kỳ một kết nối nào liên kết chúng, ví dụ: logarit của $16:0$ liên quan đến $16:1(n-7)$, sẽ tạo kết nối, dẫn đến bốn LR giải thích 92,7% phương sai logarit, giống như năm CLR của thành phần phụ.

Như nhận xét cuối cùng cho phần này, tập hợp các LR được chọn theo từng bước có thể không phải lúc nào cũng là một lựa chọn tối ưu, trong trường hợp bộ dữ liệu này việc áp dụng được cho là phù hợp nhất để biểu diễn dữ liệu thành phần. Trên thực tế, ở mỗi bước có một số LR được chọn để tham gia rất gần với việc giải thích phương sai tối đa.

1.3 Các trường hợp cần giải quyết trong việc chuẩn hóa mô hình

1.3.1 Vấn đề của những số 0

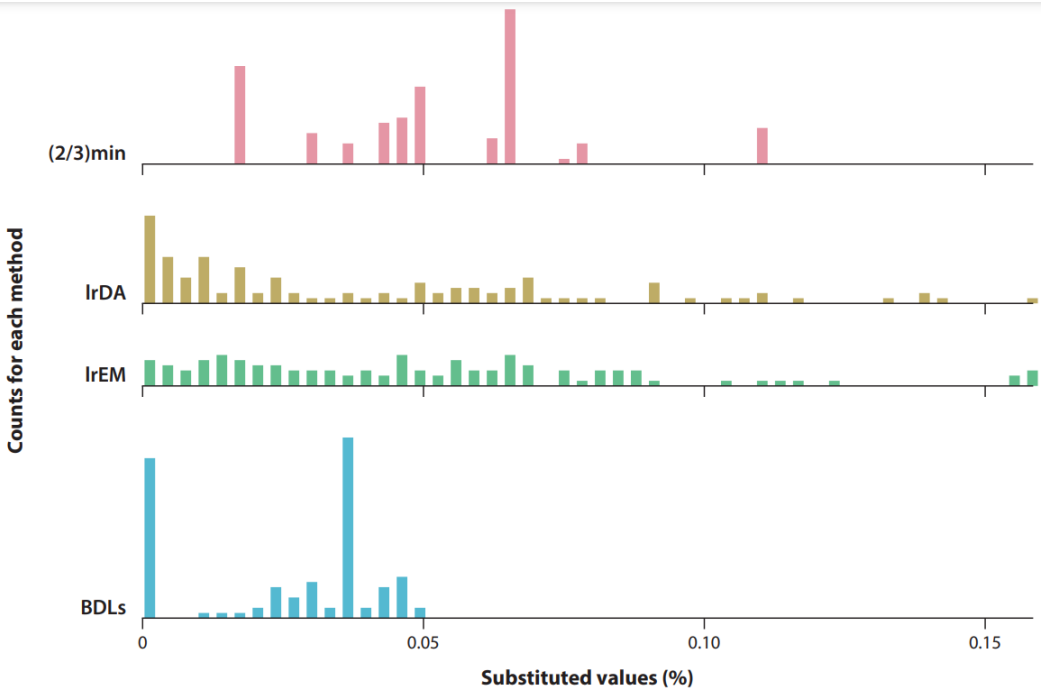
Vì phương pháp sử dụng logarit để tính toán cho nên vấn đề về sự hiện diện của các số 0 cũng như sự vắng mặt của một số giá trị trong một thành phần của dữ liệu đa hợp là vấn đề cần phải giải quyết. Cho đến nay, thành phần phụ gồm 13 phần của thành phần 40 phần ban đầu của FA đã được sử dụng, trong đó mỗi phần có tỷ lệ xuất hiện trung bình ít nhất là 0,01 (1%). Trong thành phần con này không có số 0 dữ liệu và do đó không có vấn đề gì với các phép biến đổi logarit khác nhau. Để phân tích bộ dữ liệu hoàn chỉnh gồm 40 phần, bao gồm chủ yếu là các FA hiếm hơn, cần phải đưa ra quyết định về 187 số 0 trong 42 phần này với 40 tập dữ liệu, chiếm khoảng 11% dữ liệu. Trong trường hợp này, các số 0 là do các giá trị nằm dưới giới hạn phát hiện và do đó được ghi là 0 (Palarea-Albaladejo và cộng sự, 2007). Một lựa chọn thường dùng là đưa ra quyết định dựa trên giả định về phân bố xác suất của các giá trị gần 0, ví dụ: phân bố tam giác từ 0 đến giá trị dương nhỏ nhất x phút, mang lại giá trị kỳ vọng bằng 2x phút(MartínFernández và cộng sự, 2003). Có nhiều thuật toán lặp được thiết kế để thay thế các số 0 với nhiều mức độ phức tạp khác nhau (ví dụ Martín-Fernández

và cộng sự, 2012), được tóm tắt bởi Filzmoser và cộng sự. (2018, chương 13).

Câu hỏi đặt ra là liệu phương pháp cụ thể được chọn có tạo ra bất kỳ sự khác biệt đáng kể nào đối với kết quả phân tích dữ liệu cuối cùng hay không. Vì các giá trị được thay thế sẽ là những số nhỏ (thêm một epsilon nhỏ vào các giá trị 0), tạo ra logarit âm hoặc dương lớn, nên tổng phương sai logarit của các ma trận quy nạp có thể được đánh giá qua các giá trị thay thế. Cụ thể hơn, sự thay đổi gây ra trong cấu trúc đa biến của dữ liệu có thể được đo bằng cách sử dụng tương quan Procrustes, vì cấu trúc khoảng cách của các mẫu là nền tảng cho tất cả các kết quả thu được sau đó.

Hình 1.9 hiển thị một số kết quả cho đơn giản 2x phút phương pháp cũng như ba phương pháp lặp thay thế. Một số phương pháp không thể được sử dụng do hai FA có lần lượt là 40 và 39 số 0 trong số 42 mẫu, vì vậy hai phần này đã bị loại bỏ trong phần này này, để lại 42 Ma trận với 108 số không (6,8% dữ liệu). Tổng phương sai logarit được tính theo đường chéo và ngoài đường chéo là mối tương quan Procrustes giữa các cấu hình logarit của các mẫu trong không gian đa biến. Các phương pháp lrDA (tăng dữ liệu logarit) và BDL (dưới giới hạn phát hiện) tạo ra tổng phương sai logarit lớn do số lượng lớn các giá trị quy định rất nhỏ gần bằng 0 (xem Hình 1.9) tạo logarit lớn

1.3.2 Định lượng các thành phần và sự chưa hợp lý các thành phần phụ



Hình 1.9 Biểu đồ của 108 lần thay thế số 0 được thực hiện bằng bốn thuật toán khác nhau.

ở giá trị tuyệt đối (ví dụ: 29 giá trị nhỏ hơn 0,001% được thay thế bằng phương pháp BDL). Sự phù hợp cao nhất giữa các cấu trúc đa biến là giữa phương pháp thay thế đơn giản nhất, (2/3) phút và thuật toán giảm thiểu kỳ vọng logarit lrEM (Tương quan Procrustes = 0,941). Một số mối tương quan khá thấp đối với loại so sánh này, trong đó các cấu hình phải phù hợp ở mức tương quan ít nhất là 0,9. Phải thừa nhận rằng những kết quả này chỉ áp dụng cho ví dụ cụ thể này, tuy nhiên điều này cho thấy phương pháp thay thế bằng 0 có thể có tác động mạnh mẽ đến cấu trúc của tập dữ liệu tổng hợp và ảnh hưởng đến phân tích tiếp theo của nó. Do đó, việc phân tích độ nhạy của kết quả phân tích thống kê bằng nhiều phương pháp thay thế là điều mong muốn.

1.4 Mô hình hồi quy cho dữ liệu đa hợp

Bây giờ chúng ta mô tả mô hình hồi quy cho CoDa. Mặc dù phân tích hồi quy là một trong những quy trình thống kê phổ biến và phát triển nhất, nhưng tài liệu về

dữ liệu thành phần CoDA kết hợp sử dụng phân tích hồi quy thì không nhiều.

Có ba loại mô hình tuyến tính liên quan đến các thành phần đã được định nghĩa đó là: mô hình với biến phụ thuộc là các biến thành phần, mô hình với biến độc lập là biến thành phần, và mô hình với cả biến cố định và biến độc lập đều là các biến thành phần. Phần tiếp theo sẽ xây dựng hệ thống mỗi mô hình hồi quy chỉ bằng các phép toán hình học và chỉ ra cách các mô hình này được biểu diễn trong phép biến đổi logarit đẳng cự. Trong phần phân tích hồi quy cho dữ liệu đa hợp được trình bày trong bài này sẽ tập trung vào mô hình loại 1 tức là các biến phụ thuộc là các biến thành phần.

1.4.1 Mô hình tuyến tính với biến phụ thuộc là biến thành phần

Một mô hình với biến phụ thuộc là biến thành phần giả định rằng một thành phần ngẫu nhiên Y là một hàm tuyến tính (theo nghĩa của hình học Aitchison) của một số biến ngẫu nhiên thực $X_0, X_1, ..., X_P$, điều này mang lại giá trị kỳ vọng cho một số thành phần có phân phối chuẩn,

$$\hat{Y} = \bigoplus_{i=0}^{P} X_i \odot b_i, \quad Y \sim \mathcal{N}_{\mathcal{S}^D}(\hat{Y}, \Sigma_\varepsilon), \tag{1.17}$$

trong đó $\mathcal{N}_{\mathcal{S}^D}(\hat{Y}, \Sigma_\varepsilon)$ đại diện cho phân phối chuẩn trên không gian đơn hình của Y, được tham số hóa theo vector trung bình thành phần và ma trận hiệp phương sai của thành phần ngẫu nhiên trong một biểu diễn ilr nào đó. Điều này phản ánh thực tế rằng phân phối chuẩn trên không gian đơn hình của một thành phần ngẫu nhiên tương ứng với phân phối chuẩn (thông thường) của biểu diễn ilr của nó. Mô hình hồi quy này hữu ích cho các biến giải thích thuộc loại định lượng (hồi quy), loại phân loại (ANOVA) hoặc kết hợp cả hai (ANCOVA). Lưu ý rằng có thể thiết lập mô hình hồi quy này cho dữ liệu thành phần theo nghĩa bình phương nhỏ nhất , không phụ thuộc vào giả định phân phối chuẩn. Tuy nhiên, giả định phân phối chuẩn là cần thiết trong bối cảnh kiểm định giả thuyết. Cụ thể, nó phục vụ cho việc suy ra phân phối của các thống kê kiểm định trong trường hợp hồi quy (bình phương nhỏ nhất) cổ điển và cũng

phục vụ như mô hình tham chiếu cho hồi quy bền vững robust regression.

Nếu một phép biến đổi logarit được áp dụng cho mô hình này, điều này sẽ dẫn đến một mô hình hồi quy tuyến tính đa biến thông thường trên các tọa độ:

$$\hat{Y}^* = \sum_{i=0}^{P} X_i \cdot b_i^*, \quad Y^* \sim \mathcal{N}^{D-1}(\hat{Y}^*, \Sigma_\varepsilon). \tag{9}$$

Các tham số của mô hình do đó là các độ dốc $b_0^*, b_1^*, ..., b_P^*$, và ma trận hiệp phương sai Σ_ε. Lưu ý rằng thường lấy $X_0 \equiv 1$ và sau đó b_0^* thay thế cho hệ số giao điểm của mô hình trong hệ tọa độ logarit đã chọn. Cụ thể trong (9) có lợi thế là có thể xử lý bằng phần mềm và các phương pháp giải thông thường. Khi đã có các ước lượng của các hệ số vector, chúng có thể được biến đổi ngược lại thành các hệ số thành phần, ví dụ $\hat{b}_i = \text{ilr}^{-1}(\hat{b}_i^*)$ nếu các phép tính được thực hiện trong các tọa độ ilr. Ngoài ra, các tọa độ ilr cũng có thể được chuyển đổi thành các hệ số clr với $\hat{b}_{clr_i} = V \cdot \hat{b}_i^*$.

Điều quan trọng cần nhấn mạnh là các dự đoán được cung cấp bởi mô hình hồi quy này là không thiên lệch về bất kỳ biểu diễn logarit nào, và về mặt thành phần gốc (8) với hình học Aitchison được thảo luận trong mục 1.1. Điều này trực tiếp từ tính đẳng cự của các phép biến đổi ilr hoặc clr. Nếu quan tâm đến việc hiểu các tính chất không thiên lệch của các dự đoán (8) với hình học Euclide thông thường của không gian đa biến thực $\mathbb{R}^D$, tức là về bản chất của giá trị kỳ vọng của $\hat{Y} - Y$, thì có thể sử dụng tích phân số của mô hình được giải thích bởi (8), cái cung cấp phân phối điều kiện của Y cho trước $\hat{Y}$ (Aitchison, 1986).

1.4.2 Ước lượng Bình phương nhỏ nhất cổ điển (LS)

Khi biến phụ thuộc là thành phần, chúng ta thu được các vector điểm số logarit quan sát được $y_1^*, ..., y_n^*$ có kích thước $D-1$, và các hệ số hồi quy được thu thập trong ma trận B^* kích thước $(D-1) \times (P+1)$. Cột đầu tiên của ma trận này đại diện cho vector tọa độ giao điểm b_0^*. Các cột còn lại có thể được liên kết với P biến đồng biến thực cố định. Các vector phần dư là $r_i^*(B^*) = y_i^* - B^* x_i$, với $i = 1, ..., n$.

Xét các ma trận của các biến giải thích và biến đáp ứng $X^* = [x_1^*, ..., x_n^*]$ và $Y^* = [y_1^*, ..., y_n^*]$ (mỗi hàng là một quan sát, các thành phần được biến đổi ilr), các ước lượng bình phương nhỏ nhất của các tham số mô hình là:

$$\mathbf{B}^* = [(X^*)^t \cdot X^*]^{-1} \cdot (X^*)^t \cdot Y^*$$

và

$$\Sigma_\varepsilon = \frac{1}{N-P} \sum_{i=1}^{n} r_i^*(B^*)^t \cdot r_i^*(B^*).$$

Cuối cùng, ma trận hiệp phương sai của $\mathbf{b}^*$ có thể được ước lượng là

$$\Sigma_b = \Sigma_\varepsilon \otimes [(X^*)^t \cdot X^*]^{-1},$$

trong đó $\otimes$ là tích Kronecker của hai ma trận, và $\mathbf{B}^*$ được xếp thành vector theo cột.

1.4.3 Ước lượng Robust MM trong hồi quy

Có nhiều đề xuất cho hồi quy bền vững có sẵn trong tài liệu (xem Maronna và cộng sự, 2006). Việc lựa chọn một ước lượng phù hợp phụ thuộc vào các tiêu chí khác nhau. Trước hết, ước lượng nên có các tính chất bền vững mong muốn, tức là bền vững trước mức độ nhiễu cao, đồng thời có hiệu quả thống kê cao. Các ước lượng MM cho hồi quy có điểm phá vỡ tối đa là 50% (tức là ít nhất 50% các mẫu bị nhiễm cần thiết để làm cho ước lượng trở nên vô dụng) và chúng có hiệu quả có thể điều chỉnh được. Mặc dù các ước lượng hồi quy khác cũng đạt được điểm gãy cao, như ước lượng hồi quy LTS, hiệu quả của chúng có thể khá thấp (Maronna và cộng sự 2006). Một tiêu chí khác cho việc lựa chọn là sự sẵn có của một triển khai phù hợp trong các gói phần mềm. Các ước lượng MM cho hồi quy có sẵn trong môi trường phần mềm R (R Development Core Team 2019) và Python.Đối với các trường hợp hồi quy Loại

1, các ước lượng hồi quy MM đa biến có thể được sử dụng như là các đối trọng so với của các ước lượng LS.

Ở đây, ta coi x_i là các biến phụ thuộc trong dữ liệu đa hợp, khi đó X_i thu được sau khi sử dụng phép logarit ilr tạo thành ma trận X thu được từ các biến đa hợp ban đầu x_i. Cùng với ký hiệu như trong mục 1.4.1, các ước lượng MM cho dữ liệu thành phần được định nghĩa như sau:

$$(\hat{\mathbf{B}}^*, \hat{\mathbf{C}}) = \arg\min_{\mathbf{B}} \sum_{i=1}^{n} \rho \left(\frac{\mathbf{r}_i^*(\mathbf{B})^t \mathbf{C}^{-1} \mathbf{r}_i^*(\mathbf{B})}{\hat{\sigma}} \right),$$

trong công thức trên $\mathbf{r}_i^*(\mathbf{B})^t$ là phần dư chẩn hóa cho qan sát thứ i, tính bằng $y_i - \mathbf{X}_i^T B$ với B là hệ số hồi quy, C^{-1} là ma trận nghịch đảo của ma trận covariance. với ước lượng độ lệch chuẩn $\hat{\sigma} := \det(\Sigma_S)^{1/(2D-2)}$, trong đó Σ_S được lấy từ một ước lượng S của hồi quy đa biến (xem Van Aelst và Willems 2013, để biết chi tiết). Ma trận hiệp phương sai của phần dư ước lượng sau đó được cho bởi $\Sigma_\varepsilon = \hat{\sigma}^2 C$.

CHƯƠNG 2. ỨNG DỤNG VỚI MÔ HÌNH

Từ các lý thuyết và phân tích ở phần 1, ta sẽ có các bước để tiến hành phân tích một bộ dữ liệu như sau:

- Làm sạch dữ liệu,
- Logarit hóa dữ liệu (bắt đầu tiến hành phân tích đa hợp),
- Phân tích các thành phần và cụm để đánh giá bộ dữ liệu,
- Sử dụng hồi quy tuyến tính để đưa ra các ý nghĩa mà bộ dữ liệu thể hiện.

Sau đó, ta sẽ đánh giá tác dụng của việc phân tích dữ liệu đa hợp bằng cách đánh giá 2 quá trình phân tích hồi quy tuyến tính không sử dụng và có sử dụng phân tích dữ liệu đa hợp.

2.1 Bộ dữ liệu

Bộ dữ liệu được tổng hợp từ kaggle thu thập nhằm để thống kê chất lượng không khí của Ấn Độ từ năm 2015 đến năm 2020, lấy ngữ cảnh là việc theo dõi chất lượng không khí đặc biệt ở những quốc gia đông dân là vô cùng cần thiết, do những ảnh hưởng của ô nhiễm không khí gia tăng đặc biệt gây nguy hại tới sức khỏe con người.

Bộ dữ liệu chứa dữ liệu chất lượng không khí và AQI (Chỉ số Chất lượng Không khí) ở mức hàng giờ và hàng ngày của các trạm khác nhau trên nhiều thành phố ở Ấn Độ.

Các thông tin được đề cập trong bộ dữ liệu gồm có:

StationId	mã số định danh cho các chạm đo không khí
Date	ngày xác nhận kết quả đo
PM2.5	Vật chất bụi mịn 2,5 micromet tính bằng ug / m3
PM10	Vật chất bụi mịn 10 micromet tính bằng ug / m3
NO	Nồng độ Nitric Oxide trong không khí tính bằng ug / m3
NO2	Nồng độ Nitric Dioxide trong không khí tính bằng ug / m3
NOx	Nồng độ tất cả Nitric x-oxide trong không khí tính bằng ug / m3
NH3	Nồng độ Ammonia trong không khí tính bằng ug / m3
CO	Nồng độ Carbon Monoxide trong không khí tính bằng ug / m3
SO2	Nồng độ Sulphur Dioxide trong không khí tính bằng ug / m3
O3	Nồng độ Ozone trong không khí tính bằng ug / m3
Benzene	Nồng độ Benzene trong không khí tính bằng ug / m3
Toluene	Nồng độ Toluene trong không khí tính bằng ug / m3
Xylene	Nồng độ Xylene trong không khí tính bằng ug / m3
AQI	chỉ số chất lượng không khí (Air Quality Index)
AQI_Bucket	nhóm chỉ số chất lượng không khí (Air Quality Index bucket)

Tiến hành làm sạch dữ liệu

Sau đây là mô tả một vài thông tin của các trường là các khí hoặc bụi được thu thập trong bộ dữ liệu:

	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI
count	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000	10314.000000
mean	52.482100	108.494813	12.204897	33.188329	29.911210	17.099812	0.699242	9.907130	32.224498	4.429789	12.120517	2.862333	119.847004
std	43.101142	67.881316	18.880759	22.909939	28.188032	12.737291	0.436917	7.926233	19.937162	13.419295	23.774658	7.048609	75.075472
min	1.090000	5.770000	0.020000	0.010000	0.020000	0.100000	0.000000	0.100000	0.030000	0.000000	0.000000	0.000000	16.000000
25%	25.740000	62.692500	2.650000	15.220000	12.680000	9.332500	0.420000	4.370000	18.380000	0.230000	1.390000	0.120000	72.000000
50%	43.400000	98.805000	6.270000	28.570000	22.500000	13.850000	0.630000	7.860000	28.240000	1.520000	5.010000	0.880000	104.000000
75%	66.210000	137.837500	13.430000	46.170000	36.675000	21.970000	0.900000	12.810000	41.410000	4.160000	12.467500	2.950000	139.000000
max	734.560000	830.100000	262.000000	254.780000	331.500000	269.930000	4.740000	67.260000	162.330000	165.410000	259.030000	133.600000	692.000000

Hình 2.1 Bảng thống kê mô tả các biến dữ liệu

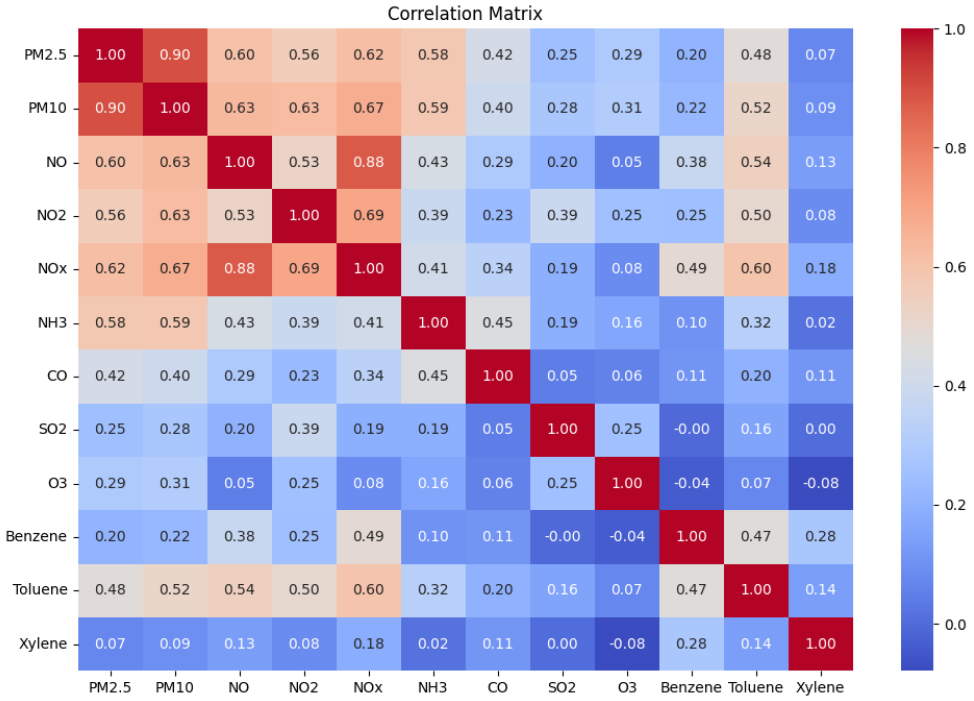
Bộ dữ liệu vẫn chứa các giá trị null (tức các giá trị chưa có thông tin) và sau đây

là thống kê tỷ lệ các giá trị null so với các giá trị đã được điền của mỗi biến dữ liệu:

	Proportion
Xylene	0.613220
PM10	0.377231
NH3	0.349734
Toluene	0.272290
Benzene	0.190410
AQI	0.158511
AQI_Bucket	0.158511
PM2.5	0.155701
NOx	0.141715
O3	0.136196
SO2	0.130507
NO2	0.121398
NO	0.121296
CO	0.069723
City	0.000000
Date	0.000000

Hình 2.2 Thông tin khái quát các biến dữ liệu

Giải thích cho sự mất mát quá nhiều của Xylene đó là việc đây là một chất khí dễ bay hơi, có nồng độ thấp và dễ bị gây nhiễu bởi các tương tác hóa học trong không khí (như các chất có tính oxi hóa mạnh).



Hình 2.3 Biểu đồ tương quan các biến dữ liệu

Biểu đồ tương quan nhằm thể hiện mối quan hệ tuyến tính giữa các chỉ số, từ đó đưa ra hướng phân tích và dự đoán cho bộ dữ liệu.

- PM2.5 và PM10: Tương quan cao (gần +1), cho thấy khi PM2.5 tăng, PM10 cũng tăng.

  Nguyên nhân: Tương quan dương cao giữa PM2.5 và PM10 có thể là do cả hai đều xuất phát từ cùng nguồn, chẳng hạn như khí thải từ giao thông, công nghiệp, và đốt cháy nhiên liệu.

- NO và NO2: Tương quan dương cao, chỉ ra rằng hai khí này có xu hướng tăng cùng nhau.

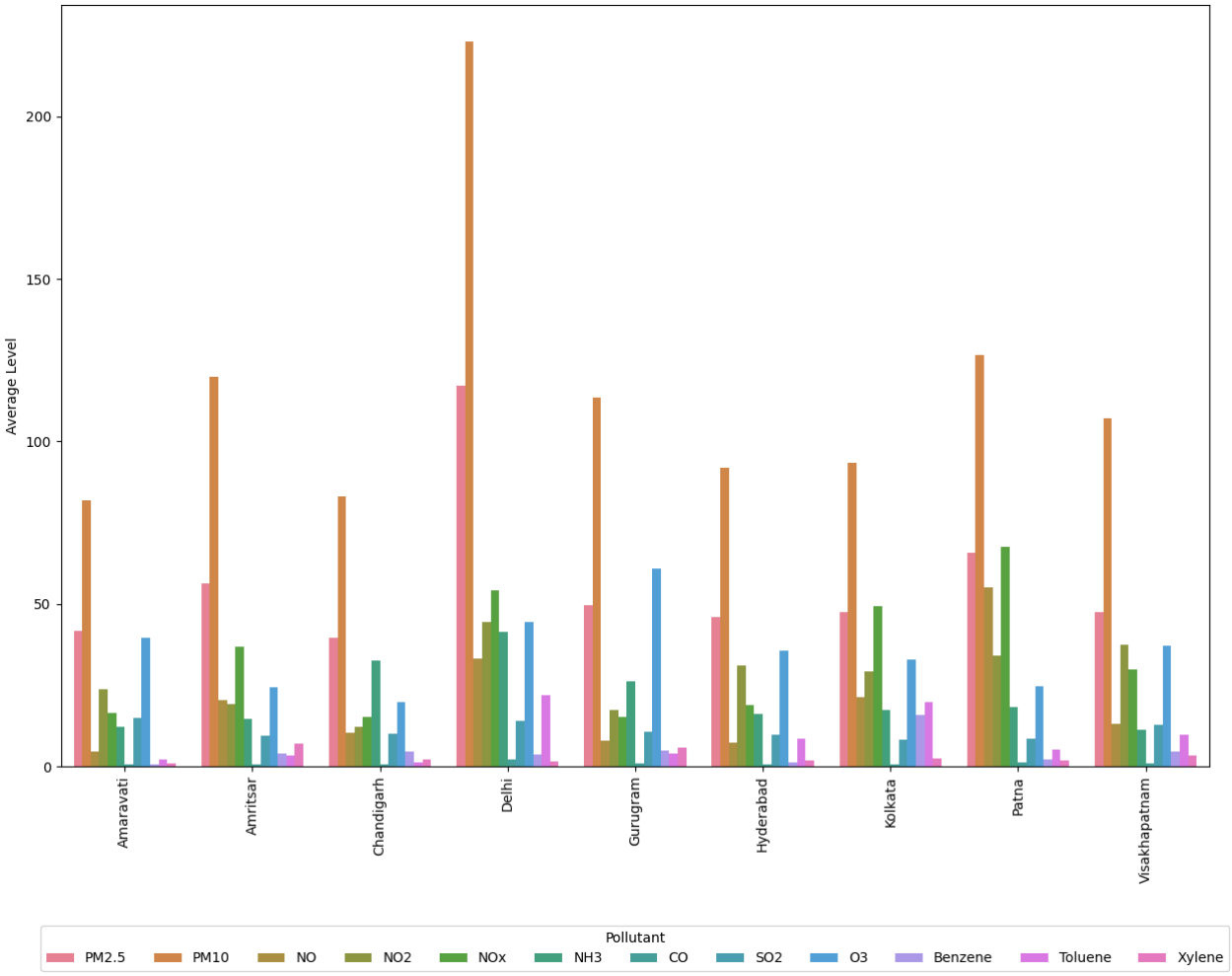
  NO và NO2 thường có tương quan dương cao vì NOx (gồm NO và NO2) là sản phẩm của quá trình đốt cháy nhiên liệu trong động cơ xe và công nghiệp.

- CO và SO2: Có thể có chỉ số tương quan dương thấp, cho thấy mức độ phụ thuộc không cao.

Độ tương quan dương của CO và SO2 do nguyên nhân hình thành của chúng, với một phần nguyên nhân chung là do các hoạt động công nghiệp.

- O3 và NOx: Có thể có tương quan âm, chỉ ra rằng khi một trong hai khí này tăng, khí kia giảm.

  Tương quan âm giữa O3 và NOx có thể được giải thích bởi quá trình hóa học trong khí quyển. Ozone (O3) thường được hình thành từ các tiền chất NOx và VOCs (Volatile Organic Compounds) dưới tác động của ánh sáng mặt trời. Khi nồng độ NOx rất cao, nó có thể phản ứng ngược lại và tiêu thụ ozone, dẫn đến tương quan âm.



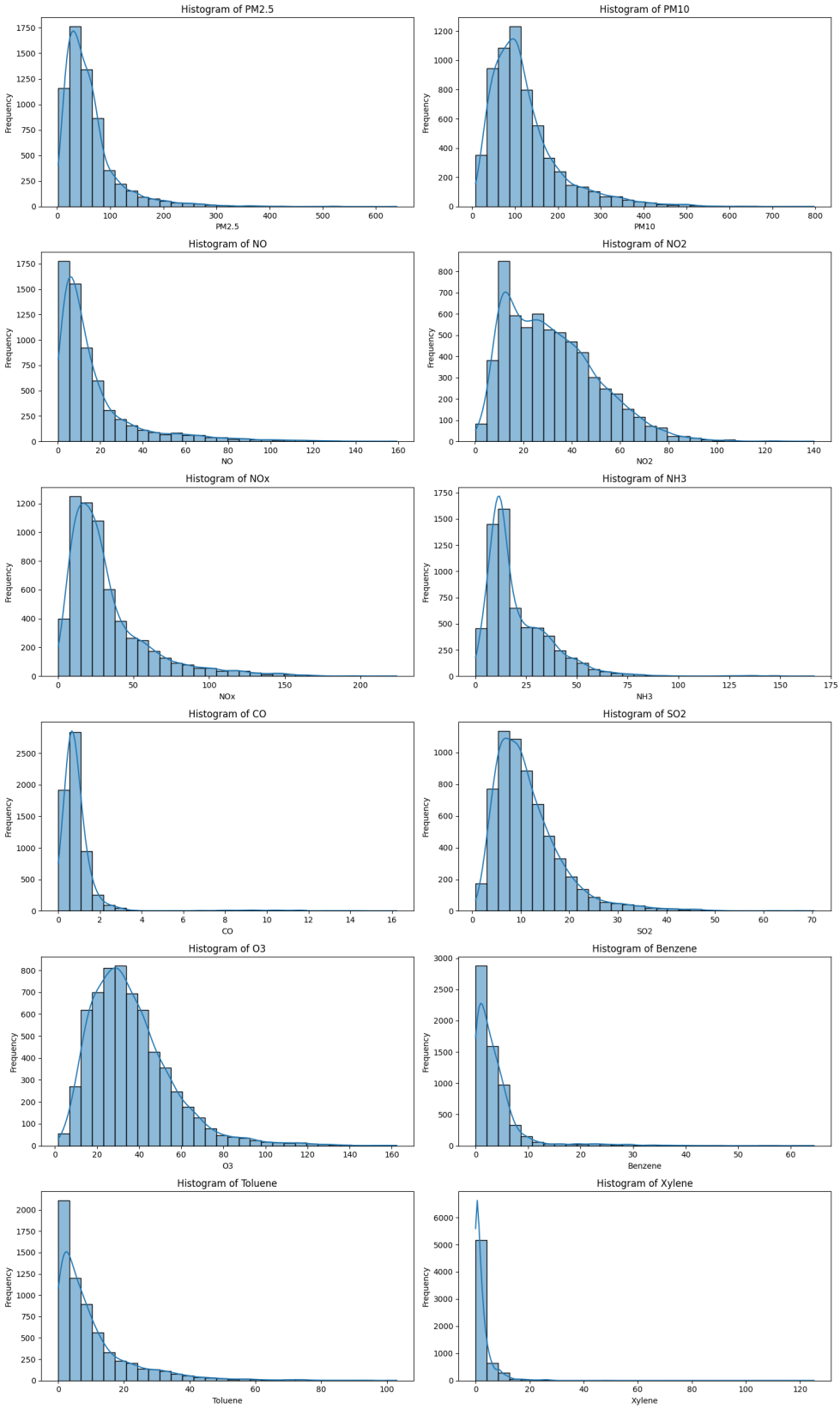
Hình 2.4 Biểu đồ mức độ trung bình của các chất ô nhiễm tại các thành phố trong dữ liệu

Biểu đồ trên thể hiện mức độ trung bình của các chất ô nhiễm (PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene) tại các thành phố trong bộ dữ liệu.

Bằng cách so sánh các thanh trên biểu đồ, ta có thể nhận ra sự khác biệt về mức độ ô nhiễm giữa các thành phố. Các chất ô nhiễm như PM2.5, PM10 thường có mức độ ô nhiễm cao hơn so với các chất khác như O3, Benzene, Toluene và Xylene. Bên cạnh đó, có thể xác định được xếp hạng ba thành phố có mức ô nhiễm cao nhất Ấn Độ trong bộ dữ liệu bao gồm Dehli, Patna, Amritsar. Nguyên nhân của thứ hạng của

các thành phố này có thể kể đến như:

- Giao thông đông đúc: Mật độ giao thông lớn với tình trạng tắc đường thường xuyên dẫn đến việc thải ra một lượng lớn khí thải ô nhiễm.
- Công nghiệp và sản xuất: Các khu công nghiệp lớn với các hoạt động sản xuất tạo ra lượng lớn khí thải và bụi mịn.
- Sự phát triển nhanh chóng và quản lý môi trường không hiệu quả: Sự phát triển nhanh chóng của dân số và kinh tế trong các thành phố này thường đi kèm với việc quản lý môi trường không hiệu quả, làm gia tăng vấn đề về ô nhiễm không khí...



Hình 2.5 Biểu đồ Histogram các biến dữ liệu

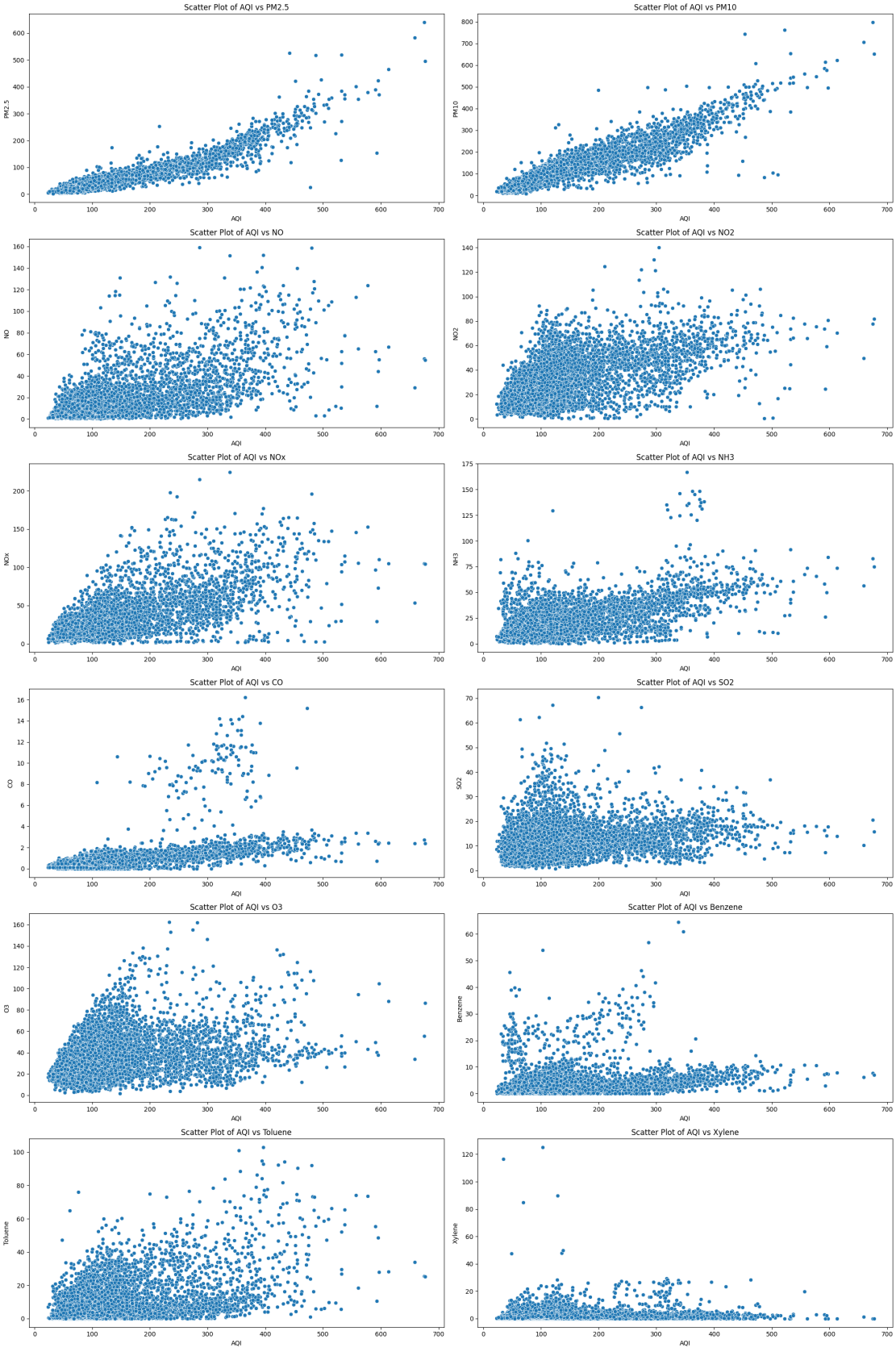
Các histogram giúp chúng ta hiểu rõ hơn về phân bố của từng biến trong bộ dữ liệu chất lượng không khí. Chúng cung cấp thông tin về dạng phân bố, giá trị ngoại lệ, và tần suất xuất hiện của các giá trị.

- Dạng phân bố: các histogram trong hình trên cho thấy hầu hết dữ liệu có phân bố không chuẩn. Ví dụ, histogram của PM2.5 có một đỉnh cao ở phía bên trái và một đuôi dài bên phải, điều đó cho thấy phân bố lệch phải. Nguyên nhân bởi một số lý do phổ biến sau: Nguồn gốc và phát thải của các chất gây ô nhiễm không đồng đều; Điều kiện thời tiết và khí hậu; Đặc điểm địa lý và nhân khẩu học; Sự can thiệp của con người.

- Giá trị ngoại lệ: Các giá trị ngoại lệ có thể được nhận diện qua các cột nhỏ hoặc cô lập ở hai đầu của histogram. Ví dụ: Nếu có các cột đơn lẻ cao ở phía ngoài của phân bố NO2, điều đó có thể chỉ ra các giá trị ngoại lệ trong dữ liệu NO2.

- Tần suất xuất hiện: Chiều cao của các cột trong histogram biểu thị tần suất xuất hiện của các giá trị trong khoảng đó. Ví dụ: Histogram của SO2 có thể cho thấy rằng phần lớn dữ liệu nằm trong khoảng nồng độ thấp, với chỉ một vài giá trị ở nồng độ cao hơn.

- So sánh giữa các biến: Bằng cách so sánh các histogram của các biến khác nhau, ta có thể nhận biết được biến nào có mức độ ô nhiễm cao hơn và biến nào ít biến động hơn. Ví dụ: Nếu histogram của PM2.5 có giá trị cao hơn nhiều so với PM10, điều này cho thấy rằng nồng độ PM2.5 thường cao hơn PM10.



Hình 2.6 Biểu đồ Scatter plot các vật chất với AQI trong bộ dữ liệu

Biểu đồ Scatter Plot (biểu đồ phân tán) là một công cụ mạnh mẽ để trực quan hóa và phân tích mối quan hệ giữa hai biến. Trong ngữ cảnh của bộ dữ liệu chất lượng không khí, việc sử dụng Scatter Plot để so sánh các biến như PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, và Xylene với AQI (Air Quality Index - Chỉ số chất lượng không khí) mang lại nhiều thông tin quan trọng.

- Mối quan hệ giữa các biến: Trong các biểu đồ trên, biến PM2.5 và PM10 là hai biến thể hiện rõ nhất mối quan hệ tuyến tính dương với AQI (điểm dữ liệu tập trung quanh một đường dốc lên). Theo sau đó là NO, NO2, NOx, SO2, O3 có mối quan hệ tuyến tính với AQI. CO, NH3, Benzene, Toluene, Xylene có mối quan hệ không rõ ràng

- Xác định độ ảnh hưởng của các biến: Vì PM2.5 có mối quan hệ mạnh với AQI, điều đó cho thấy PM2.5 là một trong những yếu tố chính ảnh hưởng đến AQI. Tương tự với biến PM10.

  NO có ảnh hưởng đến AQI nhưng không nhiều do NO thường xuyên chuyển hóa thành NO2, mối quan hệ có thể không mạnh như NO2.

  NO2 có ảnh hưởng lớn đến AQI, bởi lẽ NO2 là một chất ô nhiễm chính gây ra các vấn đề về sức khỏe và ô nhiễm không khí.

  NOx là một yếu tố quan trọng ảnh hưởng đến AQI. NOx là tiền chất của ozone tầng mặt đất và các hạt bụi mịn.

  SO2 là một chất ô nhiễm có thể gây ra các vấn đề về hô hấp và mưa axit cũng thể hiện ảnh hưởng với AQI.

  Ozone tầng mặt đất là một chất ô nhiễm nguy hiểm, và mối quan hệ giữa O3 và AQI thường rất mạnh. Một xu hướng đi lên rõ ràng cho thấy O3 là một yếu tố quan trọng ảnh hưởng đến AQI.

  Mối quan hệ không rõ ràng của NH3 có thể cho thấy đây không phải là yếu tố chính ảnh hưởng đến AQI. Tuy nhiên, NH3 có thể tương tác với các chất khác để

tạo thành các hạt bụi mịn.

Bên cạnh đó, dù các điểm dữ liệu cho thấy một mối quan hệ tuyến tính, điều này cho thấy rằng nồng độ CO có ảnh hưởng đến AQI. Tuy nhiên, nếu mối quan hệ không rõ ràng, điều này có thể chỉ ra rằng CO không phải là yếu tố chính ảnh hưởng đến AQI.

Benzene không phải là yếu tố chính ảnh hưởng đến AQI. Tuy nhiên, Benzene là một hợp chất hữu cơ dễ bay hơi và có thể gây ung thư.

Toluene không phải là yếu tố chính ảnh hưởng đến AQI, mặc dù có thể gây các vấn đề về thần kinh và hô hấp.

Xylene không phải là yếu tố chính ảnh hưởng đến AQI. Xylene có thể gây kích ứng mắt, mũi và cổ họng.

- Phát hiện giá trị ngoại lệ: Ví dụ, một vài điểm dữ liệu với nồng độ CO rất cao nhưng AQI lại thấp có thể gợi ý về một sự cố đo đạc hoặc một sự kiện cụ thể.

2.2 Sử dụng phân tích dữ liệu thành phần

Ta thay thế các giá trị bằng 0 bằng cách thêm vào một giá trị epsilon nhỏ (ở đây đặt epsilon = 1e-9) để tránh việc các dữ liệu bằng 0 sẽ ảnh hưởng tới quá trình loga hóa.

Sau đó ta đưa vào trong mô hình dữ liệu đa hợp và tiến hành chuyển hóa bằng CLR và thu được ma trận mới có dạng:

	0	1	2	3	4	5	6	7	8	9	10	11
0	2.701882	3.126812	-1.332850	1.322932	0.794058	0.674618	-3.817757	1.026430	3.147402	-3.306931	0.174309	-4.510904
1	2.538639	3.038113	-1.591053	1.435932	0.875835	0.508036	-3.788277	1.472190	2.943763	-3.336292	0.251007	-4.347893
2	2.398767	2.820474	-0.200099	1.341968	0.993364	0.470834	-4.294444	1.427060	2.629632	-3.325043	-0.055081	-4.207432
3	2.411851	2.895415	-0.809833	1.584861	1.083961	0.685526	-4.157786	1.194598	3.178717	-3.521797	-0.136411	-4.409101
4	2.464458	2.924826	-0.164303	1.325438	0.990086	0.686812	-3.651296	0.537412	2.879400	-3.379362	-0.269026	-4.344443

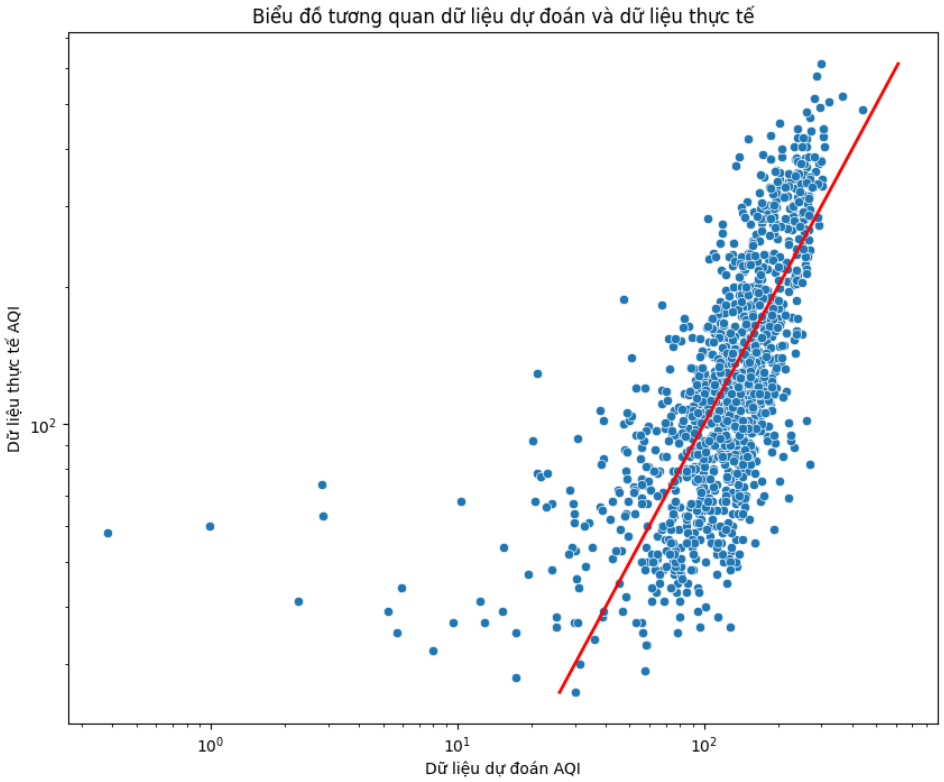
Hình 2.7 Một vài biến dữ liệu sau khi CRL

Từ ảnh ta thấy được rằng các biến dữ liệu đã được thay đổi từ dữ liệu thuần túy

(chỉ thể hiện bản thân giá trị thông tin dữ liệu đó) sang dữ liệu quan hệ giữa các biến và trung bình hình học của chúng. Ví dụ có thể thấy nếu PM mang giá trị dương tức là nếu PM2.5 tăng sẽ khiến cho trung bình hình học của trung bình hình học các biến số tăng, ngược lại đó là Xylene khi hầu hết đều mang giá trị âm.

2.3 Hồi quy mạnh RLM

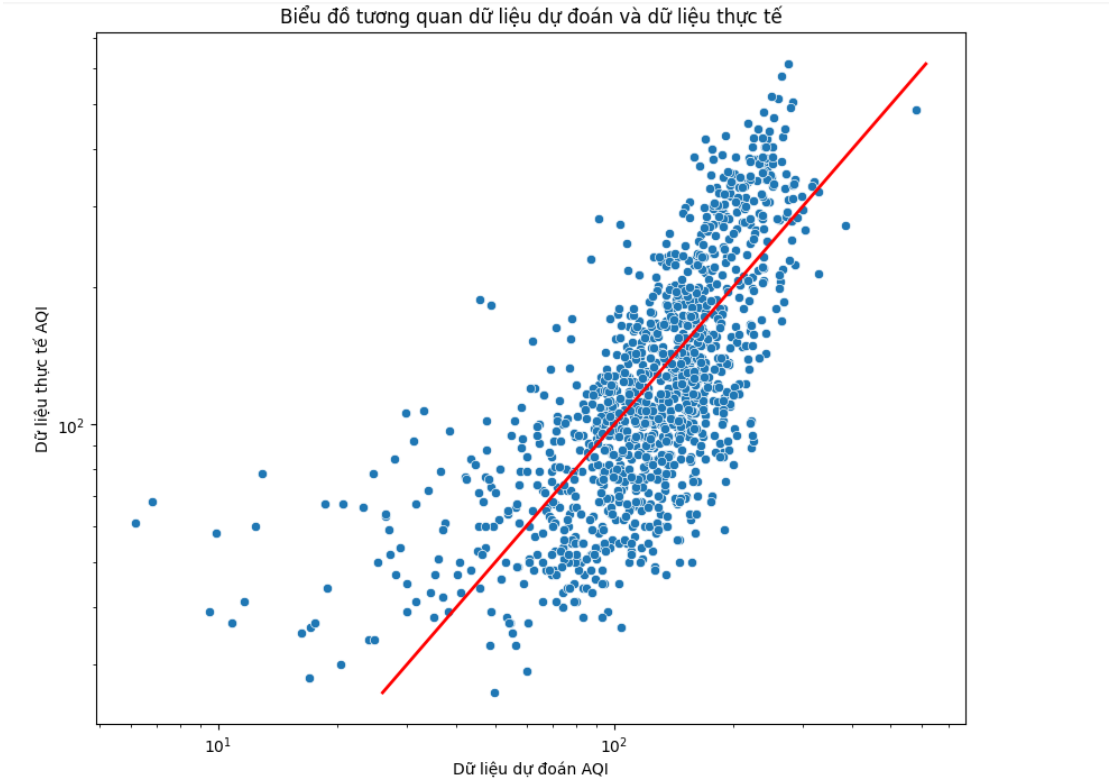
Sau khi có được dữ liệu từ việc phân tích CLR, ta đưa ma trận vào trong mô hình hồi quy tuyến tính. Kết quả của quá trình này ta thu được biểu đồ so sánh giữa bộ dữ liệu thực tế và bộ dữ liệu dự đoán (khi sử dụng hồi quy tuyến tính) như sau:



Hình 2.8 biểu đồ tương quan giá trị thực tế và giá trị dự đoán

với chỉ số giá trị trung bình bình phương dùng để đo lường độ lớn sai số giữa giá trị dự đoán và thực tế của mô hình khi có sử dụng phân tích đa hợp RMSE= 60.914120609634 66. Còn nếu không sử dụng phân tích đa hợp với dữ liệu ban đầu thì chỉ số RMSE thu được là RMSE=64.080002090968 Với biểu đồ đồ so sánh giữa

bộ dữ liệu thực tế và bộ dữ liệu dự đoán như sau:



Hình 2.9 biểu đồ tương quan giá trị thực tế và giá trị dự đoán khi không sử dụng CLR

KẾT LUẬN

Kết quả đạt được: Đồ án với mục tiêu nghiên cứu và ứng dụng phân tích dữ liệu đa hợp. Qua quá trình nghiên cứu và thực hiện đồ án, bản thân em đã rút ra được các kết quả đó là tầm quan trọng và lợi ích của việc sử dụng các phương pháp phân tích dữ liệu đa hợp trong xử lý dữ liệu có dạng tổng, đã áp dụng các biến đổi tỷ lệ log để cải thiện độ chính xác và độ tin cậy của các phân tích dữ liệu. Đồng thời, việc triển khai các phương pháp này trong một số lĩnh vực cụ thể, như phân tích kinh doanh và nghiên cứu địa chất học kết hợp với việc sử dụng mô hình hồi quy đã đem lại những kết quả đáng khích lệ, khẳng định tiềm năng ứng dụng rộng rãi của chúng.

Hướng phát triển: Nghiên cứu trong đồ án vẫn chỉ nêu ra những khái niệm cơ bản và ứng dụng trong 2 bài toán cụ thể. Vì vậy có thể cải thiện thêm về các phương diện như:

- Nghiên cứu sâu hơn về các biến đổi khác ngoài tỷ lệ log để tìm ra các phương pháp phù hợp hơn cho từng loại dữ liệu cụ thể.
- Kết hợp với nhiều ứng dụng thống kê và mô hình học máy khác như phân cụm.
- Ứng dụng các phương pháp này trong các lĩnh vực khác như y tế, tài chính, khoa học môi trường, sinh thái học, nghiên cứu địa lý dân cư và nghiên cứu vi sinh vật.
- Phát triển các công cụ phần mềm và hệ thống tự động hóa để hỗ trợ phân tích dữ liệu đa hợp.

TÀI LIỆU THAM KHẢO

[1] Aitchison, J. (1981). A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, 13, 175-189.

[2] Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J R Stat Soc Ser B*, 44, 139–77.

[3] Aitchison, J. (1986). The Statistical Analysis of Compositional Data. *Chapman & Hall, London*.

[4] Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 51(4), 375-392.

[5] Buccianti, A. (2018). Water chemistry: are new challenges possible from CoDa (compositional data analysis) point of view?. *Handbook of Mathematical Geosciences: Fifty Years of IAMG*, 299-311.

[6] Buccianti, A., & Grunsky, E. (2014). Compositional data analysis in geochemistry: are we sure to see what really occurs during natural processes?. *Journal of Geochemical Exploration*, 141, 1-5.

[7] Dumuid, D., Pedišić, Ž., Palarea-Albaladejo, J., Martín-Fernández, J. A., Hron, K., & Olds, T. (2021). Compositional data analysis in time-use epidemiology. *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*. Cham: Springer International Publishing, 383-404.

[8] Ebrahimi, P., Albanese, S., Esposito, L., Zuzolo, D., & Cicchella, D. (2021). Coupling compositional data analysis (CoDA) with hierarchical cluster analysis (HCA) for preliminary understanding of the dynamics of a complex water distribution system: The Naples (South Italy) case study. *Environmental Science: Water Research & Technology*, 7(6), 1060-1077.

[9] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3), 279-300.

[10] Filzmoser P, Hron K, Templ M., (2018). Applied Compositional Data Analysis Oxford, *UK: Oxford Univ. Press*.

[11] Gittins R. (1985). Canonical Analysis: A Review with Applications in Ecology Berlin: Springer-Verlag.

[12] Greenacre M. (2010). Biplots in Practice Bilbao. *Spain: BBVA Found* https://www.multivariatestatistics.org.

[13] Greenacre, M. (2018). Compositional data analysis in practice. *Chapman and Hall/CRC*.

[14] Greenacre, M. (2019). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*, 51(5), 649-682.

[15] 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.

[16] Greenacre, M., Martínez-Álvaro, M., & Blasco, A. (2021). Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. *Frontiers in Microbiology*, 12, 727398.

[17] Grunsky, E. C., Mueller, U. A., & Corrigan, D. (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: applications for predictive geological mapping. *Journal of Geochemical Exploration*, 141, 15-41.

[18] Leite, M. L. C. (2019). Compositional data analysis as an alternative paradigm for nutritional studies. *Clinical nutrition ESPEN*, 33, 207-212.

[19] Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). Robust statistics: theory and methods (with R). *John Wiley & Sons*.

[20] Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9), 2688-2704.

[21] McKinley, J. M., Mueller, U., Atkinson, P. M., Ofterdinger, U., Jackson, C., Cox, S. F., ... & Pawlowsky-Glahn, V. (2020). Investigating the influence of environmental factors on the incidence of renal disease with compositional data analysis using balances. *Applied Computing and Geosciences*, 6, 100024, doi:10.1016/j.acags.2020.100024.

[22] Palarea-Albaladejo, J., Martín-Fernández, J. A., & Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, 39, 625-645.

[23] Tepanosyan, G., Sahakyan, L., Maghakyan, N., & Saghatelyan, A. (2020). Combination of compositional data analysis and machine learning approaches to identify sources and geochemical associations of potentially toxic elements in soil and assess the associated human health risk in a mining city. *Environmental Pollution*, 261, 114210.

[24] Van Den Wollenberg, A. L. (1977). Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*, 42(2), 207-219.

[25] Verswijveren, S. J., Lamb, K. E., Martín-Fernández, J. A., Winkler, E., Leech, R. M., Timperio, A., ... & Ridgers, N. D. (2022). Using compositional data analysis to explore accumulation of sedentary behavior, physical activity and youth health. *Journal of Sport and Health Science*, 11(2), 234-243.

[26] Wei, Y., Wang, Z., Wang, H., & Li, Y. (2021). Compositional data techniques for forecasting dynamic change in China's energy consumption structure by 2020 and 2030. *Journal of Cleaner Production*, 284, 124702, doi:10.1016/j.jclepro.2020.124702.

[27] Zhang, S. W., Shen, C. Y., Chen, X. Y., Ye, H. C., Huang, Y. F., & Shuang, L. A. I. (2013). Spatial interpolation of soil texture using compositional kriging and regression kriging with consideration of the characteristics of compositional data and environment variables. *Journal of Integrative Agriculture*, 12(9), 1673–1683.

[28] Zheng, C., Liu, P., Luo, X., Wen, M., Huang, W., Liu, G., ... & Albanese, S. (2021). Application of compositional data analysis in geochemical exploration for concealed deposits: A case study of Ashele copper-zinc deposit, Xinjiang, China. *Applied Geochemistry*, 130, 104997, doi:10.1016/j.apgeochem.2021.104997.

[29] Zuur, A. F., Ieno, E. N., & Smith, G. M. (2007). Analysing ecological data (Vol. 680). *New York: Springer*.