

PHƯƠNG PHÁP PHÂN CỤM TRONG PHÂN TÍCH TẦN SUẤT MƯA VÙNG, ÁP DỤNG CHO KHU VỰC MIỀN TRUNG VÀ TÂY NGUYÊN

CLUSTER METHOD IN REGIONAL RAINFALL FREQUENCY ANALYSIS: A CASE STUDY IN THE CENTRAL AND HIGHLANDS

Nguyễn Chí Công

Trường Đại học Bách khoa, Đại học Đà Nẵng; chicongbkdn@gmail.com

Tóm tắt - Phân tích tần suất mưa vùng ngày càng được sử dụng phổ biến trong thủy văn, bởi vì cách tiếp cận này cho phép làm lớn kích thước mẫu thống kê cho từng trạm đo trong vùng, và từ đó tăng sự chắc chắn của suy luận thống kê. Tuy nhiên, yêu cầu mẫu dữ liệu vùng phải là đồng nhất. Khi phân tích với mẫu dữ liệu vùng với số lượng trạm đo lớn và trên diện rộng, thì mẫu dữ liệu này thường khó có thể đồng nhất. Do đó, cần tiến hành phân chia vùng nghiên cứu thành các tiểu vùng, sao cho mẫu dữ liệu mỗi vùng là đồng nhất. Nghiên cứu này trình bày hai phương pháp phân cụm trong phân tích tần suất mưa vùng và bước đầu áp dụng cho dữ liệu của 75 trạm đo mưa cho khu vực miền Trung và Tây Nguyên, Việt Nam. Kết quả nghiên cứu làm cơ sở cho việc xây dựng bản đồ mưa cực hạn cho khu vực này tiếp theo và xây dựng quy trình phân vùng đồng nhất, sử dụng trong phân tích tần suất mưa vùng.

Từ khóa - phân tích Cluster; mưa vùng; phân tích tần suất; vùng đồng nhất; miền Trung –Tây Nguyên.

Abstract - Regional rainfall frequency analysis is increasingly used in hydrology, because this approach allows us to make sample size bigger for each station in the region and thus increases the certainty of statistics probability. However, in order to apply this approach, the homogeneous sampling is required. The fact that when analysis uses regional sample data of a large quantity of gauged rainfall and a large area, it is often difficult to have a homogeneous sample. Therefore, it is necessary to divide the study regional into several sub-regions so that the sample of each sub-region can be homogeneous. The study presents two cluster methods in regional rainfall frequency analysis and is initially applied to 75 gauged rainfalls for the central and highland regions. The results form a basis for development of the extreme rainfall map for this region.

Key words - Cluster analysis; regional rainfall; frequency analysis; homogeneous region; the Central-The Highlands.

1. Đặt vấn đề

Trong thủy văn công trình, phân tích tần suất mưa thiết kế là thực sự cần thiết. Theo tiêu chuẩn thiết kế thì các tần suất thiết kế công trình thường nằm ở vùng đuôi của đường cong tần suất. Tuy nhiên với số năm quan trắc ngắn của mỗi trạm đo, sẽ dẫn đến sự không chắc chắn của giá trị suy luận ứng với tần suất thiết kế. Để khắc phục hạn chế này, phương pháp phân tích tần suất vùng (RFA-Regional Frequency Analysis) đã được áp dụng rộng rãi trong những năm gần đây, tiêu biểu như [5], [6], [7].

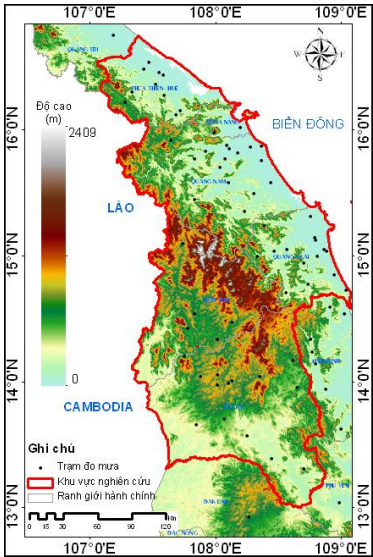
Trong RFA, có 2 bước cơ bản là (i) phân chia vùng đồng nhất và (ii) phân tích tần suất vùng sau khi được phân chia [1]. Nghiên cứu này giới thiệu các phương pháp phân chia vùng đồng nhất và điều kiện chấp nhận vùng đồng nhất. Phạm vi nghiên cứu là các tỉnh thuộc khu vực Miền Trung và Tây Nguyên (MT-TN). Đây là khu vực có chế độ mưa khá phức tạp và lượng mưa của một số trạm là lớn nhất trong nước [7]. Các nghiên cứu trong nước về vấn đề này đã chỉ ra rằng khi thực hiện RFA cho từng tiểu vùng tỉnh Quảng Nam thì mẫu dữ liệu mưa hoàn toàn đồng nhất cho 16 trạm đo mưa [7] và không đồng nhất cho 26 trạm đo mưa thuộc tỉnh Gia Lai [6]. Với cách tiếp cận đơn lẻ theo từng tiểu vùng này thì mẫu dữ liệu mưa vùng sẽ có thể chưa đại diện cho mưa vùng và khi không đồng nhất sẽ bị chia nhỏ tiểu vùng, điều này sẽ làm giảm kích thước mẫu dữ liệu mưa và dẫn đến độ tin cậy ước tính tần suất thiết kế là chưa cao. Để khắc phục điều này nghiên cứu xem xét số lượng trạm đo mưa trên diện rộng, đại diện cho khu vực MT-TN với tổng số lượng trạm đo mưa lên đến 75 trạm và số năm đo đạc liên tục trung bình là 31 năm.

Để phân chia vùng có N trạm nhưng không đồng nhất thành N_k tiểu vùng (K chính là số lượng tiểu vùng, $K = 2$ hoặc 3 ...), nghiên cứu sử dụng phương pháp phân cụm (phân tích Cluster - Cluster analysis) để thể hiện quan hệ mật thiết giữa các trạm trong một tiểu vùng. Phương pháp phân cụm bao gồm: (i) phương pháp không thứ bậc (K-Means) [3], [5] và (ii) phương pháp thứ bậc (Ward) [1], [4]. Tiếp theo, mẫu dữ liệu của mỗi tiểu vùng K sẽ được kiểm tra tính đồng nhất thông qua phương pháp của Hosking và Wallis (1997) [1], [3], [4], [5], [6], [7]. Mục tiêu của nghiên cứu này là so sánh kết quả của 2 phương pháp phân vùng K-Mean và Ward cho mẫu dữ liệu mưa ngày lớn nhất (NLN) với các thời đoạn mưa tính toán 1, 3, 5 và 7 ngày của 75 trạm đo mưa trên khu vực MT-TN.

2. Giới thiệu vùng nghiên cứu và dữ liệu

2.1. Vùng nghiên cứu

Vùng nghiên cứu bao gồm 6 tỉnh thuộc khu vực MT-TN gồm: Thừa Thiên Huế, Đà Nẵng, Quảng Nam, Quảng Ngãi, Kun Tom và Gia Lai (Hình 1). Đây là các tỉnh có địa hình khá phức tạp, do dãy Trường Sơn chia cắt khu vực miền Trung và Tây Nguyên theo hướng Đông Nam. Ở khu vực miền Trung, các dãy núi cao chia cắt ranh giới các tỉnh theo hướng Đông Bắc như: dãy Bạch Mã, dãy Ngọc Linh (Hình 1). Hàng năm các hoạt động gió Đông Bắc từ tháng 11 đến tháng 1, áp thấp nhiệt đới ở biển Đông từ tháng 9 đến tháng 12 và gió Tây Nam từ tháng 5 đến tháng 9, kết hợp với yếu tố địa hình này đã tạo nên những trận mưa lớn gây lũ lụt trên các lưu vực sông trong khu vực.



Hình 1. Địa hình và trạm đo mưa vùng nghiên cứu

2.2. Dữ liệu

Trong RFA, mẫu số liệu thực đo của các trạm đóng vai trò rất quan trọng. Theo đó, mẫu số liệu đo phải thỏa mãn các điều kiện: tính đại diện về không gian, thời gian đo phải liên tục và đủ dài (ít nhất trên 15 năm), chất lượng đo phải tin cậy và đặc biệt là mẫu dữ liệu vùng phải đồng nhất. Vùng nghiên cứu có số liệu đo rất đa dạng như: trạm đo truyền thống, trạm đo tự động. Các trạm này có thể đặt tại các trạm khí tượng thủy văn hoặc tại các hồ chứa nước hoặc trạm đo nhân dân. Sự phân bố các trạm đo không đồng đều, vùng đồng bằng có mật độ trạm dày, còn vùng núi có mật độ trạm thưa, đặc biệt vùng tiếp giáp với 2 nước Lào và Campuchia đều không có trạm đo, vùng giáp biển Đông chỉ có duy nhất 1 trạm tại đảo Lý Sơn (Hình 1). Qua phân tích và đánh giá, nghiên cứu đã lựa chọn được 75 trạm thỏa mãn điều kiện. Trong đó, thông tin của các trạm đo mưa là lượng mưa ngày, thời gian đo ngắn nhất là 15 năm (trạm IaLy) và dài nhất là 59 năm (trạm Pleiku), thời gian đo liên tục trung bình là 31 năm (Bảng 1).

2.3. Tham số thuộc tính trạm đo

Các nghiên cứu trước đây [1], [3], [5] cho thấy việc lựa chọn các thuộc tính thích hợp là một trong những bước quan trọng nhất trong phân vùng mưa. Các thuộc tính này có thể bao gồm: (i) tham biến khí tượng (độ ẩm, nhiệt độ, tốc độ gió...) và (ii) tham số vị trí (kinh độ, vĩ độ và cao độ trạm đo).

Trong nghiên cứu này, tác giả sử dụng tham số thuộc tính mưa là (kinh độ, vĩ độ và cao độ) của trạm quan trắc mưa. Các thuộc tính này phù hợp với hiện trạng số liệu của khu vực nghiên cứu và đặc biệt là tham số cao độ trạm có ảnh hưởng rất lớn đến lượng mưa (Bảng 1).

Bảng 1. Tên trạm và tham số thuộc tính của trạm đo mưa

ID	Tên trạm	Kinh độ	Vĩ độ	Cao độ	Số năm
1	Thạch Hãn	107,185	16,749	11	39
2	Phú Ốc	107,467	16,533	4	36
3	Cỏ Bi	107,433	16,483	43	23

4	Kim Long	107,550	16,450	9	39
5	Huế	107,583	16,433	20	40
6	Bình Điền	107,524	16,356	22	39
7	Tà Lương	107,337	16,301	87	21
8	Dương Hòa	107,591	16,273	112	21
9	A Lưới	107,283	16,217	589	40
10	Nam Đông	107,717	16,150	242	40
11	Thượng Nhật	107,683	16,117	106	37
12	Cầm Lệ	108,200	16,014	6	38
13	Bà Nà	107,990	15,974	724	18
14	Hiên	107,646	15,917	420	28
15	Ái Nghĩa	108,093	15,871	9	35
16	Hội An	108,365	15,870	7	35
17	Câu Lâu	108,279	15,860	10	35
18	Giao Thủy	108,109	15,846	10	35
19	Hội Khách	107,919	15,827	12	18
20	Thành Mỹ	107,833	15,767	200	35
21	Thăng Bình	108,367	15,750	11	18
22	Nông Sơn	108,033	15,700	24	35
23	Hiệp Đức	108,102	15,578	30	25
24	Tam Kỳ	108,439	15,575	12	35
25	Tiên Phước	108,308	15,488	49	26
26	Trà My	108,233	15,350	128	34
27	Phước Sơn	107,832	15,446	234	33
28	Thạch Bàn	108,076	15,768	38	23
29	Vĩnh Trinh	108,179	15,814	33	33
30	Lý Sơn	109,150	15,383	9	30
31	Châu Ổ	108,759	15,306	10	21
32	Trà Bồng	108,383	15,233	8	39
33	Trà Khúc	108,793	15,143	7	39
34	Quảng Ngãi	108,800	15,117	8	35
35	Sông Vệ	108,865	15,047	9	20
36	Sơn Giang	108,571	15,044	39	39
37	Sơn Hà	108,467	15,033	79	39
38	Mộ Đức	108,883	15,033	9	39
39	Sơn Tây	108,335	14,990	380	19
40	An Chi	108,81	14,989	7	39
41	Minh Long	108,706	14,938	66	25
42	Đức Phổ	108,942	14,845	56	38
43	Ba Tơ	108,733	14,767	60	38
44	Sa Huỳnh	109,037	14,720	8	35
45	Giá Vực	108,562	14,706	140	39
46	An Hoà	108,912	14,585	20	26

47	Hoài Nhơn	109,029	14,532	7	26
48	Vĩnh Sơn	108,755	14,332	135	20
49	Bình Tường	108,879	13,935	26	38
50	Bình Quang	108,800	14,133	86	17
51	Vĩnh Kim	108,767	14,233	85	32
52	Vân Canh	108,997	13,621	153	24
53	Đèo Cù Mông	109,183	13,667	98	38
54	Củng Sơn	108,983	13,033	41	36
55	Đắk Glei	107,739	15,093	697	20
56	Sa Thầy	107,783	14,417	571	27
57	Đắk Tô	107,833	14,533	645	20
58	ComPlông	108,133	14,467	721	19
59	Kon Tum	108,017	14,333	517	39
60	Ia Ly	107,750	14,200	606	15
61	Kbang	108,617	14,167	574	26
62	Biên Hồ	108,017	14,050	751	22
63	Ia Hrug	107,883	14,033	648	15
64	PomoreTV	108,35	14,033	686	37
65	Thôn 4	108,133	14,000	705	22
66	Đắk Đoa	108,100	13,983	689	35
67	Pleiku	108,017	13,967	796	59
68	An Khê	108,650	13,950	404	38
69	Chư Sê	108,067	13,700	496	21
70	Chư Prông	107,850	13,650	289	37
71	Ayun Hạ	108,250	13,567	190	16
72	Ayun Pa	108,450	13,383	157	38
73	Krông Pa	108,700	13,300	426	35
74	Buôn Hồ	108,267	12,917	646	33
75	Krông Buk	108,383	12,750	474	37

3. Phương pháp

Giả sử trong vùng nghiên cứu có N trạm đo mưa, khi đó vector thuộc tính biểu diễn cho N trạm đo mưa là: $\mathbf{Y} = \{y_i/i = 1, \dots, N\}$ (1)

Trong đó: $y_i = [y_{i1}, \dots, y_{ij}, \dots, y_{in}] \in \mathbb{R}^n$ là vector thuộc tính của trạm thứ i. Vector thuộc tính này bao gồm: kinh độ, vĩ độ và cao độ của trạm thứ i [3], [5]; y_{ij} là giá trị thuộc tính j của vector y_i .

Cho tập x_i biểu thị cho giá trị tỷ lệ thứ i trong không gian n chiều {i.e., $x_i = [x_{i1}, \dots, x_{ij}, \dots, x_{in}] \in \mathbb{R}^n$ }, trong đó x_{ij} thu được từ sự thay đổi tỷ lệ y_i bằng phương trình:

$$x_{ij} = \frac{(y_{ij} - \bar{y}_j)}{\sigma_j} \quad \text{với } 1 \leq j \leq n \quad (2)$$

Trong đó: x_{ij} biểu thị cho sự thay đổi tỷ lệ của y_{ij} ; σ_j là

độ lệch chuẩn của thuộc tính j, $\bar{y}_j$ là giá trị trung bình của tất cả vector thuộc tính.

3.1. Phương pháp phân cụm không thứ bậc (K-Means)

Phương pháp phân cụm không thứ bậc là phân bố các trạm đo vào số lượng các cụm đã xác định trước, trong đó số lượng cụm K hay còn gọi là tiêu vùng phải được giả định trước ($K = 1, 2, 3, \dots$). Phương pháp này bao gồm: (i) phương pháp ngưỡng tuần tự; (ii) phương pháp ngưỡng song song, (iii) phương pháp phân chia tối ưu. Nghiên cứu này sử dụng phương pháp phân chia tối ưu. Theo đó, thuật toán K-Means được McQueen đề xuất năm 1967 [2] và được sử dụng phổ biến trong RFA [3], [5]. Phương pháp này sử dụng thuật toán lặp để phân thành các cụm sao cho hàm số F đạt cực tiểu. F được xác định như sau:

$$F = \sum_{k=1}^K \sum_{j=1}^n \sum_{i=1}^{N_k} d^2(x_{ij}^k + x_j^k) \quad (3)$$

Trong đó: K là số tiêu vùng và được giả định ban đầu $K=1$; N_k là số trạm trong tiêu vùng thứ k; d là khoảng cách của mỗi vector đến trung tâm của nhóm; x_{ij}^k là giá trị đã thay đổi tỷ lệ thứ j của vector i được quy định bởi nhóm k; x_j^k là giá trị trung bình của thuộc tính j của nhóm thứ K:

$$x_j^k = \frac{\sum_{i=1}^{N_k} x_{ij}^k}{N_k} \quad (4)$$

Để F đạt cực tiểu, thì khoảng cách của mỗi vector đến tâm của nhóm phải nhỏ nhất (d đạt cực tiểu). Chúng ta có thể lựa chọn phương pháp khoảng cách như Euclidean hoặc Mahalanobis. Trong nghiên cứu này tác giả chọn phương pháp khoảng cách như Euclidean và sử dụng ngôn ngữ lập trình R triển khai thuật toán thông qua hàm **K-Means**.

3.2. Phương pháp phân cụm thứ bậc (Ward)

Phương pháp này là một thủ tục được xây dựng theo một cấu trúc thứ bậc hình cây và được tiến hành theo cách tích tụ hoặc phân chia ra. Nghiên cứu này sử dụng phương pháp phân cụm thứ bậc tích tụ dựa vào khoảng cách trung tâm. Phương pháp này được Ward đề xuất năm 1963 và được sử dụng khá phổ biến trong RFA hiện nay [1], [4], đây là một thuật toán phân cấp mà ban đầu giả định một trạm là một tiêu vùng. Thuật toán sau đó sát nhập các tiêu vùng bằng cách xem xét mối quan hệ giữa các đối tượng bằng phương pháp bình phương cực tiểu (Error Sum of Squares - ESS) khoảng cách giữa các đối tượng, ESS được tính như sau:

$$ESS_j = \sum_{i=1}^N (x_{ij} - \bar{y}_j)(x_{ij} - \bar{y}_j) \quad (5)$$

Trong đó: các tham số x_{ij} và $\bar{y}_j$ như trình bày ở công thức (2). Nghiên cứu sử dụng ngôn ngữ lập trình R triển khai thuật toán này thông qua hàm **hclust**.

3.3. Kiểm tra tính đồng nhất mẫu dữ liệu

Theo Hosking và Wallis (1997) vùng đồng nhất được đánh giá bởi tính đồng nhất dữ liệu thông qua L-moment. Lượng mưa trong vùng sẽ được mô phỏng ($N_{sim} = 500$ lần) từ phân phối kappa dựa trên các giá trị trung bình L-

moment: I_1^R , τ^R , τ_3^R and τ_4^R . Các giá trị thống kê này được ước tính theo ba chỉ số đồng nhất H_n ($n = 1, 2$ và 3):

$$\begin{cases} H_1 = \frac{(V-\mu_V)}{\sigma_V} \\ H_2 = \frac{(V_2-\mu_{V_2})}{\sigma_{V_2}} \\ H_3 = \frac{(V_3-\mu_{V_3})}{\sigma_{V_3}} \end{cases} \quad (6)$$

Trong đó: V là độ lệch chuẩn trọng số của mỗi trạm L-CV, V_2 là khoảng cách trung bình trọng số từ trạm đến trung bình trọng số trong không gian ba chiều L-CV, L-CA và L-Kurtosis, và V_3 là khoảng cách trung bình trọng số từ trạm đến trung bình trọng số trong không gian hai chiều L-CA and L-kurtosis. Với μ_V , μ_{V_2} và μ_{V_3} là giá trị trung bình, σ_V , σ_{V_2} , và σ_{V_3} là độ lệch chuẩn ứng với N lần mô phỏng của V , V_2 and V_3 . Một vùng có thể xem “chấp nhận đồng nhất” nếu $H_n < 1$, “có thể đồng nhất” nếu $1 \leq H_n \leq 2$, và “không đồng nhất” nếu $H_n > 2$. Nghiên cứu sử dụng ngôn ngữ lập trình R triển khai thuật toán này thông qua hàm *HW.original*.

4. Kết quả và bàn luận

Các kết quả được trình bày dưới đây được thực hiện theo một quy trình và tiêu chí như sau: Trong RFA, để tăng kích thước mẫu thống kê và giảm sự không chắc chắn suy luận ứng với các tần suất thiết kế, thì số lượng các trạm đo trong vùng hay tiểu vùng càng nhiều càng tốt. Do đó khi phân cụm cần hạn chế phân quá nhiều tiểu vùng, và số lượng các trạm trong các tiểu vùng tương đối đồng đều nhau. Quy trình thực hiện như sau: (i) với 4 mẫu dữ liệu tương ứng với thời đoạn mưa tính toán là 1 NLN, 3 NLN, 5 NLN và 7 NLN của 75 trạm, đầu tiên cho $K=1$ và kiểm tra tính đồng nhất của 4 mẫu dữ liệu này; (ii) nếu mẫu dữ liệu nào không đồng nhất thì tăng $K=2$ và tiếp tục kiểm tra tính đồng nhất các tiểu vùng của mẫu dữ liệu đó. Tương tự như vậy cho đến khi các tiểu vùng là đồng nhất.

Bảng 2 trình bày kết quả kiểm tra tính đồng nhất khi $K=1$ của 4 mẫu dữ liệu (1 NLN, 3 NLN, 5 NLN, 7 NLN). Kết quả cho thấy mẫu dữ liệu ứng với thời đoạn 3 NLN và 5 NLN cho kết quả đồng nhất với chỉ số $H_n < 2$. Đây là một lợi thế rất lớn khi thực hiện RFA với thời đoạn mưa 3 và 5 NLN, vì kích thước mẫu dữ liệu vùng trung bình là $75 \times 31 = 2.325$ năm.

Bảng 2. Chỉ số (H_n) kiểm tra tính đồng nhất của 4 mẫu dữ liệu ($K=1$)

H_n	1 NLN	3 NLN	5 NLN	7 NLN
H_1	2,073	-0,380	-0,667	1,451
H_2	1,163	1,561	1,446	2,902
H_3	0,774	0,868	1,072	2,647

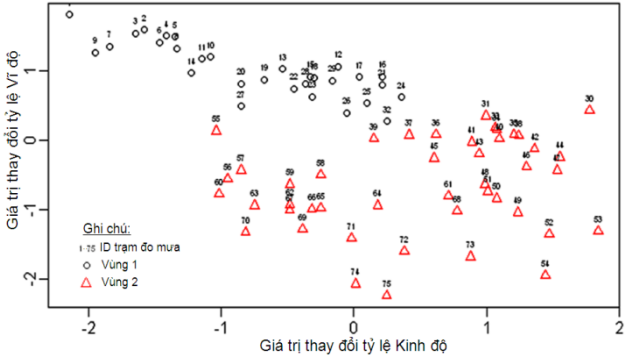
Hai mẫu dữ liệu 1 NLN và 7 NLN là không đồng nhất ($H_n > 2$) nên cần phân cụm dữ liệu với $K=2$ (chia thành 2 tiểu vùng). Sử dụng 2 phương pháp K-Means và Ward lần lượt cho 2 mẫu dữ liệu này và thu được kết quả như sau:

Hình 2 và Hình 3 thể hiện kết quả phân cụm ($K=2$) theo phương pháp K-Means và Ward cho mẫu dữ liệu 1 NLN và 7 NLN với thuộc tính các trạm là kinh độ, vĩ độ và cao

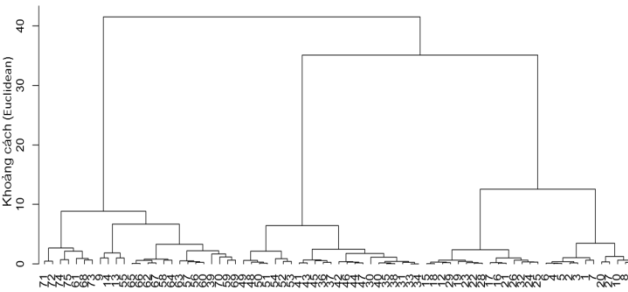
độ. Tuy nhiên kết quả kiểm tra tính đồng nhất của mẫu dữ liệu 1 NLN và 7 NLN cho thấy một trong hai tiểu vùng là không đồng nhất (Bảng 3 và Bảng 4). Do vậy, cần tiếp tục tăng số lượng tiểu vùng.

Biểu đồ phân cụm theo phương pháp Ward (Hình 4) cũng cho thấy với mẫu dữ liệu 1 NLN, các tiểu vùng đồng thời đạt được đồng nhất khi $K=5$. Tuy nhiên kích thước mẫu dữ liệu của 5 tiểu vùng đã bị giảm đáng kể ($N = 23, 15, 12, 7$ và 18 trạm). Trong khi đó, phương pháp K-Means phân chia thành 3 tiểu vùng và mẫu dữ liệu 3 tiểu vùng đều đồng nhất (Bảng 5) và kích thước mẫu dữ liệu mỗi trạm khá đồng đều ($N= 30, 19$ và 26 trạm), tức là giữ nguyên vùng 1 (đồng nhất) và phân chia vùng 2 thành vùng 2' và vùng 2''. Do đó, đối với mẫu dữ liệu 1 NLN, chọn phương pháp K-Means để phân cụm và số tiểu vùng $K=3$ là hợp lý nhất.

Đối với mẫu dữ liệu 7 NLN khi $K=3$ (Bảng 6): theo phương pháp Ward (Hình 4) chia vùng A (50 trạm) thành vùng A' (27 trạm) và vùng A'' (23 trạm). Theo phương pháp K-Means (Hình 5) chia tiểu vùng 2 thành vùng 2' (19 trạm) và vùng 2'' (26 trạm). Bảng 6 cho thấy chỉ số H_n của các tiểu vùng đều đồng nhất, ngoại trừ tiểu vùng A'' (Ward) là có thể đồng nhất ($H_3=2$).



Hình 2. Biểu đồ phân cụm theo phương pháp K-Means (Với $K=2$)



Hình 3. Biểu đồ phân cụm theo phương pháp Ward ($K= 2, 3, 4$ và 5)

Bảng 3. Chỉ số (H_n) kiểm tra tính đồng nhất của mẫu dữ liệu 1 NLN ($K=2$)

Phương pháp	Ward		K-Means	
Vùng/ Số trạm	Vùng A/ ($N_A = 50$ trạm)	Vùng B/ ($N_B = 25$ trạm)	Vùng 1/ ($N_1 = 30$ trạm)	Vùng 2/ ($N_2 = 45$ trạm)
H_1	0,040	2,163	-0,113	2,130
H_2	0,544	1,059	-0,046	1,516
H_3	0,714	0,388	-0,159	1,116

Bảng 4. Chỉ số (H_n) kiểm tra tính đồng nhất của mẫu dữ liệu 7 NLN ($K=2$)

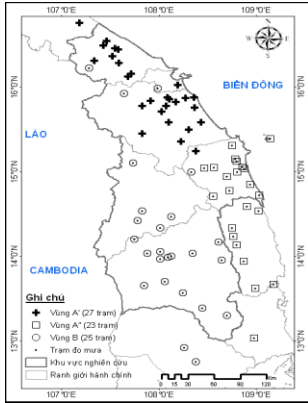
Phương pháp	Ward		K-Means	
Vùng/ Số trạm	Vùng A/ ($N_A = 50$ trạm)	Vùng B/ ($N_B = 25$ trạm)	Vùng 1/ ($N_1 = 30$ trạm)	Vùng 2/ ($N_2 = 45$ trạm)
H_1	-1,575	1,691	-2,347	3,692
H_2	2,288	1,060	-0,407	3,412
H_3	2,650	0,765	-0,235	2,773

Bảng 5. Chỉ số (H_n) kiểm tra tính đồng nhất của mẫu dữ liệu 1 NLN ($K=3$) theo phương pháp K-Means

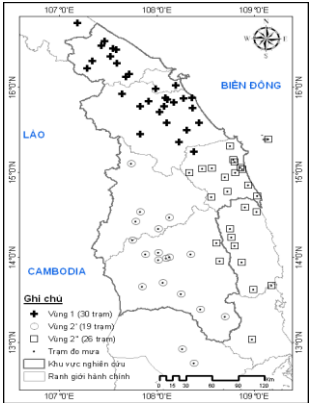
H_n	Vùng 1 ($N= 30$ trạm)	Vùng 2' ($N= 19$ trạm)	Vùng 2'' ($N= 26$ trạm)
H_1	-0,113	1,828	-0,262
H_2	-0,046	0,780	1,009
H_3	-0,159	0,106	1,404
Kết luận	Đồng nhất	Đồng nhất	Đồng nhất

Bảng 6. Chỉ số (H_n) kiểm tra tính đồng nhất của mẫu dữ liệu 7 NLN ($K=3$) theo phương pháp Ward và K-Means

Vùng / Số trạm	Ward			K-Means		
	Vùng A' / ($N=27$ trạm)	Vùng A'' / ($N= 23$ trạm)	Vùng B / ($N= 25$ trạm)	Vùng 1 ($N= 30$ trạm)	Vùng 2' ($N= 19$ trạm)	Vùng 2'' ($N= 26$ trạm)
H_1	-2,546	0,637	1,422	-2,499	0,430	-1,142
H_2	-0,775	1,997	1,070	-0,374	1,684	-0,229
H_3	-0,600	2,000	0,705	-0,179	1,597	-0,311
Kết luận	Đồng nhất	Có thể đồng nhất	Đồng nhất	Đồng nhất	Đồng nhất	Đồng nhất



Hình 4. Bản đồ tiểu vùng theo Ward cho 7 NLN



Hình 5. Bản đồ tiểu vùng theo K-Means cho 7 NLN

5. Kết luận

Nghiên cứu đã thực hiện phân cụm dữ liệu 1 NLN, 3 NLN, 5 NLN và 7 NLN cho 75 trạm đo mưa trên khu vực MT-TN theo 2 phương pháp phổ dụng là Ward và K-Means để quan sát ứng xử của các phương pháp này đối với các mẫu dữ liệu mưa ngày tại khu vực MT-TN, Việt Nam. Kết quả nghiên cứu là rất quan trọng cho bước tiếp theo trong RFA mưa của vùng nghiên cứu, nhằm xây dựng bản đồ mưa cực hạn với độ tin cậy cao.

Kết quả nghiên cứu đã chỉ ra rằng với thời đoạn mưa tính toán 3 NLN và 5 NLN, thì 2 mẫu dữ liệu này hoàn toàn đồng nhất với số lượng trạm $N=75$ trạm (không phân tiểu vùng), đây là lợi thế rất lớn khi thực hiện RFA cho vùng nghiên cứu. Tuy nhiên, với thời đoạn mưa tính toán 1 NLN và 7 NLN thì phương pháp K-Means cho kết quả phân cụm tốt hơn phương pháp Ward. Tác giả khuyến nghị nên dùng kết quả của phương pháp K-Means trong phân cụm cho 2 mẫu dữ liệu này, với số lượng tiểu vùng $K=3$ và số lượng trạm đo của mỗi tiểu vùng lần lượt là 30, 19 và 26 trạm theo vị trí như Hình 5.

TÀI LIỆU THAM KHẢO

[1] H,Wazned, F, Chebana, T, B, M, J, Ouarda (2015), “Delineation of homogeneous regions for regional frequency analysis using statistical depth function”, *Journal of Hydrology*, ISSN: 0022-1694, Vol: 521, p 232-244.

[2] McQueen, J. (1967), *Some methods for classification and analysis of multivariate observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297, Univ. of Calif. Press, Berkeley.

[3] Satyanarayana P, Srinivas VV (2008), “Regional frequency analysis of recipitation using large-scale atmospheric variables”, *Journal of Geophysical Research*, 113:D24110, doi:10.1029/2008JD010412.

[4] Fredline Ilorme, Veronica Webster Griffis (2013), “A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation”, *Journal of Hydrology* 492, ISSN: 0022-1694, 151–162.

[5] Ngogondo CS, C-Y, Xu, L,M,Tallaksen, B, Alemaw and T, Chirwa (2011), *Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches*, Stoch, Env, Res, Risk A, 25, 939-955.

[6] Nguyễn Chí Công (2017), “Phân tích tần suất mưa cực hạn cho tỉnh Gia Lai dựa trên cách tiếp cận vùng”, *Tạp chí Khoa học Kỹ thuật Thủy lợi và Môi trường*, ISSN 1859-3941, Vol: 57.

[7] Nguyễn Chí Công, Nguyễn Vinh Long (2016), “Phân tích tần suất mưa vùng cho tỉnh Quảng Nam”, *Tạp chí Khoa học & Công nghệ - Đại học Đà Nẵng*, ISSN 1859-1531, Vol:5(102), p 10-13.