



DOI:10.22144/ctujos.2026.209

ỨNG DỤNG PHÂN PHỐI HỖN HỢP DIRICHLET PHÂN CỤM CÁC DẠNG NGÀY BỨC XẠ MẶT TRỜI

Trần Văn Lý^{1,*}, Nguyễn Thị Hồng Dân¹, Trương Thành An¹, Vương Thị Mỹ Tiên¹ và Lê Đại Nghiệp²

¹Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam

²Khoa Cơ bản, Trường Đại học Nam Cần Thơ, Việt Nam

*Tác giả liên hệ (Corresponding author): tvly@ctu.edu.vn

Thông tin chung (Article Information)

Nhận bài (Received): 24/06/2026

Sửa bài (Revised): 15/07/2026

Duyệt đăng (Accepted): 26/08/2026

Title: Application of Dirichlet mixture distribution for clustering solar radiation day types

Author(s): Tran Van Ly^{1,*}, Nguyen Thi Hong Dan¹, Truong Thanh An¹, Vuong Thi My Tien¹ and Le Dai Nghiep²

Affiliation(s): ¹College of Natural Science, Can Tho University, Viet Nam;
²Basis College, Nam Can Tho University, Viet Nam

TÓM TẮT

Bài báo được thực hiện nhằm khảo sát khả năng áp dụng mô hình phân phối hỗn hợp Dirichlet kết hợp với thuật toán xấp xỉ ngẫu nhiên cực đại hóa kỳ vọng (mixD-SAEM) để phân loại các dạng ngày bức xạ mặt trời dựa trên phân phối của chỉ số sáng. Hiệu quả của phương pháp được đánh giá thông qua so sánh với hai phương pháp phân cụm phổ biến là k-means và phân cụm thứ bậc. Thử nghiệm được thực hiện trên hai nhóm dữ liệu gồm bộ dữ liệu bức xạ mặt trời có tần suất lấy mẫu cao được sử dụng trong nghiên cứu trước đây và các bộ dữ liệu bức xạ mặt trời ghi nhận tại hai trạm khí tượng ở Đồng bằng sông Cửu Long với chỉ 9 quan sát mỗi ngày. Đối với trường hợp dữ liệu thưa, các histogram phục vụ phân loại được xây dựng từ các hàm mật độ xác suất ước lượng bằng phương pháp hàm hạt nhân. Kết quả đánh giá thông qua các chỉ số Silhouette và Davies–Bouldin cho thấy phương pháp mixD-SAEM có khả năng nhận dạng tốt các dạng ngày bức xạ mặt trời, đặc biệt trên các dữ liệu mà đặc trưng phân biệt chủ yếu nằm ở hình dạng phân phối.

Từ khóa: Bức xạ mặt trời, EM, k-means, phân phối Dirichlet

ABSTRACT

The article investigates the applicability of a Dirichlet mixture distribution model combined with the Stochastic Approximation Expectation Maximization algorithm (mixD-SAEM) for clustering daily solar radiation patterns based on the distribution of the clearness index. The performance of the proposed approach is evaluated and compared with two widely used clustering methods, namely k-means and hierarchical clustering. Experiments are conducted on two datasets: a high-frequency solar radiation dataset previously used in the literature and solar radiation datasets collected from two meteorological stations in the Mekong Delta region of Vietnam, where only nine observations are available per day. For the sparse datasets, histograms used in the clustering process are constructed from probability density functions estimated by kernel density estimation. Clustering performance is assessed using the Silhouette and Davies–Bouldin indices. The results indicate that the mixD-SAEM approach is capable of effectively identifying distinct daily solar radiation patterns, particularly when the discriminative information is embedded in the distributional shape of the data.

Keywords: Dirichlet distribution, EM, k-means, solar radiation

1. GIỚI THIỆU

Sự biến đổi của bức xạ mặt trời gây ra sự biến động của hệ thống sản xuất điện sử dụng các tế bào quang điện. Trong lưới phân phối điện mặt trời với mật độ các tế bào quang điện cao, sự biến động nhanh chóng giữa các dạng phân bố bức xạ khác nhau có thể dẫn đến sự biến đổi khó lường của điện áp nút và công suất trong mạng lưới. Điều này dẫn đến sự cần thiết phải nghiên cứu, khảo sát các loại, các dạng phân phối bức xạ mặt trời có thể biến động thay đổi qua lại lẫn nhau trong ngày hay trong chuỗi các ngày trong năm.

Dữ liệu bức xạ mặt trời đo được trong ngày được chuyển qua khảo sát dưới dạng biểu đồ tần suất (histogram) của các chỉ số sáng tương ứng và các dạng ngày bức xạ mặt trời được phân loại theo các histogram này. Qua quá trình cập nhật tham số của phân phối hỗn hợp Dirichlet bằng thuật toán xấp xỉ ngẫu nhiên cực đại hóa kỳ vọng (SAEM – Stochastic Approximation Expectation Maximization), các histogram được phân bố vào các cụm phân loại khác nhau. Đây là cách tiếp cận phân loại được giới thiệu bởi Soubdhan et al. (2009) - được nêu tên viết gọn là *mixD-SAEM* trong các xem xét, đề cập có liên quan ở phía sau. Trong công bố này, các tác giả cũng đã tiến hành các thực nghiệm trên dữ liệu thực; tuy nhiên, chưa có các đánh giá về hiệu quả của kết quả phân loại và cũng không có các so sánh kết quả với các phương pháp phân loại khác. Hơn nữa, dữ liệu bức xạ mặt trời được sử dụng trong thực nghiệm của Soubdhan et al. (2009) là khá lý tưởng với tốc độ lấy mẫu là 1 Hz (dữ liệu được ghi mỗi nửa giây một lần). Nhờ đó, phương pháp được thực nghiệm trên các histogram có chất lượng cao, được hình thành từ số lượng lớn các giá trị dữ liệu đo được mỗi ngày. Câu hỏi được đặt ra để xem xét ở đây là liệu phương pháp này có thể được áp dụng tốt trên dữ liệu bức xạ mặt trời đo được ở các trạm đo hiện nay ở vùng Đồng bằng sông Cửu Long của Việt Nam, với dữ liệu được ghi đo khá thưa, mỗi giờ ghi lưu một lần?

Dữ liệu được sử dụng để thử nghiệm trong bài báo này là các giá trị bức xạ được ghi lưu ở hai trạm khí tượng đặt tại Cần Thơ và Rạch Giá, ghi đo hàng giờ mỗi ngày trong khung thời gian 9:00 – 17:00 (chi tiết về việc xử lý dữ liệu được trình bày ở mục 5.2.1 phía sau). Do điều kiện dữ liệu quan sát ở mỗi ngày quá ít, với chỉ 9 quan sát mỗi ngày nên các histogram sử dụng phân loại cho mỗi ngày không được lập từ 9 quan sát này mà được thiết lập từ các hàm mật độ xác suất (PDF) được ước lượng từ phương pháp hàm hạt nhân. Chất lượng kết quả thực nghiệm phân loại được đánh giá theo các tiêu chuẩn

Silhouette và Davies–Bouldin. Qua đó, các kết quả nhận được từ phương pháp *mixD-SAEM* được so sánh với các phương pháp phân tích cụm thường được sử dụng phổ biến khác. Trước đó, việc đánh giá và so sánh các phương pháp trên các dữ liệu thực nghiệm được công bố bởi Soubdhan et al. (2009) cũng được tiến hành để kiểm chứng các phương pháp trên dữ liệu có chất lượng ghi đo tốt.

Trong các phương pháp phân tích cụm phổ biến (bao gồm cả dạng có giám sát và không có giám sát) thường được sử dụng như phân tích cụm dựa vào trọng tâm *k-means*, phân tích cụm thứ bậc và không thứ bậc, phân tích cụm dựa trên mô hình, phân tích cụm mờ và không mờ, phân tích cụm tự động,... thì hai phương pháp là phân tích cụm *k-means* (có yêu cầu biết số lượng cụm) và phân tích cụm thứ bậc (ký hiệu *clustergram*, không yêu cầu phải biết số lượng cụm) được phát triển từ rất sớm và có tính đại diện khá tốt. Các kết quả thực nghiệm của phương pháp *mixD-SAEM* được đánh giá và so sánh với các kết quả nhận được từ hai phương pháp này.

2. PHƯƠNG PHÁP PHÂN LOẠI TỪ PHÂN PHỐI HỖN HỢP DIRICHLET

2.1. Phân phối Dirichlet

Xét l biến ngẫu nhiên độc lập $Y_1, Y_2, \dots, Y_l$ lần lượt có phân phối Gamma tương ứng $\gamma(\alpha_1, 1), \gamma(\alpha_2, 1), \dots, \gamma(\alpha_l, 1)$. Khi chuẩn hóa các biến Y_i theo tổng $Y = Y_1 + Y_2 + \dots + Y_l$ thì vector $\left(\frac{Y_1}{Y}, \frac{Y_2}{Y}, \dots, \frac{Y_l}{Y}\right)$ sẽ tuân theo phân phối Dirichlet $D(\alpha_1, \alpha_2, \dots, \alpha_l)$.

Với K biến $X_k \sim D_k(\alpha_{k1}, \alpha_{k2}, \dots, \alpha_{kl})$, $k = 1, 2, \dots, K$, có phân phối Dirichlet được hình thành như trên, một phân phối hỗn hợp Dirichlet là một biến tổng được xác định bởi $\sum_{k=1}^K p_k D_k(\alpha_{k1}, \alpha_{k2}, \dots, \alpha_{kl})$, trong đó $p_1, p_2, \dots, p_K$ là các trọng số không âm thỏa mãn $\sum_{k=1}^K p_k = 1$.

Vấn đề đặt ra khi mô hình hóa theo dạng phân phối hỗn hợp Dirichlet là phải ước lượng tập tham số của mô hình:

$$\theta = \{\alpha_{k1}, \dots, \alpha_{kl}, p_k; k = 1, \dots, K\}.$$

Trong nghiên cứu này, khi phân phối hỗn hợp Dirichlet $\sum_{k=1}^K p_k D_k(\alpha_{k1}, \alpha_{k2}, \dots, \alpha_{kl})$ được sử dụng để mô hình hóa cho dãy histogram các chỉ số sáng bức xạ mặt trời với K nhóm, trong đó thành phần $D_k(\alpha_{k1}, \alpha_{k2}, \dots, \alpha_{kl})$ xác định phân phối của nhóm thứ k ($k = 1, 2, \dots, K$), thì tập tham số θ sẽ được ước lượng dựa vào mẫu dữ liệu quan sát gồm n histogram $h_i = \{h_{i1}, h_{i2}, \dots, h_{il}\}$, $i = 1, 2, \dots, n$, theo thuật toán lặp SAEM, là một dạng thuật toán

lập được cải tiến từ thuật toán EM (Expectation Maximization) (Dempster et al., 1977), bằng cách thêm các bước mô phỏng xấp xỉ ngẫu nhiên để khởi tạo các giá trị tham số phù hợp ban đầu.

2.2. Thuật toán EM

Qua thuật toán EM tập tham số mới sẽ nhận được từ việc cực đại hóa hàm log-likelihood dữ liệu đầy đủ được xác định bởi $Q(\hat{\theta}, \theta) = E_{\theta}(\ln(\Lambda_n^{\hat{\theta}\theta}))$; trong đó, $\ln(\cdot)$ là hàm logarit tự nhiên, $\Lambda_n^{\hat{\theta}\theta} = \frac{dP_{\hat{\theta}}}{dP_{\theta}}|\mathcal{H}_n$ với E_{θ} , P_{θ} lần lượt là kỳ vọng và độ đo xác suất phụ thuộc theo θ , $\mathcal{H}_n = \sigma\{h_1, h_2, \dots, h_n\}$ là σ -đại số lọc dữ liệu của các histogram quan sát được $h_i = \{h_{i,1}, h_{i,2}, \dots, h_{i,l}\}$, $i = 1, 2, \dots, n$.

Xuất phát từ tập tham số được chọn ban đầu $\hat{\theta}^{(0)}$, các bước lặp của thuật toán EM tạo ra được dãy $(\hat{\theta}^{(q)})_{q \geq 1}$ cập nhật ước lượng cho θ . Mỗi bước lặp của thuật toán EM bao gồm 2 bước sau:

Bước E (Expectation): Đặt $\theta = \hat{\theta}^{(q)}$ và tính

$$Q(\hat{\theta}, \hat{\theta}^{(q)}) = E_{\theta}(\ln(\Lambda_n^{\hat{\theta}\hat{\theta}^{(q)}})).$$

Bước M (Maximization): Tìm điều kiện cực đại

$$\hat{\theta}^{(q+1)} = \arg \max_{\hat{\theta}} Q(\hat{\theta}, \hat{\theta}^{(q)}).$$

bước E được lặp lại với $q = q + 1$ cho đến khi các tham số ước lượng không còn biến động thay đổi.

Theo Dembo and Zeitouni (1986), ở mỗi lần lặp của thuật toán EM luôn có $Q(\cdot, \hat{\theta}^{(q+1)}) - Q(\cdot, \hat{\theta}^{(q)}) \geq \lambda(\hat{\theta}^{(q+1)} - \hat{\theta}^{(q)})^2$ với $\lambda > 0$. Đây là cơ sở chính để nhận được kết quả rằng dãy cập nhật $\{\hat{\theta}^{(q)}\}_{q=0,1,2,\dots}$ hội tụ hình học về một tham số θ^* (thuộc trong một bao đóng chứa miền biến động giá trị của tập các tham số) mà không phụ thuộc vào sự lựa chọn tham số xuất phát ban đầu $\hat{\theta}^{(0)}$.

2.3. Thuật toán SAEM

Mục đích sử dụng thuật toán SAEM trong nghiên cứu này là phân loại các ngày bức xạ mặt trời vào K nhóm khác nhau, qua quá trình ước lượng cập nhật cho các tham số của phân phối hỗn hợp Dirichlet. Mỗi dãy dữ liệu bức xạ đo được ở từng ngày được chuyển sang dạng dữ liệu chỉ số sáng rồi lập histogram tương ứng với l khoảng chia đều nhau, các dạng ngày bức xạ sẽ được phân loại theo các histogram này. Giả sử có một mẫu gồm n ngày dữ liệu bức xạ mặt trời, thuật toán SAEM phân loại theo các histogram chỉ số sáng tương ứng $h_i =$

$\{h_{i,1}, h_{i,2}, \dots, h_{i,l}\}$, $i = 1, \dots, n$, xếp vào K nhóm theo các bước lặp được trình bày dưới đây.

Bước khởi tạo ban đầu $q = 0$: Gán ngẫu nhiên mỗi histogram vào một nhóm, khởi tạo ngẫu nhiên các xác suất hậu nghiệm $t_{i,l_i}^{(0)}$ và tập tham số ban đầu theo các phân phối đều

$$\theta^{(0)} = \{\alpha_{k1}^{(0)}, \dots, \alpha_{kl}^{(0)}, p_k^{(0)}; k = 1, \dots, K\}.$$

Bước xấp xỉ ngẫu nhiên (SA): Sinh nhân ngẫu nhiên $e_{qi} = (e_{qi}^k)$ cho mỗi ngày i theo phân phối đa thức theo các xác suất tiên nghiệm $t_{i,l_i}^{(q)}$, từ đó tạo phân hoạch $C = (C_k)_{\{k=1,\dots,K\}}$ xếp các histogram vào các nhóm.

Nếu có tỷ lệ phân hoạch của nhóm nào đó quá nhỏ sẽ trở lại bước $q = 0$ và bước SA.

Tại vòng lặp $q = 1, 2, \dots, Q$; thuật toán thực hiện các bước lặp **E** và **M** dưới đây.

Bước E: Tính các trọng số và các tham số Dirichlet thành phần

$$p_k^{(q)} = \frac{1}{n} [(1 - \gamma_q) \sum_{i=1}^N t_{ik}^q + \gamma_q \sum_{i=1}^N e_{qi}^k], \quad (1)$$

$$\alpha_k^{(q)} = (1 - \gamma_q) \frac{\sum_{i=1}^N t_{ik}^q b(f_i)}{\sum_{i=1}^N t_{ik}^q} + \gamma_q \frac{\sum_{i=1}^N e_{qi}^k b(f_i)}{\sum_{i=1}^N e_{qi}^k}. \quad (2)$$

Bước M: Cập nhật các xác suất hậu nghiệm xếp histogram i vào nhóm k

$$t_{ik}^{(q+1)} = \frac{p_k^{(q+1)} D(\alpha_{k,1}^{(q+1)}, \alpha_{k,2}^{(q+1)}, \dots, \alpha_{k,l}^{(q+1)})(y_i)}{\sum_{k=1}^K p_k^{(q+1)} D(\alpha_{k,1}^{(q+1)}, \alpha_{k,2}^{(q+1)}, \dots, \alpha_{k,l}^{(q+1)})(y_i)}.$$

Thuật toán lặp đến bước Q được ấn định hoặc cho đến khi các tham số ước lượng (các trọng số và các tham số Dirichlet) không còn biến động thay đổi. Sự hội tụ của thuật toán SAEM được đảm bảo bởi sự hội tụ của thuật toán EM.

2.4. Ước lượng hàm hạt nhân

Trong nghiên cứu của Soubdhan et al. (2009), dữ liệu bức xạ mặt trời được lấy mẫu với tần suất rất cao (1 Hz), nhờ đó mỗi ngày quan sát có thể thu được hàng chục nghìn giá trị chỉ số sáng. Các histogram được xây dựng trực tiếp từ các quan sát này có khả năng phản ánh khá đầy đủ đặc điểm phân phối của từng dạng ngày bức xạ mặt trời.

Tuy nhiên, đối với các bộ dữ liệu bức xạ mặt trời hiện có tại các trạm khí tượng ở Đồng bằng sông Cửu Long, dữ liệu chỉ được ghi nhận theo giờ trong khoảng thời gian từ 9:00 đến 17:00, tương ứng chỉ có 9 quan sát mỗi ngày. Với số lượng quan sát nhỏ như vậy, việc xây dựng histogram trực tiếp từ dữ

liệu có thể dẫn đến hiện tượng phân bố rời rạc, không ổn định và khó phân ánh chính xác đặc điểm phân phối của các dạng ngày bức xạ mặt trời.

Để khắc phục hạn chế này, phương pháp ước lượng hàm mật độ xác suất không tham số bằng hàm hạt nhân (Kernel Density Estimation – KDE) đã được sử dụng. Đây là phương pháp xây dựng hàm mật độ xác suất từ dữ liệu mẫu mà không cần giả thiết trước dạng phân phối của dữ liệu.

Giả sử $k_1, k_2, \dots, k_n$ là mẫu gồm n giá trị chỉ số sáng của một ngày quan sát. Hàm mật độ xác suất tại điểm k được ước lượng theo công thức

$$\hat{f}(k) = \frac{1}{nh} \sum_{i=1}^n \psi\left(\frac{k-k_i}{h}\right) \quad (3)$$

Trong đó:

- $\psi(\cdot)$ là hàm hạt nhân,
- h là tham số băng thông ($h > 0$),
- n là kích thước mẫu.

Trong bài báo này, hàm hạt nhân được sử dụng là hàm Gauss. Tham số băng thông h quyết định mức độ làm trơn của hàm mật độ ước lượng. Nếu h quá nhỏ thì hàm mật độ nhận được dao động mạnh và phản ánh nhiều nhiễu của dữ liệu; ngược lại, nếu h quá lớn thì các đặc trưng phân biệt giữa các dạng ngày bức xạ có thể bị làm mờ. Trong nghiên cứu này, giá trị băng thông được lựa chọn tự động theo quy tắc Silverman nhằm cân bằng giữa độ trơn và khả năng phản ánh cấu trúc của dữ liệu.

Sau khi nhận được hàm mật độ xác suất $\hat{f}(x)$, đoạn giá trị của chỉ số sáng $[0;1]$ được chia thành l khoảng đều nhau. Giá trị histogram tương ứng của mỗi khoảng được xác định từ tích phân của hàm mật độ trên khoảng đó.

Các histogram thu được từ quá trình ước lượng hàm hạt nhân này được sử dụng làm dữ liệu đầu vào cho mô hình hỗn hợp Dirichlet và thuật toán SAEM. Cách tiếp cận này cho phép khai thác thông tin về hình dạng phân phối của dãy chỉ số sáng ngay cả khi số lượng quan sát trong mỗi ngày rất nhỏ, từ đó mở rộng khả năng áp dụng phương pháp *mixD-SAEM* đối với các bộ dữ liệu bức xạ mặt trời được ghi nhận trong điều kiện thực tế ở Việt Nam.

3. PHƯƠNG PHÁP PHÂN TÍCH CỤM K-MEANS, PHƯƠNG PHÁP THỨ BẬC

3.1. Phương pháp phân tích cụm *k-means*

Phân tích cụm *k-means* là một trong những phương pháp phân cụm không giám sát được sử dụng phổ biến nhất hiện nay. Ý tưởng cơ bản của

phương pháp là phân chia tập dữ liệu gồm n đối tượng quan sát thành K nhóm sao cho các đối tượng trong cùng một nhóm có mức độ tương đồng cao, đồng thời sự khác biệt giữa các nhóm là lớn nhất.

Giả sử tập dữ liệu gồm n đối tượng quan sát $(X_1, X_2, \dots, X_n)$, trong đó mỗi đối tượng được biểu diễn bởi một vector l chiều. Với số cụm K đã biết trước, phương pháp *k-means* tìm cách cực tiểu hóa tổng bình phương khoảng cách từ các đối tượng đến tâm cụm tương ứng:

$$W(K) = \sum_{k=1}^K \sum_{X_i \in C_k} |X_i - \mu_k|^2$$

Trong đó, C_k là cụm thứ k và μ_k là vector tâm của cụm đó, $k = 1, 2, \dots, K$.

Thuật toán *k-means* được thực hiện theo các bước lặp sau:

Bước 1. Chọn ngẫu nhiên K tâm cụm ban đầu.

Bước 2. Gán mỗi đối tượng vào cụm có tâm gần nhất theo khoảng cách Euclid.

Bước 3. Cập nhật lại tâm của từng cụm bằng trung bình cộng các đối tượng thuộc cụm:

$$\mu_k = \frac{1}{n_k} \sum_{X_i \in C_k} X_i$$

Trong đó, n_k là số phần tử thuộc cụm C_k , $k = 1, 2, \dots, K$.

Bước 4. Lặp lại bước 2 và bước 3 cho đến khi đạt số lần lặp tối đa hoặc khi các tâm cụm không còn thay đổi đáng kể, tức là khi giá trị của hàm tổng khoảng cách giữa các quan sát tới tâm cụm mà nó thuộc về (gọi là hàm biến dạng, là hàm giảm, bị chặn dưới bởi 0) không còn biến động thay đổi nhiều (giá trị hàm tiệm cận gần về giá trị nhỏ nhất).

Trong nghiên cứu này, phương pháp *k-means* được áp dụng trên ma trận dữ liệu các chỉ số sáng của các ngày bức xạ mặt trời cần phân loại. Mỗi ngày được biểu diễn bởi một vector gồm các giá trị chỉ số sáng theo thời gian trong ngày. Số cụm K được xác định trước dựa trên thống kê Gap trình bày ở mục 4.1. Kết quả phân cụm nhận được từ *k-means* được sử dụng làm cơ sở đối sánh với phương pháp *mixD-SAEM* nhằm đánh giá khả năng nhận dạng các dạng ngày bức xạ mặt trời khác nhau.

3.2. Phương pháp thứ bậc *clustergram*

Phân tích cụm thứ bậc là phương pháp phân cụm xây dựng cấu trúc phân cấp giữa các đối tượng dữ

liệu thông qua quá trình gộp hoặc tách cụm liên tiếp. Khác với *k-means*, phương pháp này không yêu cầu phải xác định trước số lượng cụm và thường được sử dụng để khảo sát cấu trúc tiềm ẩn của dữ liệu.

Có hai hướng tiếp cận chính là phân cụm kết tụ và phân cụm phân chia. Trong nghiên cứu này, phương pháp kết tụ đã được sử dụng, bắt đầu với n cụm đơn lẻ, mỗi đối tượng dữ liệu tạo thành một cụm riêng biệt. Ở mỗi bước lặp, hai cụm có khoảng cách gần nhau nhất được hợp nhất thành một cụm mới. Quá trình này tiếp tục được thực hiện cho đến khi toàn bộ dữ liệu được gộp thành một cụm duy nhất.

Khoảng cách giữa hai cụm có thể được xác định theo nhiều tiêu chuẩn khác nhau như liên kết đơn, liên kết đầy đủ, liên kết trung bình. Trong nghiên cứu này, khoảng cách Euclid được sử dụng kết hợp tiêu chuẩn liên kết đầy đủ nhằm giảm phương sai nội cụm và tạo ra các cụm có tính đồng nhất cao.

Kết quả của phân tích cụm thứ bậc được biểu diễn dưới dạng biểu đồ cây. Dựa trên chiều cao của các nhánh cây và khoảng cách hợp nhất giữa các cụm, người phân tích có thể xác định số lượng cụm thích hợp bằng cách lựa chọn một mức cắt phù hợp trên biểu đồ.

Đối với bộ dữ liệu bức xạ mặt trời, phương pháp phân cụm thứ bậc được áp dụng trên cùng tập dữ liệu được sử dụng cho *k-means*. Ngoài vai trò là một phương pháp đối sánh với *mixD-SAEM*, biểu đồ cây nhận được còn được sử dụng kết hợp với thống kê Gap để hỗ trợ lựa chọn số lượng cụm tối ưu trước khi thực hiện các thực nghiệm phân cụm ở mục 5.

4. SỐ CỤM TỐI ƯU VÀ CÁC CHỈ SỐ ĐÁNH GIÁ PHÂN TÍCH CỤM

Phương pháp *mixD-SAEM* được đề xuất để phân loại các dạng ngày bức xạ mặt trời được sử dụng với giả thiết đã biết số lượng cụm K . Trong các thực nghiệm, phương pháp thống kê Gap (Gap Statistic) được sử dụng cùng với việc quan sát biểu đồ phân cụm thứ bậc để xác định số cụm. Cơ sở và nguyên tắc cơ bản của phương pháp thống kê Gap cùng hai chỉ số đánh giá hiệu quả phân tích cụm là chỉ số Silhouette và chỉ số Davies-Bouldin sẽ được trình bày lần lượt tiếp theo dưới đây.

4.1. Xác định số cụm tối ưu theo thống kê Gap

Có nhiều tiêu chí và phương pháp thống kê (phương pháp khuỷu tay, thống kê Gap, phương pháp silhouette,...) có thể được sử dụng để xác định số lượng cụm tối ưu nhưng hiện chưa có lựa chọn

nào được xem là tốt nhất cho vấn đề này. Số lượng cụm tối ưu được xác định có phần chủ quan (dựa vào kinh nghiệm phân tích, trực quan hóa dữ liệu,...) và phụ thuộc vào phương pháp được sử dụng để đo lường sự tương đồng và các tham số được sử dụng để phân cụm.

Giải pháp được sử dụng ở đây là dựa vào thống kê Gap; sau đó, việc quan sát kiểm tra biểu đồ cây được tạo ra được thực hiện bằng phương pháp phân cụm thứ bậc để xem liệu có phù hợp với gợi ý một số lượng cụm quan sát từ đó hay không.

Tiếp cận theo thống kê Gap là một phương pháp định lượng chặt chẽ để xác định số cụm tối ưu, được đề xuất bởi Tibshirani et al. (2001). Phương pháp này so sánh mức độ tập trung của dữ liệu thực với mức độ tập trung của dữ liệu tham chiếu được sinh ngẫu nhiên.

Giả sử dữ liệu được phân thành K cụm $C_1, C_2, \dots, C_K$. Với D_r là tổng khoảng cách giữa tất cả các cặp điểm trong cụm C_r , và n_r là số phần tử của cụm C_r . Tiêu chí đánh giá tổng bình phương khoảng cách bên trong cụm được định nghĩa là:

$$W_K = \sum_{r=1}^K \frac{1}{2n_r} D_r.$$

Tiếp cận của thống kê Gap là chuẩn hóa sự biến thiên của $\ln(W_K)$ bằng cách so sánh nó với giá trị kỳ vọng của nó dưới một phân phối tham chiếu đồng nhất không có cấu trúc cụm. Giá trị thống kê Gap cho K cụm được tính bằng công thức:

$$Gap_K = E_n^*[\ln(W_K)] - \ln(W_K)$$

Trong đó, $E_n^*[\ln(W_K)]$ chỉ kỳ vọng toán học trên v tập dữ liệu mẫu tham chiếu được sinh ra ngẫu nhiên.

Để tính toán sai số mô phỏng, gọi s_K là độ lệch chuẩn của v lần mô phỏng mẫu tham chiếu:

$$s_K = t_K \sqrt{1 + \frac{1}{v}},$$

với t_K là độ lệch chuẩn mẫu của các đại lượng $\ln(W_K)$.

Số lượng cụm tối ưu $\hat{K}$ sẽ là giá trị K nhỏ nhất thỏa mãn điều kiện chỉ số Gap tại đó không bị vượt qua đáng kể bởi chỉ số Gap của số cụm kế tiếp $K+1$. Quy tắc toán học được Tibshirani et al. (2001) đề xuất là chọn $\hat{K}$ sao cho:

$$Gap_K \geq Gap_{K+1} - s_{K+1}. \quad (4)$$

Số cụm tối ưu là điểm mà tại đó giá trị Gap đạt cực đại địa phương hoặc bắt đầu có xu hướng bão hòa, đảm bảo cấu trúc phân cụm vừa đủ chặt chẽ, vừa tránh được hiện tượng quá khớp dữ liệu.

4.2. Chỉ số Silhouette đánh giá hiệu quả phân tích cụm

Chỉ số Silhouette được đề xuất bởi Rousseeuw (1987) và hiện là một trong những tiêu chuẩn được sử dụng khá phổ biến để đánh giá chất lượng phân cụm. Chỉ số này đồng thời xem xét mức độ tương đồng của một đối tượng với các phần tử trong cùng cụm và mức độ tách biệt của đối tượng đó với các cụm còn lại.

Giả sử đối tượng thứ i thuộc cụm C_k , $i = 1, 2, \dots, n$. Ký hiệu $a(i)$ là khoảng cách trung bình từ đối tượng này đến tất cả các đối tượng còn lại trong cùng cụm:

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} d(i, j)$$

Trong đó, $d(i, j)$ là khoảng cách giữa hai đối tượng i và j .

Đối với mỗi cụm khác C_r , $r \neq k$, tính khoảng cách trung bình từ đối tượng i ($i \in C_k$) đến các phần tử thuộc cụm đó:

$$d_{ir} = \frac{1}{|C_r|} \sum_{j \in C_r} d(i, j).$$

Giá trị $b(i) = \min_{r \neq k} d_{ir}$ được gọi là khoảng cách lân cận gần nhất của đối tượng i đến một cụm khác.

Khi đó chỉ số Silhouette của đối tượng thứ i được xác định bởi

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad -1 \leq s(i) \leq 1.$$

Giá trị Silhouette trung bình của toàn bộ dữ liệu được tính bởi

$$\bar{s} = \frac{1}{n} \sum_{i=1}^n s(i).$$

Ý nghĩa của chỉ số Silhouette được sử dụng để đánh giá như sau:

- $s(i)$ dương: đối tượng được phân vào đúng cụm,
- $s(i)$ gần bằng 1: đối tượng được phân vào đúng cụm và nằm tách biệt tốt với các cụm khác,
- $s(i)$ gần bằng 0: đối tượng nằm gần ranh giới giữa hai cụm,
- $s(i)$ âm: đối tượng có thể đã được gán nhầm cụm.

Trong đánh giá phân cụm thực nghiệm phía sau, ngoài giá trị Silhouette trung bình, tỷ lệ các phần tử có chỉ số Silhouette dương cũng được sử dụng để đánh giá hiệu quả phân cụm. Tỷ lệ này càng cao chứng tỏ phần lớn các đối tượng được gán vào đúng cụm tương ứng. Khi so sánh các phương pháp phân cụm trên cùng một bộ dữ liệu, phương pháp có giá trị Silhouette trung bình lớn hơn và tỷ lệ Silhouette dương cao hơn được xem là cho kết quả phân cụm tốt hơn.

4.3. Chỉ số Davies-Bouldin đánh giá hiệu quả phân tích cụm

Chỉ số Davies-Bouldin (Davies & Bouldin, 1979) là một tiêu chuẩn đánh giá nội tại của kết quả phân cụm dựa trên sự kết hợp giữa độ phân tán nội cụm và khoảng cách tách biệt giữa các cụm. Không giống chỉ số Silhouette được tính trên từng đối tượng, chỉ số Davies-Bouldin được xây dựng trực tiếp từ các đặc trưng của từng cụm.

Giả sử dữ liệu được phân thành K cụm. Với cụm C_i có tâm cụm là phần tử μ_i , độ phân tán nội cụm được xác định bởi

$$S_i = \frac{1}{|C_i|} \sum_{X \in C_i} |X - \mu_i|.$$

Khoảng cách giữa hai tâm cụm C_i và C_j được xác định bởi $M_{ij} = |\mu_i - \mu_j|$. Từ đó xây dựng hệ số tương đồng giữa hai cụm $R_{ij} = \frac{S_i + S_j}{M_{ij}}$. Đối với mỗi cụm C_i , xác định $R_i = \max_{j \neq i} R_{ij}$. Chỉ số Davies-Bouldin của toàn bộ kết quả phân cụm được tính bởi

$$DB = \frac{1}{K} \sum_{i=1}^K R_i.$$

Giá trị DB luôn không âm. Chỉ số này càng nhỏ thì kết quả phân cụm càng tốt do các cụm có mức độ phân tán nội cụm thấp đồng thời khoảng cách giữa các cụm lớn. Ngược lại, giá trị DB lớn cho thấy các cụm bị chồng lấn hoặc chưa được tách biệt rõ ràng.

Trong bài báo, chỉ số Davies-Bouldin được sử dụng đồng thời với chỉ số Silhouette để đánh giá hiệu quả của các phương pháp phân cụm. Khi so sánh trên cùng một bộ dữ liệu, phương pháp có giá trị Silhouette lớn hơn và giá trị Davies-Bouldin nhỏ hơn được xem là cho kết quả phân cụm tốt hơn. Việc sử dụng đồng thời hai tiêu chuẩn này giúp đánh giá đầy đủ hơn cả mức độ đồng nhất bên trong cụm lẫn khả năng tách biệt giữa các cụm.

5. ÁP DỤNG THỰC NGHIỆM

5.1. Dữ liệu thực nghiệm, chỉ số sáng

Trong phần thực nghiệm này, phương pháp phân tích cụm được đề xuất *mixD-SAEM* cùng hai phương pháp đối sánh *k-means*, *clustergram* được thực nghiệm trên hai dạng dữ liệu bức xạ mặt trời với số quan sát đo được trong ngày nhiều ít khác nhau. Thứ nhất là bộ dữ liệu được sử dụng ở Soubdhan et al. (2009). Đây là bộ dữ liệu có số quan sát bức xạ mặt trời trong ngày lớn với các histogram của các chỉ số sáng tương ứng có chất lượng phân biệt tốt. Dạng dữ liệu thứ hai là các dữ liệu đo được ở các trạm đo đặt ở Đồng bằng sông Cửu Long, có số quan sát mỗi ngày nhỏ, phương pháp *mixD-SAEM* khi này được áp dụng trên các histogram lập từ các PDF nhận được từ phương pháp ước lượng hàm hạt nhân.

Do biên độ biến động của dữ liệu bức xạ mặt trời rất lớn nên thường được thay thế trong các nghiên cứu thống kê bằng đại lượng chuẩn hóa vô hướng (không đơn vị) tương ứng là chỉ số sáng. Chỉ số sáng K_t tại thời điểm t là tỷ lệ giữa lượng bức xạ mặt trời đo được trên bề mặt nằm ngang G_t (W/m^2) và lượng bức xạ ngoài khí quyển H_t (W/m^2) được tính theo các phương trình thiên văn (Sen, 2008) ở thời điểm đó:

$$k_t = \frac{G_t}{H_t}. \quad (5)$$

Đây là loại chỉ số dùng để đo mức độ sáng trong của bầu trời. Các chỉ số sáng có cùng các thuộc tính thống kê với dãy dữ liệu bức xạ mặt trời và lấy giá trị trong giới hạn từ 0 đến 1, chỉ số này càng gần 1 thì bầu trời càng trong sáng.

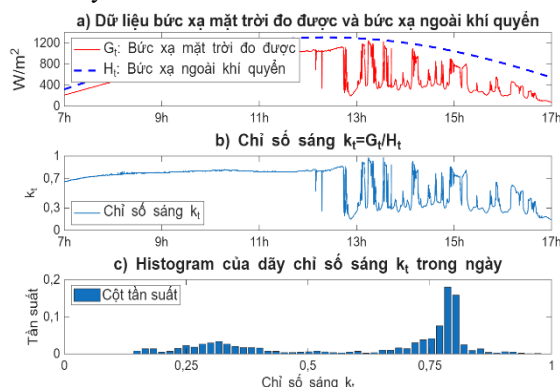
5.2. Dữ liệu có số quan sát trong ngày lớn

5.2.1. Giới thiệu dữ liệu

Dữ liệu bức xạ mặt trời được sử dụng trong Soubdhan et al. (2009) được đo tại Guadeloupe, một hòn đảo ở Tây Ấn thuộc Pháp, nằm ở vĩ độ $16^\circ 15' B$ và kinh độ $60^\circ 30' Đ$. Mẫu ghi đo ở tần số 1 Hz được thực hiện từ tháng 9 năm 2005 đến tháng 12 năm 2006 bằng máy đo bức xạ mặt trời của KIPP&ZONEN, model SP-Lite. Cảm biến này có thời gian phản hồi nhỏ hơn 1 giây, ghi lưu dữ liệu mỗi nửa giây 1 lần. Thực nghiệm áp dụng các phương pháp phân loại ở đây được chọn thực hiện trên 225 ngày quan sát của bộ dữ liệu này, từ ngày 21/03/2006 đến ngày 31/10/2026. Đây là khoảng thời gian với đa số các ngày dữ liệu được ghi lưu đầy đủ hoặc ít bị khuyết thiếu nhất.

Xử lý đối với mỗi ngày dữ liệu: Trước hết, dãy bức xạ ngoài khí quyển (H_t , W/m^2) được tính theo

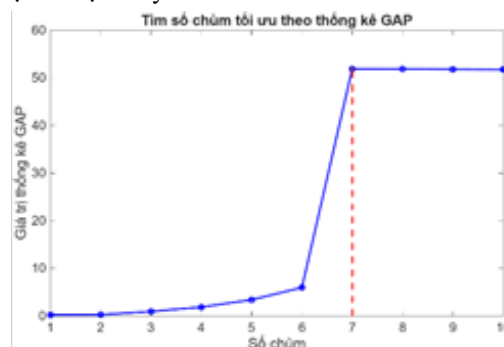
các phương trình thiên văn, cùng với dãy bức xạ mặt trời đo được (G_t , W/m^2), dãy chỉ số sáng (k_t) nhận được theo tỷ lệ ở (5), sau đó lập biểu đồ tần suất histogram của (k_t) theo $l = 50$ khoảng chia đều nhau trên đoạn giá trị giới hạn chỉ số sáng từ 0 đến 1. Hình 1 giới thiệu các dải bức xạ ngoài khí quyển, bức xạ mặt trời đo được tại máy đo, dải chỉ số sáng và histogram tương ứng của một ngày dữ liệu. Thực nghiệm các phân tích cụm được tiến hành trên dãy $n = 225$ Histogram được lập theo cách thức được thể hiện ở Hình 1. Phương pháp phân cụm *mixD-SAEM* và hai phương pháp đối sánh được áp dụng trên bộ dữ liệu này và được trình bày ở mục tiếp theo dưới đây.



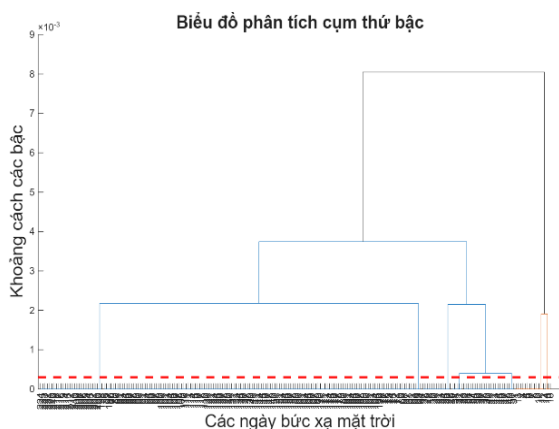
Hình 1. Xác định histogram của dãy chỉ số sáng từ các dãy dữ liệu bức xạ mặt trời trong ngày

5.2.2. Kết quả thực nghiệm

Phương pháp thống kê Gap xác định số cụm tối ưu cho bộ dữ liệu là $K = 7$ (Hình 2). Đồng thời, việc quan sát biểu đồ cây được tạo ra bằng phương pháp phân cụm thứ bậc ở Hình 3 củng cố thêm cho việc chọn số cụm này.



Hình 2. Xác định số cụm tối ưu với bộ dữ liệu có số quan sát trong ngày lớn



Hình 3. Biểu đồ phân cụm thứ bậc với bộ dữ liệu có số quan sát trong ngày lớn

Bảng 1. Đánh giá kết quả phân cụm ngày bức xạ mặt trời trên dữ liệu có số quan sát trong ngày lớn

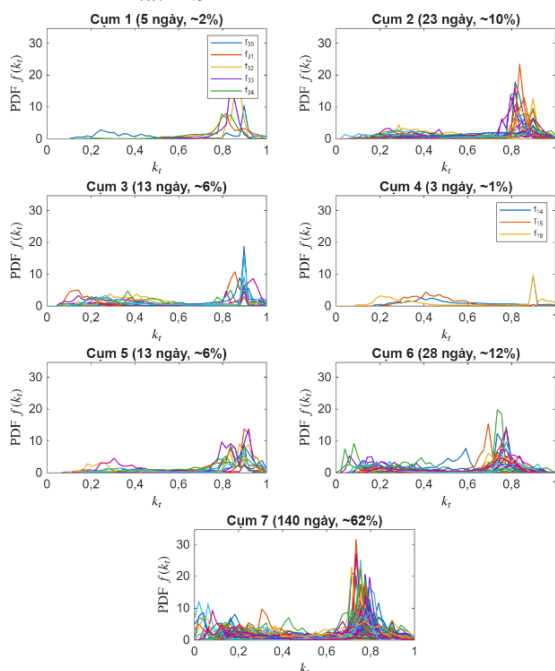
Phương pháp	Chỉ số Silhouette		Chỉ số Davies-Bouldin
	Tỷ lệ dương (%)	Trung bình	
<i>k-means</i>	14,22	- 0,1869	5,2453
<i>clustergram</i>	14,22	- 0,1869	5,2453
<i>mixD-SAEM</i> (7 cụm)	48,89	- 0,0153	3,5572
<i>mixD-SAEM</i> (4 cụm)	80,89	0,2625	2,6206

Kết quả phân cụm nhận được theo *k-means* và *clustergram* là hoàn toàn giống nhau về số lượng, tỷ lệ và hình dạng của các PDF ở các cụm (Hình 4). Kết quả giống nhau này còn thể hiện ở các chỉ số đánh giá Shihouette và Davies-Bouldin ở Bảng 1. Tỷ lệ phân cụm tốt tương ứng theo chỉ số Shihouette dương là không cao, chỉ 14,22%. Trung bình chỉ số Shihouette bị âm chứng tỏ có nhiều phần tử dữ liệu bị phân cụm sai hoặc bị chồng lấn.

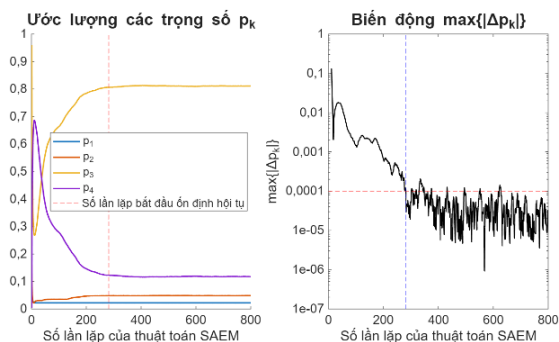
Với cùng số cụm $K = 7$, phương pháp *mixD-SAEM* cho kết quả tốt hơn về chỉ số Silhouette. So sánh về chỉ số Davies-Bouldin cũng nhận được kết quả tốt hơn (thấp hơn) so với chỉ số tương ứng của hai phương pháp *k-means* và *clustergram*.

Dòng cuối cùng của Bảng 1 liệt kê các chỉ số đánh giá phân cụm theo phương pháp *mixD-SAEM* với $K = 4$ cụm. Đây là số cụm được sử dụng trong thực nghiệm của Soubdhan et al. (2009). Số cụm thực nghiệm này được các tác giả xác định từ những khảo sát và tham khảo nhiều phân loại trước đó trên các dữ liệu tương tự. Như đã nêu ở mục 4.1, số lượng cụm tối ưu được xác định có phần phụ thuộc vào phương pháp được sử dụng để đo lường sự tương

đồng, có thể do cách tiếp cận không gần lắm với phương pháp *mixD-SAEM* khai thác mạnh thuộc tính phân phối của đối tượng phân loại nên phương pháp thống kê Gap gợi ý số cụm $K = 7$ chưa là tối ưu đối với *mixD-SAEM*.



Hình 4. Kết quả phân cụm theo các phương pháp *k-means*, *clustergram* trên bộ dữ liệu với số quan sát trong ngày lớn

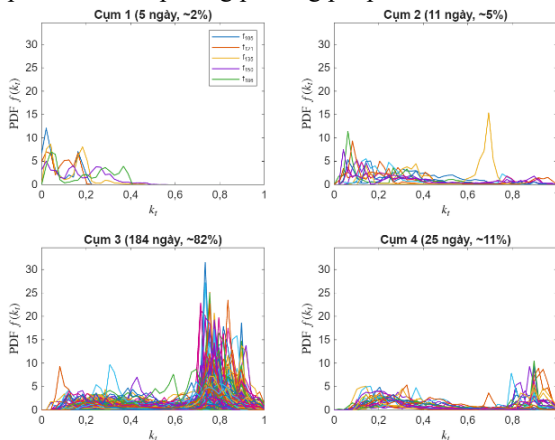


Hình 5. Quá trình biến thiên đến biến thiên ước lượng các trọng số trong phân phối hỗn hợp Dirichlet qua thuật toán SAEM

Hình 5 thể hiện sự hội tụ của thuật toán SAEM trong phân cụm *mixD-SAEM* với số cụm $K = 4$. Tính hội tụ thể hiện ở sự ổn định của các cập nhật thay đổi các trọng số p_k ($k = 1, 2, 3, 4$) ở các lần lặp của thuật toán. Sự cập nhật các trọng số bắt đầu ổn định từ gần lần lặp thứ 300 của thuật toán. Ở các lần lặp bắt đầu từ đây biến động thay đổi các giá trị trọng số không nhiều, thể hiện qua quan sát giá trị sai lệch

thay đổi lớn nhất của các trọng số $\max\{|\Delta p_k|\}$ hầu như ổn định thấp hơn mức 10^{-4} (đồ thị bên phải tại Hình 5).

Hình 6 biểu diễn các dạng PDF (được xác định từ dữ liệu chỉ số sáng theo cùng số lượng số gia đều nhau $l = 50$ tương ứng với các histogram) ở 4 cụm nhận được từ phương pháp *mixD-SAEM*. Các dạng PDF được phân cụm tương đối tốt, mặc dù cũng còn một phần không nhiều chồng lấn của các cụm với nhau. Mặt khác, thể hiện qua các chỉ số đánh giá Shihouette và Davies-Bouldin ở Bảng 1, cụ thể phương pháp *mixD-SAEM* với 4 cụm có tỷ lệ chỉ số Shihouette dương cao nhất (80,89%) với mức trung bình cũng ở mức cao nhất (0,2625), đồng thời chỉ số Davies-Bouldin có giá trị đạt được ở mức thấp nhất (2,6206), nên có thể kết luận (theo nguyên tắc đã nêu ở cuối mục 4.3) rằng chất lượng phân cụm theo *mixD-SAEM* với $K = 4$ cụm tốt hơn nhiều so các phương pháp *k-means*, *clustergram* (và cả với trường hợp áp dụng cùng phương pháp này theo $K = 7$ cụm). Điều này thể hiện rằng phương pháp *mixD-SAEM* có thể được áp dụng tốt hơn với dạng dữ liệu phân loại có đặc tính và chất lượng phân biệt khác biệt về dạng phân phối. Qua đây cũng chỉ ra rằng cần có cách xác định số cụm phù hợp trong việc phát triển và áp dụng phương pháp *mixD-SAEM*.



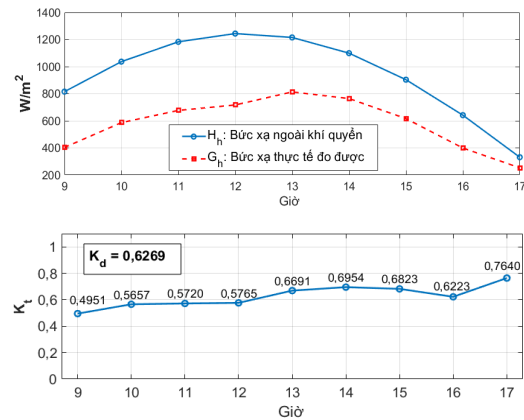
Hình 6. Kết quả phân cụm theo các phương pháp *mixD-SAEM* trên bộ dữ liệu với số quan sát trong ngày lớn

5.3. Dữ liệu có số quan sát trong ngày nhỏ

5.3.1. Giới thiệu dữ liệu

Dạng dữ liệu thứ hai sử dụng thực nghiệm được lấy từ Visual Crossing (2026) là các giá trị bức xạ mặt trời đo được tại hai trạm khí tượng thuộc Đồng bằng sông Cửu Long - Việt Nam. Một trạm đặt tại Cảng hàng không Quốc tế Cần Thơ (10.0851° B, 105.7120° Đ) thuộc thành phố Cần Thơ, và một trạm

đặt tại phường Rạch Giá (10.0033° B, 105.0459° Đ) thuộc tỉnh An Giang. Phạm vi dữ liệu được lấy từ ngày 01/01/2020 đến ngày 31/12/2025, mỗi ngày lấy dữ liệu hàng giờ trong khung thời gian 9:00 – 17:00, tương ứng một mẫu dữ liệu gồm 19.728 quan sát với 9 quan sát mỗi ngày ở mỗi bộ dữ liệu.



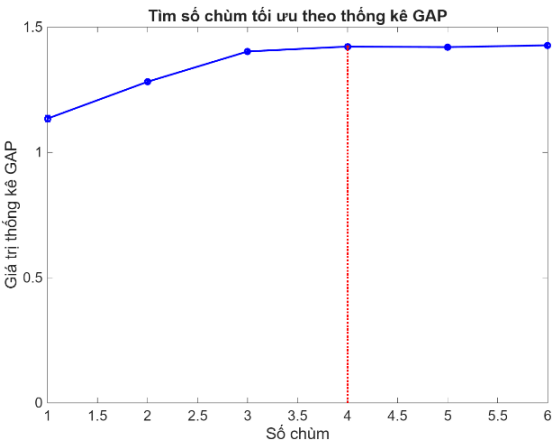
Hình 7. Bức xạ ngoài khí quyển, bức xạ đo được (đồ thị trên) và dãy các chỉ số sáng tính được ở một ngày tại trạm đo Cần Thơ (đồ thị dưới)

Ở mỗi bộ dữ liệu đo được ở 2 trạm, từ dữ liệu bức xạ mặt trời ghi lưu hàng giờ G_h trong khung thời gian 9:00 – 17:00 với 9 quan sát mỗi ngày (dãy các điểm quan sát ở đường đứt khúc màu đỏ thuộc phần đồ thị trên trong minh họa ở Hình 7), cùng với các giá trị bức xạ ngoài khí quyển H_h được tính từ các phương trình thiên văn (dãy các điểm giá trị ở đường liền màu xanh phần đồ thị trên của Hình 7), tính được dãy các chỉ số sáng hàng giờ $K_h = \frac{G_h}{H_h}$ (dãy 9 điểm giá trị ở đồ thị dưới trong Hình 7). Sau đó, việc ước lượng PDF cho từng dãy chỉ số sáng của mỗi ngày được tiến hành theo phương pháp hàm hạt nhân rồi lập các histogram tương ứng từ các PDF này với $l = 20$ khoảng chia đều nhau trên đoạn $[0;1]$. Từ đó, kết quả nhận được hai bộ dữ liệu histogram tương ứng với hai trạm khí tượng Cần Thơ và Rạch Giá với kích thước $n = 2.192$ histogram mỗi bộ.

Kết quả áp dụng phân cụm cho từng bộ dữ liệu ở hai trạm Cần Thơ và Rạch Giá lần lượt được trình bày tiếp theo dưới đây; trong đó, các phương pháp *k-means*, *clustergram* được áp dụng trên ma trận dữ liệu các chỉ số sáng kích thước 2.192×9 , còn phương pháp *mixD-SAEM* được áp dụng trên dãy $n = 2.192$ histogram lập được như trên.

5.3.2. Kết quả phân cụm trên bộ dữ liệu ở trạm Cần Thơ

Phương pháp thống kê Gap được sử dụng trên bộ dữ liệu ở trạm Cần Thơ. Kết quả xác định được số cụm tối ưu là $K = 4$ (Hình 8).



Hình 8. Xác định số cụm tối ưu với bộ dữ liệu ở trạm Cần Thơ

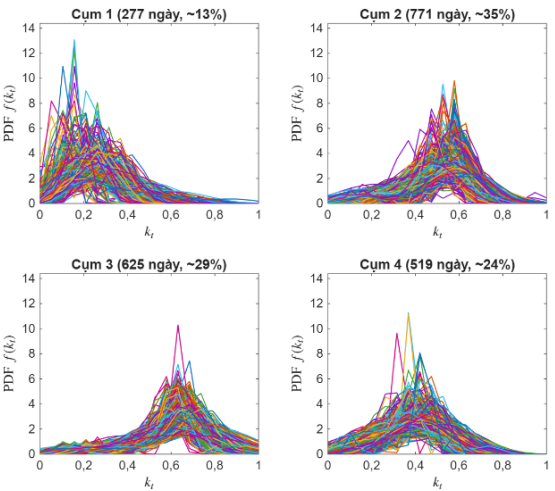
Bảng 2. Đánh giá kết quả phân cụm ngày bức xạ mặt trời trên dữ liệu có số quan sát trong ngày nhỏ do được ở trạm Cần Thơ

Phương pháp	Chỉ số Silhouette		Chỉ số Davies-Bouldin
	Tỷ lệ dương (%)	Trung bình	
<i>k-means</i>	91,06	0,3554	1,4175
<i>clustergram</i>	89,01	0,4040	0,8699
<i>mixD-SAEM</i>	77,87	0,2975	0,6777

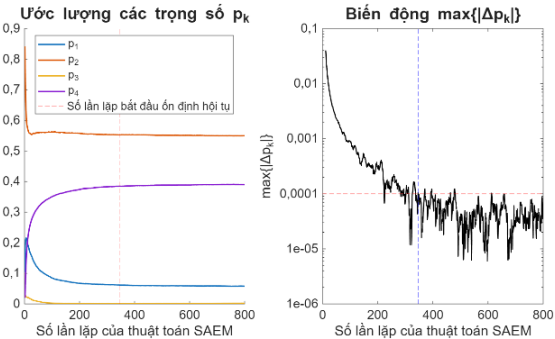
Các phương pháp phân cụm *k-means*, *mixD-SAEM* (với $K = 4$ cụm) và phương pháp *clustergram* được áp dụng trên bộ dữ liệu này. Kết quả đánh giá hiệu quả theo các chỉ số Shilhouette và Davies-Bouldin được trình bày ở Bảng 2. Theo đó, phương pháp *k-means* có tỷ lệ phân cụm tốt (chỉ số Silhouette dương) là cao nhất, còn phương pháp *clustergram* thì lại có trung bình chỉ số Silhouette cao nhất, chứng tỏ phương pháp này mặc dù có thể có phần phân cụm không tốt nhưng chắc chắn kèm theo sẽ có tỷ lệ vùng phân cụm rất tốt. Còn đối với chỉ số Davies-Bouldin thì phương pháp *mixD-SAEM* được đánh giá là tốt nhất (chỉ số Davies-Bouldin thấp nhất); bên cạnh đó, các thông tin về chỉ số Silhouette đánh giá phương pháp này cũng không quá thấp so với các giá trị tương ứng của hai phương pháp đối sánh, đặc biệt là với giá trị trung bình Silhouette.

Kết quả phân cụm (Hình 9) nhận được từ phương pháp *mixD-SAEM* nhờ vào thuật toán SAEM sau

800 bước lặp với sự ổn định hội tụ được thể hiện ở Hình 10.



Hình 9. Kết quả phân cụm theo các phương pháp *mixD-SAEM* trên bộ dữ liệu trạm Cần Thơ

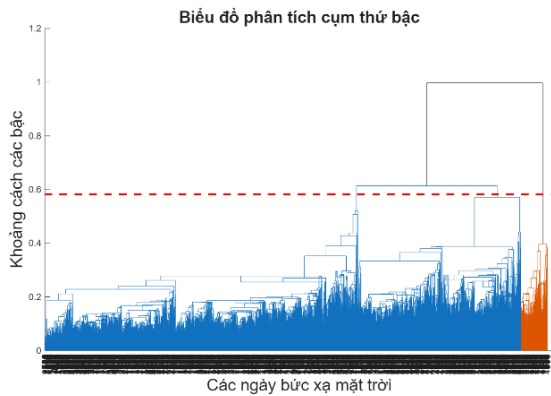


Hình 10. Quá trình biến thiên đến biến thiên ước lượng các trọng số qua thuật toán SAEM với dữ liệu ở trạm Cần Thơ

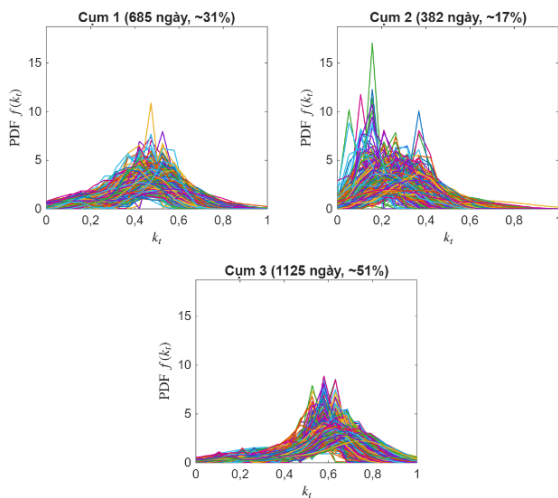
5.3.3. Kết quả phân cụm trên bộ dữ liệu ở trạm Rạch Giá

Bảng 3. Đánh giá kết quả phân cụm ngày bức xạ mặt trời trên dữ liệu có số quan sát trong ngày nhỏ do được ở trạm Rạch Giá

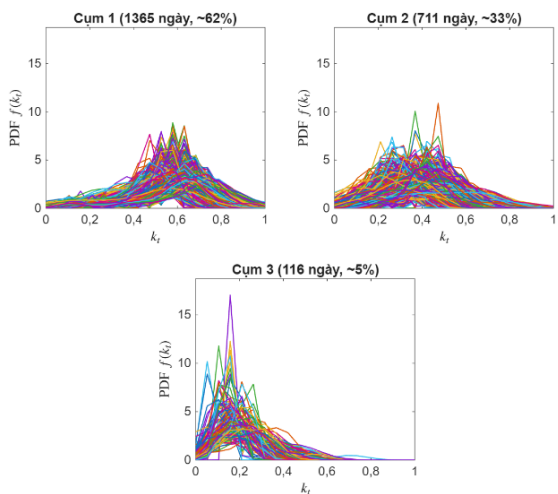
Phương pháp	Chỉ số Silhouette		Chỉ số Davies-Bouldin
	Tỷ lệ dương (%)	Trung bình	
<i>k-means</i>	95,48	0,5070	1,1584
<i>clustergram</i>	90,65	0,5357	1,0126
<i>mixD-SAEM</i>	76,64	0,2986	1,0115



Hình 11. Biểu đồ phân cụm thứ bậc với bộ dữ liệu ở trạm Rạch Giá

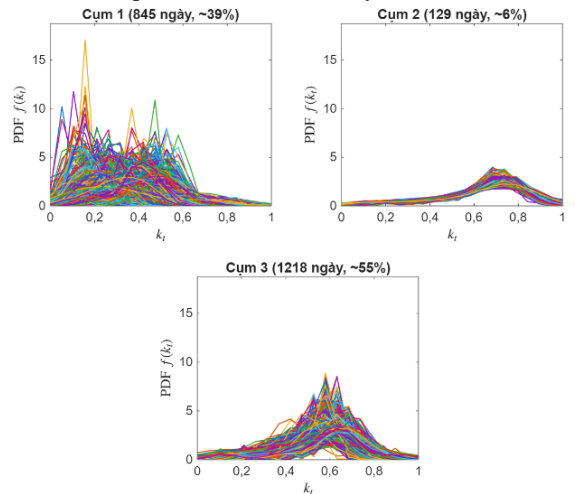


Hình 12. Kết quả phân cụm theo các phương pháp *k-means* trên bộ dữ liệu trạm Rạch Giá



Hình 13. Kết quả phân cụm theo các phương pháp *clustergram* trên bộ dữ liệu trạm Rạch Giá

Phương pháp thống kê Gap được áp dụng trên bộ dữ liệu này nhận được chỉ định số cụm tối ưu là $K = 3$. Kết hợp với quan sát biểu đồ cây phân cụm thứ bậc ở Hình 11, các phương pháp *k-mean* và *mixD-SAEM* được áp dụng thực nghiệm với số cụm này. Các kết quả phân cụm của các phương pháp được thể hiện ở Hình 12 (*k-means*), Hình 13 (*clustergram*) và Hình 14 (*mixD-SAEM*). Các chỉ số Silhouette và Davies-Bouldin đánh giá hiệu quả phân cụm được trình bày ở Bảng 3. Việc so sánh cho thấy các chỉ số ở Bảng 2 và Bảng 3 là tương tự nhau, điều này cho thấy sự ổn định trong thực nghiệm các phương pháp và sự tin cậy trong các nhận định đúc kết được ở phần kết luận dưới đây.



Hình 14. Kết quả phân cụm theo các phương pháp *mixD-SAEM* trên bộ dữ liệu trạm Rạch Giá

6. KẾT LUẬN

Kết quả nghiên cứu này đã khảo sát khả năng áp dụng mô hình phân phối hỗn hợp Dirichlet kết hợp với thuật toán xấp xỉ ngẫu nhiên cực đại hóa kỳ vọng *mixD-SAEM* trong bài toán phân loại các dạng ngày bức xạ mặt trời. Điểm khác biệt của cách tiếp cận này là khai thác trực tiếp đặc điểm phân phối xác suất của dữ liệu thông qua các histogram chỉ số sáng, thay vì dựa chủ yếu vào khoảng cách hình học giữa các đối tượng như các phương pháp phân cụm truyền thống. Để đánh giá hiệu quả của phương pháp, các kết quả nhận được từ *mixD-SAEM* đã được so sánh với hai phương pháp phân cụm phổ biến là *k-means* và phân cụm thứ bậc.

Các thực nghiệm được tiến hành trên hai nhóm dữ liệu có đặc tính khác nhau. Nhóm thứ nhất là bộ dữ liệu bức xạ mặt trời được sử dụng trong nghiên cứu của Soubdhan et al. (2009); trong đó, số lượng quan sát thu được trong mỗi ngày rất lớn nên các

histogram chỉ số sáng phản ánh khá đầy đủ đặc điểm phân phối của dữ liệu. Nhóm thứ hai là các bộ dữ liệu bức xạ mặt trời ghi nhận tại hai trạm khí tượng ở Đồng bằng sông Cửu Long, nơi chỉ có 9 quan sát mỗi ngày. Đối với trường hợp này, các histogram được xây dựng từ các hàm mật độ xác suất ước lượng bằng phương pháp hàm hạt nhân nhằm khắc phục hạn chế về số lượng dữ liệu quan sát.

Kết quả thực nghiệm trên bộ dữ liệu có số lượng quan sát lớn cho thấy phương pháp *mixD-SAEM* đạt hiệu quả phân cụm vượt trội so với *k-means* và phân cụm thứ bậc. Các chỉ số đánh giá Silhouette và Davies–Bouldin đều cho thấy các cụm nhận được từ *mixD-SAEM* có mức độ đồng nhất nội cụm cao hơn và khả năng tách biệt giữa các cụm tốt hơn. Đặc biệt, với số cụm $K = 4$ như đề xuất trong nghiên cứu gốc của Soubdhan et al. (2009), phương pháp *mixD-SAEM* cho kết quả phân loại rõ ràng và ổn định hơn đáng kể so với các phương pháp đối sánh.

Đối với các bộ dữ liệu đo được tại Cần Thơ và Rạch Giá, mặc dù số lượng quan sát trong mỗi ngày rất hạn chế nhưng kết quả thực nghiệm cho thấy phương pháp *mixD-SAEM* vẫn có khả năng nhận dạng được các dạng ngày bức xạ mặt trời khác nhau thông qua các histogram được xây dựng từ các hàm mật độ xác suất ước lượng. Các chỉ số đánh giá cho thấy hiệu quả của *mixD-SAEM* nhìn chung tương đương với các phương pháp phân cụm truyền thống. Trong một số trường hợp, phương pháp này đạt giá trị Davies–Bouldin tốt nhất, phản ánh khả năng tạo ra các cụm có tính phân tách cao. Điều này cho thấy

việc mô hình hóa dữ liệu dưới dạng phân phối xác suất vẫn mang lại những lợi thế nhất định ngay cả khi số lượng quan sát ban đầu không lớn.

Một kết quả đáng chú ý khác là số lượng cụm tối ưu được xác định bởi thống kê Gap không phải lúc nào cũng tương ứng với chất lượng phân cụm tốt nhất đối với mô hình hỗn hợp Dirichlet. Kết quả quan sát này cho thấy việc lựa chọn số cụm đối với phương pháp *mixD-SAEM* cần được nghiên cứu riêng, có thể dựa trên các tiêu chuẩn thông tin hoặc các thước đo đánh giá phù hợp hơn với dữ liệu dạng phân phối.

Các kết quả thu được cho thấy phương pháp *mixD-SAEM* là một công cụ hiệu quả trong phân loại các dạng ngày bức xạ mặt trời, đặc biệt đối với các bài toán mà đặc trưng quan trọng nằm ở hình dạng phân phối của dữ liệu. Nghiên cứu cũng cho thấy tiềm năng áp dụng phương pháp này trong các bài toán phân tích và dự báo năng lượng mặt trời tại Việt Nam. Trong các nghiên cứu tiếp theo, việc phát triển các tiêu chí lựa chọn số cụm phù hợp hơn cho mô hình hỗn hợp Dirichlet, kết hợp với các mô hình chuyển trạng thái như chuỗi Markov hoặc mô hình Markov ẩn để mô tả động lực chuyển đổi giữa các dạng ngày bức xạ, là những hướng nghiên cứu có nhiều triển vọng.

LỜI CẢM ƠN

Đề tài nghiên cứu này được tài trợ bởi Đại học Cần Thơ, mã số: CTCS2024-10-05.

TÀI LIỆU THAM KHẢO

- Davies, D. L., & Bouldin, D. W. (1979). A cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 1(2), 224 – 227.
doi Bookmark: 10.1109/tpami.1979.4766909
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1), 1–38.
<https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
- Dembo, A., & Zeitouni, O. (1986). Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stochastic Processes and Their Applications*, 23(1), 91–113.
[https://doi.org/10.1016/0304-4149\(86\)90018-9](https://doi.org/10.1016/0304-4149(86)90018-9)
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65
[https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Soubdhan, T., Emilion, R., & Calif, R. (2009). Classification of daily solar radiation distributions using a mixture of Dirichlet distributions. *Solar Energy*, 83(7), 1056–1063.
<https://doi.org/10.1016/j.solener.2009.01.010>
- Sen, Z. (2008). *Solar Energy Fundamentals and Modeling Techniques*. Springer.
<https://doi.org/10.1007/978-1-84800-134-3>
- Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the Gap statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(2), 411–423.
<https://doi.org/10.1111/1467-9868.00293>
- Visual Crossing. (2026). *Weather Data for Solar Energy: 2020–2025*.
<https://www.visualcrossing.com/solar-energy>