



A STUDY ON THE APPLICATION OF UNSUPERVISED CLUSTERING ALGORITHMS IN GNSS-RTK DATA ANALYSIS FOR CABLE-STAYED BRIDGES MONITORING

Tran Duc Cong, Ho Thi Lan Huong, Le Khanh Giang*, Le Van Hien

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article

Received: 19/08/2025

Revised: 11/10/2025

Accepted: 13/10/2025

Published online: 15/10/2025

<https://doi.org/10.47869/tcsj.76.8.8>

* *Corresponding author*

Email: giangk@utc.edu.vn; Tel: +84983663031

Abstract. GNSS-RTK satellite positioning technology has been widely applied in structural health monitoring (SHM) of cable-stayed bridges due to its ability to provide accurate and continuous displacement data. However, a major challenge is that the monitoring data often contain noise, outliers, and lack labels, which makes early detection of abnormal structural states difficult. This study focuses on analyzing vertical GNSS-RTK displacement time series from the Nhat Tan cable-stayed bridge (Hanoi) using two unsupervised clustering algorithms: Hierarchical Agglomerative Clustering (HAC) and the Gaussian Mixture Model (GMM). The analysis procedure includes data preprocessing (outlier removal using the Hampel filter, interpolation, segmentation, and z-score normalization) and cluster validation through the Silhouette index. The results show that both HAC and GMM identify the optimal number of clusters as $k = 2$, representing two distinct displacement states, with GMM achieving better clustering performance (Silhouette ≈ 0.596) compared to HAC (≈ 0.520). These findings confirm the feasibility of applying unsupervised clustering for early detection of abnormal states in cable-stayed bridges, thereby enhancing proactive maintenance efficiency and operational safety.

Keywords: GNSS-RTK, HAC clustering, Gaussian Mixture Model (GMM), Silhouette index, unsupervised machine learning, cable-stayed bridge monitoring

@ 2025 University of Transport and Communications



NGHIÊN CỨU ỨNG DỤNG THUẬT TOÁN PHÂN CỤM KHÔNG GIÁM SÁT TRONG PHÂN TÍCH DỮ LIỆU GNSS-RTK PHỤC VỤ QUAN TRẮC CẦU DÂY VĂNG

Trần Đức Công, Hồ Thị Lan Hương, Lê Khánh Giang*, Lê Văn Hiến

Trường Đại học Giao thông vận tải, Số 3 Cầu Giấy, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

CHUYÊN MỤC: Công trình khoa học

Ngày nhận bài: 19/08/2025

Ngày nhận bài sửa: 11/10/2025

Ngày chấp nhận đăng: 13/10/2025

Ngày xuất bản Online: 15/10/2025

<https://doi.org/10.47869/tcsj.76.8.8>

* Tác giả liên hệ

Email: gianglk@utc.edu.vn; Tel: +84983663031

Tóm tắt. Công nghệ định vị vệ tinh GNSS-RTK đã và đang được ứng dụng phổ biến trong quan trắc sức khỏe kết cấu cầu dây văng nhờ khả năng cung cấp dữ liệu dịch chuyển chính xác và liên tục. Tuy nhiên, thách thức lớn đặt ra là dữ liệu thường chứa nhiều nhiễu, giá trị ngoại lai và không có nhãn, gây khó khăn cho việc phát hiện sớm các trạng thái bất thường của công trình. Nghiên cứu này tập trung vào phân tích chuỗi dịch chuyển GNSS-RTK theo phương đứng tại cầu Nhật Tân (Hà Nội) thông qua hai thuật toán phân cụm không giám sát: Phân cụm phân cấp kết hợp dần (HAC) và Mô hình hỗn hợp Gaussian (GMM). Quy trình phân tích bao gồm tiền xử lý dữ liệu (loại bỏ ngoại lai bằng bộ lọc Hampel, nội suy, phân đoạn và chuẩn hóa z-score) và đánh giá kết quả bằng chỉ số Silhouette. Kết quả cho thấy cả HAC và GMM đều xác định số cụm tối ưu $k = 2$, phản ánh hai trạng thái dịch chuyển đặc trưng, trong đó GMM đạt chất lượng phân cụm tốt hơn (Silhouette $\approx 0,596$) so với HAC ($\approx 0,520$). Kết quả nghiên cứu khẳng định tính khả thi của việc ứng dụng phân cụm không giám sát trong phát hiện sớm các trạng thái bất thường của cầu dây văng, góp phần nâng cao hiệu quả bảo trì chủ động và an toàn khai thác.

Từ khóa: GNSS-RTK, phân cụm HAC, mô hình hỗn hợp Gaussian (GMM), chỉ số Silhouette, học máy không giám sát, quan trắc cầu dây văng

@ 2025 Trường Đại học Giao thông vận tải

1. ĐẶT VẤN ĐỀ

Hệ thống định vị vệ tinh toàn cầu – Định vị động thời gian thực (Global Navigation Satellite System – Real-Time Kinematic - GNSS-RTK) đang được ứng dụng ngày càng rộng rãi trong hệ thống theo dõi biến dạng công trình cầu, đặc biệt là các cầu dây văng lớn (cầu dây văng, cầu treo) trên thế giới. Ưu điểm nổi bật của GNSS-RTK là cung cấp dữ liệu dịch chuyển với độ chính xác cao và gần thời gian thực, cho phép đánh giá mức độ ổn định kết cấu dưới tác động của tải trọng và điều kiện môi trường đa dạng [1]. Tại Việt Nam, từ sau năm 2010, GNSS-RTK đã được tích hợp vào hệ thống quan trắc sức khỏe kết cấu (Structural Health Monitoring – SHM) trên nhiều cầu dây văng lớn như Cầu Thơ, Bạch Đằng, Nhật Tân, Vàm Cống... Một số nghiên cứu trong nước tập trung xử lý dữ liệu GNSS-RTK nhằm khử nhiễu, loại bỏ dữ liệu ngoại lai và bù dữ liệu thiếu để nâng cao độ tin cậy phép đo [2].

Tuy nhiên, việc phân tích chuyên sâu chuỗi dữ liệu GNSS-RTK dài hạn vẫn còn hạn chế, dẫn đến chưa khai thác hết tiềm năng của công nghệ này trong thực tiễn vận hành cầu [3]. Trong nhiều trường hợp, lượng dữ liệu khổng lồ thu thập được chỉ được lưu trữ dưới dạng thô, không qua xử lý phân tích, gây ra tình trạng “bội thực dữ liệu” và khó khăn trong phát hiện sớm các bất thường [3]. Bên cạnh đó, do sự cố hư hỏng nghiêm trọng hiếm khi xảy ra, dữ liệu gần như trạng thái hư hỏng thực tế gần như không tồn tại, khiến các phương pháp học máy có giám sát khó áp dụng. Điều này đặt ra nhu cầu áp dụng phương pháp học máy không giám sát để tự động phát hiện các thay đổi bất thường trong hành vi kết cấu dựa trên dữ liệu quan trắc liên tục [4].

Phân cụm (clustering) là một kỹ thuật học máy không giám sát quan trọng, cho phép phân nhóm dữ liệu theo mức độ tương đồng mà không cần nhãn trước [5]. Trong bối cảnh quan trắc cầu, phân cụm chuỗi thời gian dịch chuyển có thể giúp nhận diện các trạng thái vận hành khác nhau, từ bình thường ổn định đến dao động bất thường vượt ngưỡng. Các nghiên cứu trước đã chỉ ra tiềm năng của phương pháp này, ví dụ Diez và cộng sự áp dụng phân cụm để phát hiện các vị trí bất thường trên cầu Sydney [4]. Trong bối cảnh quốc tế, các thuật toán phân cụm phân cấp kết hợp dần (HAC) và mô hình hỗn hợp Gaussian (GMM) đã được ứng dụng trong nhiều bài toán phân cụm dữ liệu kỹ thuật, bao gồm cả dữ liệu dịch chuyển cầu. Tuy nhiên, tại Việt Nam, theo hiểu biết của nhóm tác giả, chưa có nghiên cứu nào đánh giá so sánh trực tiếp hiệu quả của HAC và GMM trên dữ liệu GNSS-RTK thực tế từ các cầu dây văng, đặc biệt với dữ liệu quan trắc dài hạn có đặc trưng nhiễu cao, phân bố không đồng đều và không nhãn. Xuất phát từ khoảng trống nghiên cứu trên, nghiên cứu này đề xuất và đánh giá khả năng của hai thuật toán phân cụm tiêu biểu thuộc hai hướng tiếp cận khác nhau:

- HAC đại diện cho phương pháp phân cấp, cung cấp trực quan dạng dendrogram (sơ đồ cây phân cụm) và không yêu cầu biết trước số cụm.
- GMM đại diện cho phương pháp mô hình xác suất, cho phép phân cụm mềm và đánh giá xác suất thuộc cụm của từng điểm.

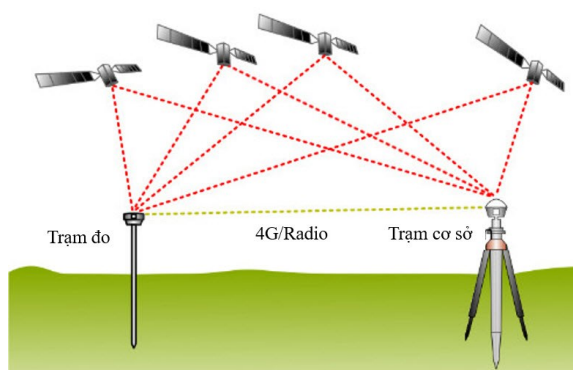
Điểm mới của nghiên cứu nằm ở việc tích hợp một quy trình tiền xử lý dữ liệu GNSS-RTK bài bản (lọc ngoại lai bằng Hampel, nội suy, phân đoạn, chuẩn hóa z-score) trước khi phân cụm, và so sánh toàn diện HAC và GMM trên dữ liệu GNSS-RTK thực tế từ cầu Nhật Tân (Việt Nam). Chỉ số Silhouette được sử dụng làm thước đo hiệu quả phân cụm, hỗ trợ xác định số cụm tối ưu khi không có nhãn dữ liệu. Nghiên cứu này không chỉ cung cấp bằng chứng thực nghiệm về hiệu quả của các phương pháp phân cụm không giám sát trong giám sát biến dạng cầu dây văng, mà còn gợi mở khả năng tích hợp chúng vào các mô-đun cảnh báo sớm trong hệ thống SHM tại Việt Nam.

2. CƠ SỞ LÝ THUYẾT NGHIÊN CỨU

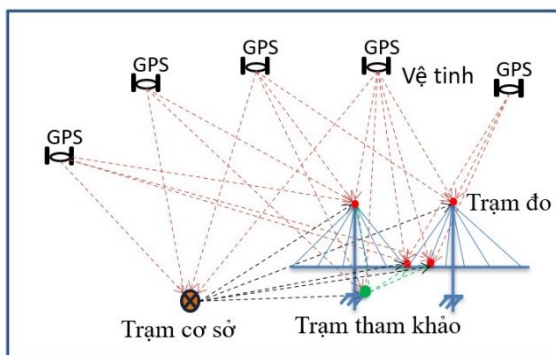
2.1. Nguyên lý hoạt động của GNSS-RTK trong quan trắc cầu dây văng

Phương pháp định vị GNSS-RTK thực hiện dựa trên nguyên tắc định vị tương đối của GNSS. Theo phương pháp này, trạm cơ sở được đặt tại một điểm không chế đã biết tọa độ, một hoặc nhiều trạm đo đặt tại các điểm cần xác định. Hai trạm cơ sở và trạm đo liên kết với nhau thông qua sóng viễn thông 4G (phổ biến hiện nay) hoặc sóng radio và cùng thu nhận tín hiệu từ các vệ tinh tại cùng một thời điểm đo theo thời gian thực. Trạm cơ sở thu nhận tín hiệu vệ tinh, xử lý và xác định tọa độ tức thời của điểm cơ sở, so sánh với tọa độ đã biết để xác định đại lượng hiệu chỉnh và truyền đại lượng hiệu chỉnh đến các trạm đo thông qua phương thức 4G hoặc radio. Hình 1a mô tả nguyên tắc cơ bản của phương pháp định vị động GNSS-RTK.

Nguyên tắc định vị GNSS-RTK ứng dụng trong các hệ thống quan trắc SHM của cầu dây văng gồm có một trạm cơ sở thường được xây dựng gần với công trình cầu và các trạm đo được đặt cố định tại các điểm đặc trưng cần quan trắc trên cầu như: tại đỉnh các tháp; chính giữa các nhịp của cầu; tại đỉnh các trụ... Các đầu thu GNSS (sensors) được kết nối về đầu thu dữ liệu thông qua cáp nối và dữ liệu đo được quản lý trên một hệ thống chung. Hình 1b mô tả sơ đồ quan trắc GNSS-RTK của một cầu dây văng có hai trụ tháp.



a) Nguyên tắc định vị GNSS-RTK [6].



b) Sơ đồ quan trắc GNSS-RTK của cầu dây văng [2].

Hình 1. Nguyên lý định vị GNSS-RTK và ứng dụng trong quan trắc cầu dây văng.

2.2. Cơ sở lý thuyết về thuật toán HAC và GMM

Phân cụm dữ liệu là một nhiệm vụ cốt lõi trong lĩnh vực học không giám sát (unsupervised learning). Mục tiêu chính của nó là phân chia một tập hợp các đối tượng thành các nhóm, được gọi là các cụm (clusters), sao cho các đối tượng trong cùng một cụm có mức độ tương đồng với nhau cao hơn so với các đối tượng trong các cụm khác [7]. Về bản chất, phân cụm là một quá trình khám phá cấu trúc, tìm kiếm các mẫu hoặc các nhóm tiềm ẩn trong dữ liệu mà không có bất kỳ kiến thức tiên nghiệm nào về các nhãn của dữ liệu [8].

Thách thức cơ bản và cũng là đặc điểm xác định của học không giám sát là sự thiếu vắng của các nhãn "sự thật nền" (ground-truth labels). Do đó, việc đánh giá chất lượng của một kết quả phân cụm không thể dựa vào việc so sánh với một đáp án có sẵn. Thay vào đó, việc đánh giá này trở nên chủ quan hơn và đòi hỏi phải sử dụng các chỉ số xác thực nội tại (internal validation metrics). Các chỉ số này đánh giá "tính tốt" của cấu trúc cụm chỉ dựa trên chính dữ liệu và kết quả phân cụm, bằng cách đo lường các đặc tính như độ cô đọng (compactness) trong cụm và độ tách biệt (separation) giữa các cụm [9].

Điều quan trọng là phải phân biệt rõ ràng giữa phân cụm không giám sát và phân loại có giám sát (supervised classification). Trong khi phân loại là một nhiệm vụ dự đoán, sử dụng dữ liệu đã được gán nhãn để huấn luyện một mô hình có khả năng gán nhãn cho dữ liệu mới, thì phân cụm lại mang tính chất khám phá. Nó không dự đoán một nhãn đã biết trước, mà thay vào đó là phát hiện ra các nhóm có ý nghĩa trong dữ liệu, từ đó cung cấp những hiểu biết sâu sắc về cấu trúc vốn có của dữ liệu đó [5].

2.2.1. Thuật toán phân cụm HAC

HAC là một thuật toán phân cụm không giám sát thuộc nhóm phân cụm phân cấp. Phương pháp hoạt động theo hướng kết hợp dần từ dưới lên (bottom-up), khởi đầu với mỗi điểm dữ liệu là một cụm riêng biệt. Ở mỗi bước, hai cụm gần nhau nhất sẽ được hợp nhất, quá trình này lặp lại cho đến khi chỉ còn một cụm duy nhất hoặc đạt được số cụm mong muốn [10].

Khoảng cách giữa các cụm có thể được xác định theo nhiều tiêu chí khác nhau, bao gồm: khoảng cách gần nhất (single linkage), khoảng cách xa nhất (complete linkage), khoảng cách trung bình (average linkage), hoặc khoảng cách giữa các trọng tâm cụm (centroid linkage). Tuy nhiên, trong nghiên cứu này, thuật toán HAC được triển khai với tiêu chí liên kết theo phương sai tối thiểu (ward linkage) [11].

Phương pháp liên kết ward là một tiêu chí đặc biệt trong phân cụm phân cấp. Thay vì chỉ xét khoảng cách hình học giữa các điểm hoặc cụm, ward linkage thực hiện việc hợp nhất hai cụm sao cho làm giảm tối thiểu sự gia tăng tổng bình phương sai số trong cụm (Within-Cluster Sum of Squares - WCSS). Phương pháp này có ưu điểm là tạo ra các cụm có kích thước tương đối đồng đều và giảm thiểu sự phân tán nội cụm, do đó rất phù hợp với các tập dữ liệu dạng số có độ phân tán thấp hoặc vừa phải [11].

Công thức tính khoảng cách ward giữa hai cụm A và B được định nghĩa như sau [11]:

$$D_{\text{ward}}(A, B) = \frac{|A| \cdot |B|}{|A| + |B|} \cdot \|\bar{x}_A - \bar{x}_B\|^2 \quad (1)$$

Trong đó:

$|A|, |B|$: số lượng phần tử trong cụm A và B,

$\bar{x}_A, \bar{x}_B$: vector trung bình (centroid) của cụm A và B,

$\|\cdot\|$: chuẩn Euclidean.

Kết quả của HAC có thể được biểu diễn dưới dạng cây phân cấp (dendrogram), từ đó có thể "cắt cây" ở các ngưỡng khác nhau để xác định số cụm phù hợp. HAC có ưu điểm là không yêu cầu xác định số cụm trước và cung cấp trực quan phân cấp giữa các cụm, nhưng hạn chế về hiệu suất tính toán với dữ liệu lớn và nhạy cảm với nhiễu.

2.2.2. Thuật toán phân cụm GMM

GMM là phương pháp phân cụm xác suất, giả định rằng dữ liệu được tạo ra từ sự kết hợp của nhiều phân phối Gaussian (chuẩn) khác nhau. Mỗi cụm tương ứng với một thành phần Gaussian trong mô hình [12].

Hàm mật độ xác suất của GMM là [12]:

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) \quad (2)$$

Trong đó:

x : điểm dữ liệu cần phân cụm

K : số cụm

π_k : trọng số của cụm thứ k , thỏa mãn $\sum_{k=1}^K \pi_k = 1$

μ_k : vector trung tâm (mean) của cụm k

Σ_k : ma trận hiệp phương sai (covariance) của cụm k

$\mathcal{N}(x | \mu_k, \Sigma_k)$: mật độ xác suất Gaussian đa chiều:

$$\mathcal{N}(x | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right) \quad (3)$$

Ước lượng tham số: Thuật toán Expectation-Maximization (EM) [12]

Bước 1 (E-step): Tính xác suất điểm dữ liệu x_i thuộc về cụm k

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)} \quad (4)$$

Bước 2 (M-step): Cập nhật tham số dựa trên γ_{ik} :

$$N_k = \sum_{i=1}^N \gamma_{ik}, \mu_k = \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} x_i \quad (5)$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^N \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T, \pi_k = \frac{N_k}{N} \quad (6)$$

Lặp lại hai bước trên cho đến khi hội tụ. GMM linh hoạt và hiệu quả trong việc mô hình hóa các cụm có hình dạng phức tạp, nhưng dễ bị ảnh hưởng bởi nhiễu và yêu cầu chọn số cụm trước, đồng thời có thể hội tụ vào nghiệm cục bộ.

2.2.3. Đánh giá hiệu quả phân cụm bằng chỉ số Silhouette

Vì phân cụm là một nhiệm vụ không giám sát, không có nhãn bên ngoài để đánh giá. Do đó, các chỉ số xác thực nội tại là cần thiết để đánh giá chất lượng của các cụm kết quả chỉ dựa trên dữ liệu và chính sự phân hoạch. Mục tiêu của các chỉ số này là cung cấp một điểm số định lượng để giúp xác định số lượng cụm tối ưu (k) cho một tập dữ liệu và thuật toán nhất định. Phương pháp điển hình là chạy thuật toán phân cụm với một loạt các giá trị k và vẽ biểu đồ các điểm số chỉ số kết quả, tìm kiếm một giá trị tối ưu (ví dụ: một giá trị cực đại hoặc cực tiểu).

Chỉ số Silhouette là một phương pháp để diễn giải và xác thực tính nhất quán bên trong các cụm dữ liệu. Nó đo lường mức độ tương đồng của một đối tượng với cụm của chính nó (gắn kết) so với các cụm khác (tách biệt) [13].

Điểm số cho một điểm dữ liệu duy nhất i được tính như sau [14]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (7)$$

Trong đó :

$a(i)$ (Cohesion - Độ gắn kết): Khoảng cách trung bình từ điểm i đến tất cả các điểm khác trong cùng một cụm. Giá trị $a(i)$ nhỏ có nghĩa là điểm đó rất phù hợp với cụm của nó.

$b(i)$ (Separation - Độ tách biệt): Khoảng cách trung bình nhỏ nhất từ điểm i đến tất cả các điểm trong bất kỳ cụm nào khác. Giá trị $b(i)$ lớn có nghĩa là điểm đó ở xa cụm lân cận.

Chỉ số $s(i)$ nằm trong khoảng từ -1 đến +1.

- Gần +1: Cho thấy điểm đó ở xa cụm lân cận và rất gần với cụm của chính nó (phân cụm hoàn hảo).
- Bằng 0: Cho thấy điểm đó nằm trên ranh giới giữa hai cụm.

- Gần -1: Cho thấy điểm đó có khả năng bị gán sai cụm.

Điểm số tổng thể là giá trị trung bình của $s(i)$ trên tất cả các điểm. Điểm số trung bình cao hơn cho thấy các cụm được xác định rõ ràng hơn.

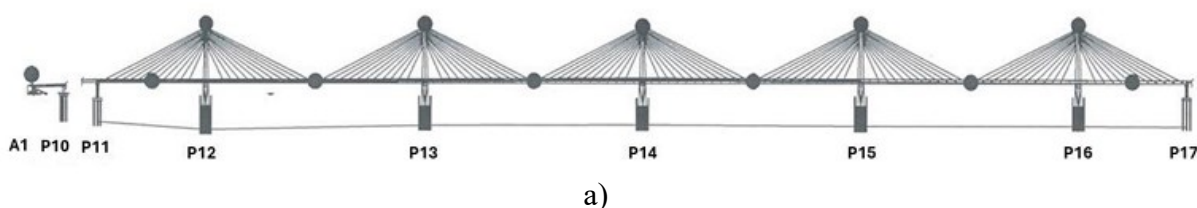
3. TÍNH TOÁN THỰC NGHIỆM

3.1. Dữ liệu nghiên cứu

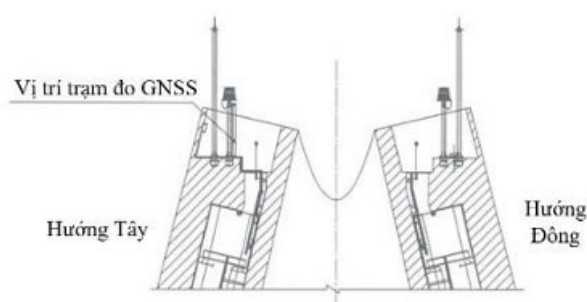
Cầu Nhật Tân là một cầu dây văng nhiều trụ tháp được xây dựng bắc qua sông Hồng và đưa vào khai thác từ tháng 1 năm 2015 (Hình 2). Cầu có tổng chiều dài 3,9 kilomet (trong đó chiều dài vượt sông là 2,5 kilomet) với 5 trụ tháp hình thoi và 6 nhịp dây văng [15]. Hệ thống quan trắc SHM của cầu Nhật Tân được đánh giá là hệ thống đầy đủ các loại cảm biến và hiện đại nhất ở Việt Nam. Trong đó hệ thống quan trắc GNSS bao gồm 1 trạm cơ sở và 17 trạm đo gắn trên 5 đỉnh tháp và 12 điểm trên dầm cầu. Các máy thu GNSS của cầu Nhật Tân sử dụng máy thu của hãng Leica, gồm ăng-ten Leica AS10 cho trạm đo và Leica AR10 cho trạm cơ sở [15]. Hình 3a mô tả sơ đồ vị trí lắp đặt các máy thu GNSS, hình 3b mô tả vị trí lắp đặt máy thu GNSS trên đỉnh tháp của cầu.



Hình 2. Phối cảnh cầu Nhật Tân.



a)

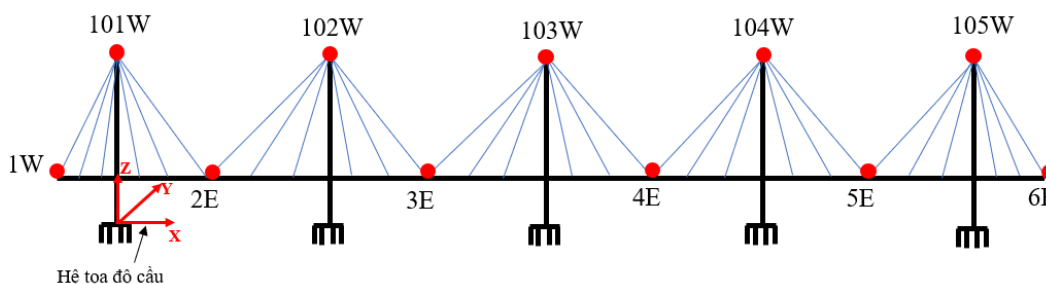


b)

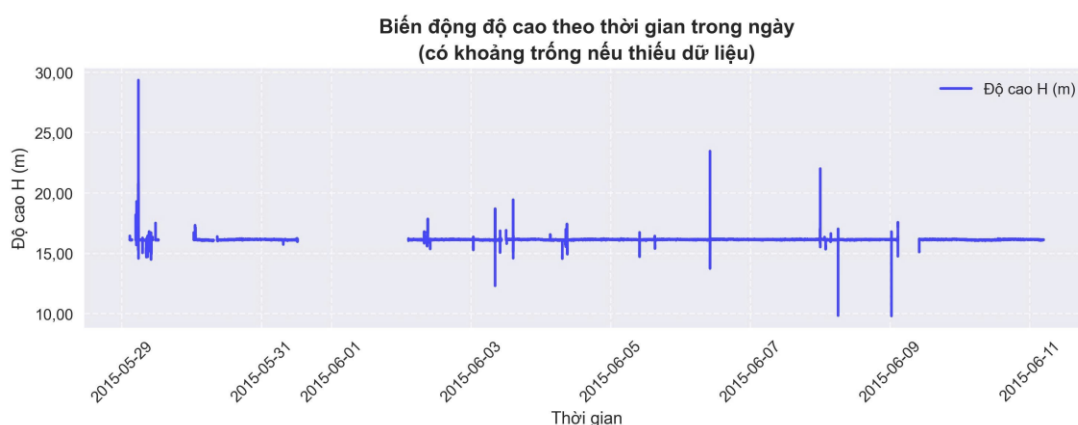
Hình 3. Sơ đồ lắp đặt máy thu GNSS trên cầu dây văng Nhật Tân [16].

Trong nghiên cứu này, nhóm tác giả tập trung vào chuỗi dữ liệu GNSS-RTK chuyển dịch theo thẳng đứng (phương Z) đặt trên dầm giữa nhịp cầu Nhật Tân. Cụ thể, dữ liệu được

lấy tại điểm ký hiệu 3E (Hình 4) trên dầm chính - đây là một trong các vị trí rover trên cầu với tần suất lấy mẫu 1Hz. Chuỗi dữ liệu GNSS-RTK của điểm 3E được thu thập liên tục trong 13 ngày, từ 29/05 đến 11/06 (thời gian này đủ dài để bao quát nhiều điều kiện khai thác cầu khác nhau). Hình 5 biểu diễn một đoạn chuỗi dữ liệu thô thu được, cho thấy dao động chuyển vị theo thời gian của điểm quan trắc.



Hình 4. Sơ đồ điểm quan trắc nghiên cứu (3E).



Hình 5. Biểu diễn dữ liệu thô thu được trong 13 ngày.

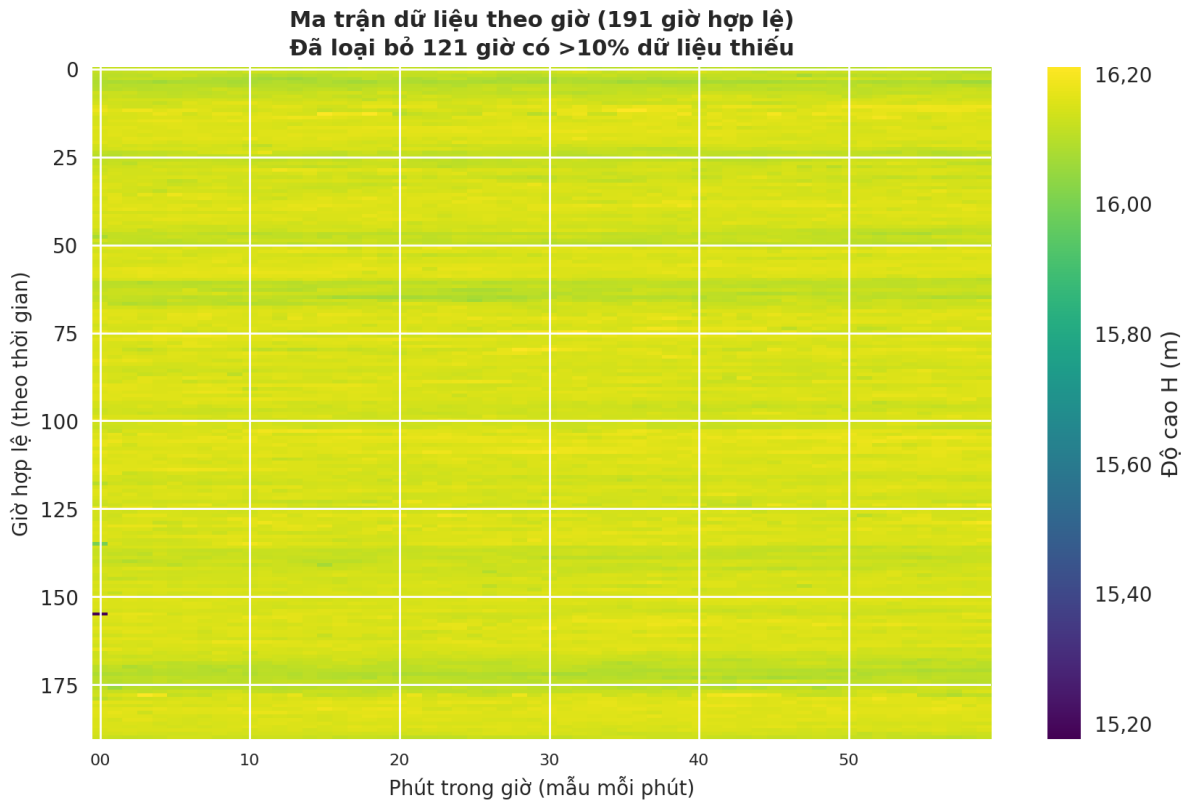
Hình 5 cho thấy dữ liệu GNSS chứa nhiều giá trị bất thường, ảnh hưởng đáng kể đến độ chính xác và độ tin cậy của phân tích kết quả. Do đó, việc tiền xử lý dữ liệu là bước bắt buộc và có ý nghĩa quyết định trong toàn bộ quy trình phân tích.

3.2. Tiền xử lý dữ liệu

Dữ liệu GNSS-RTK thu được tại điểm 3E với tần suất 1 Hz trong 13 ngày (29/05–11/06), tương ứng 1.123.200 mẫu thô. Do một số giờ quan trắc có tỷ lệ dữ liệu thiếu lớn, các giờ có >10% dữ liệu thiếu đã được loại bỏ (121 giờ bị loại), còn lại 191 giờ hợp lệ tương ứng 687.600 mẫu (191×3600). Trên tập hợp 191 giờ hợp lệ này, quá trình tiền xử lý bao gồm ba bước: (i) phát hiện và loại bỏ ngoại lai bằng bộ lọc Hampel, (ii) nội suy các khoảng thiếu nhỏ (<10%), và (iii) rút mẫu (down-sampling) theo bước thời gian 10 s nhằm giảm chi phí tính toán. Tổng số mẫu cuối cùng dùng cho phân cụm là 68.760 ($191 \text{ giờ} \times 360 \text{ điểm/giờ}$).

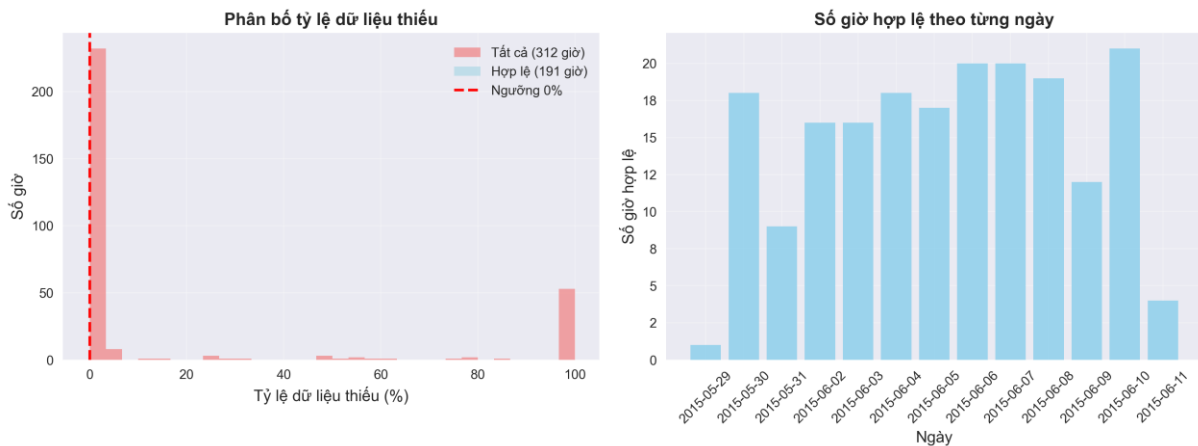
3.2.1. Xử lý dữ liệu bị thiếu (Missing Data Handling)

Sau khi loại bỏ các khoảng thời gian có thiếu dữ liệu lớn hơn 10%, dữ liệu GNSS theo phương Z được tổ chức lại thành ma trận 191 giờ liên tục, mỗi giờ gồm 60 mẫu (1 mẫu/phút).



Hình 6. Ma trận dữ liệu GNSS hợp lệ theo giờ.

Hình 6 minh họa ma trận dữ liệu với trực quan hóa giá trị độ cao qua thang màu, cho thấy dữ liệu đủ điều kiện để áp dụng các thuật toán phân cụm. Từ ma trận dữ liệu có thể thấy các giờ quan trắc có mức độ đầy đủ tương đối cao.



a, Phân bố tỷ lệ dữ liệu thiếu theo giờ.

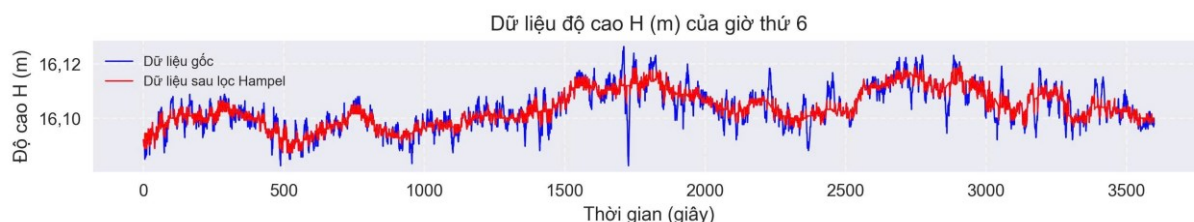
b, Số giờ hợp lệ theo từng ngày.

Hình 7. Phân tích tỷ lệ dữ liệu thiếu và phân bố số giờ hợp lệ theo ngày.

Để làm rõ hơn đặc điểm chất lượng dữ liệu, Hình 7a cho thấy phần lớn các giờ thu thập có tỷ lệ dữ liệu thiếu dưới 10%, đáp ứng điều kiện hợp lệ để đưa vào phân tích. Đồng thời, Hình 7b cũng phản ánh mức độ phân bố số giờ hợp lệ theo từng ngày, qua đó có thể nhận diện được những khoảng thời gian có nguy cơ gián đoạn hoặc thiếu dữ liệu rõ rệt hơn.

3.2.2. Xử lý dữ liệu nhiễu (Outlier Detection and Removal)

Trước khi tiến hành phân tích, dữ liệu GNSS thô đã được tiền xử lý nhằm nâng cao chất lượng. Trước hết, các giá trị ngoại lai (những đo đạc bất thường do nhiễu hoặc lỗi thiết bị) trong chuỗi dữ liệu được nhận diện và hiệu chỉnh bằng phương pháp nội suy Hampel - một kỹ thuật lọc nhiễu dựa trên trung vị trượt giúp loại bỏ các dữ liệu bất thường và nội suy khoảng trống do mất số liệu (các trường hợp dưới 10%). Kết quả nội suy cho thấy chuỗi dữ liệu mượt hơn và phản ánh chính xác xu hướng dịch chuyển của kết cấu hơn (Hình 8 minh họa chuỗi dữ liệu trước và sau khi lọc ngoại lai).



Hình 8. Kết quả nội suy dữ liệu bất thường của điểm 3E.

3.2.3. Chuẩn hoá dữ liệu (Data Normalization)

Tiếp theo, toàn bộ dữ liệu sau khi xử lý được chuẩn hóa về dạng không thứ nguyên, với giá trị trung bình bằng 0 và phương sai bằng 1 (z-score normalization). Việc chuẩn hóa này đảm bảo rằng các thuật toán phân cụm HAC và GMM vận hành trên cùng một thang đo, tránh hiện tượng bị lệch do sự chênh lệch về đơn vị hoặc độ lớn tuyệt đối giữa các đặc trưng. Sau bước chuẩn hóa, tập dữ liệu đã sẵn sàng để đưa vào giai đoạn phân cụm.

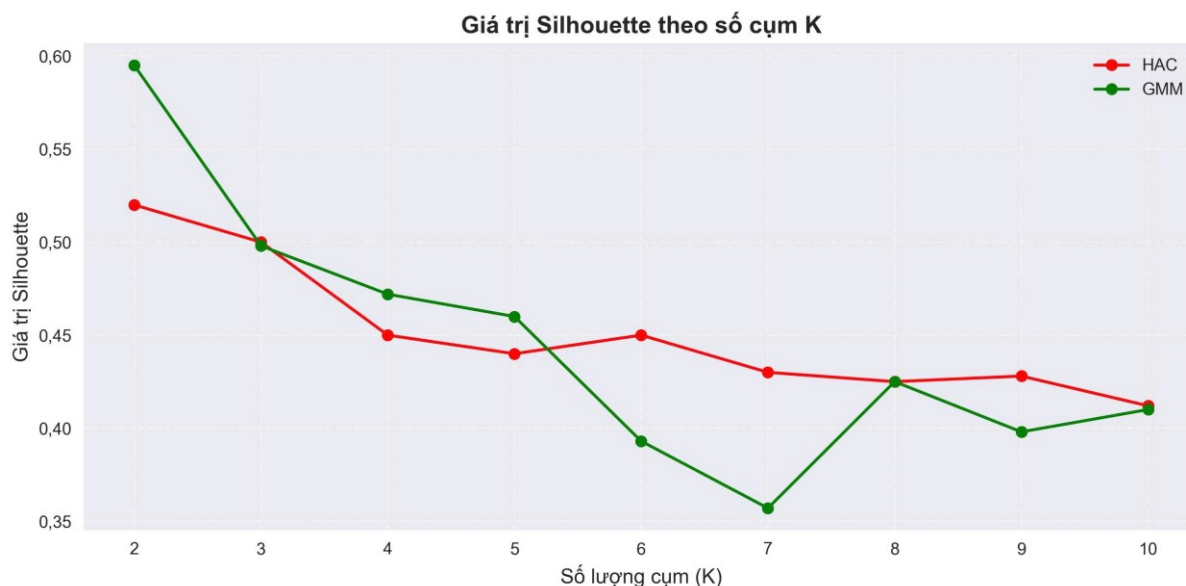
3.3. Phân cụm dữ liệu

Hai thuật toán phân cụm HAC và GMM được triển khai trên bộ dữ liệu GNSS đã qua xử lý và chuẩn hóa. Do chuỗi thời gian có số lượng điểm dữ liệu rất lớn, để giảm tải tính toán nhóm nghiên cứu đã lấy mẫu lại (down-sampling) dữ liệu với bước thời gian thưa hơn trong giai đoạn phân cụm (mục tiêu nhằm đến các xu hướng dài hạn hơn là dao động từng giây). Cụ thể, chuỗi dữ liệu sau tiền xử lý được lấy mẫu lại mỗi 10 giây một điểm, rút gọn còn khoảng 110.000 mẫu, giúp việc tính toán phân cụm khả thi trong thời gian cho phép.

Việc lấy mẫu lại dữ liệu mỗi 10 giây giúp tập trung vào các xu hướng chuyển vị dài hạn và trung hạn do thay đổi nhiệt độ hoặc tải trọng giao thông kéo dài, đồng thời giảm đáng kể chi phí tính toán. Các dao động tần số cao tức thời không phải là mục tiêu chính của phân tích này.

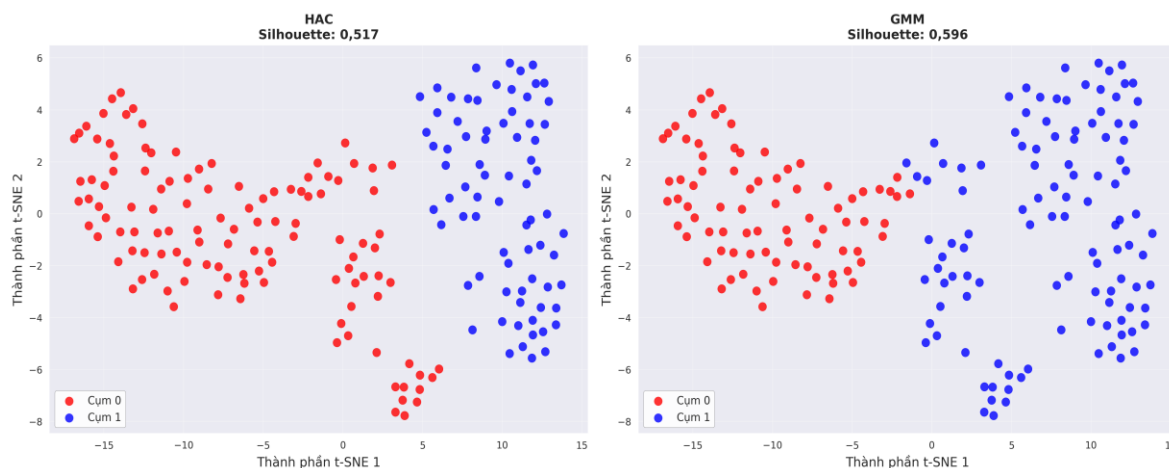
Trên tập dữ liệu rút gọn, khoảng cách Euclid được sử dụng làm thước đo độ tương đồng giữa các điểm. Do dữ liệu là chuỗi một chiều (vị trí theo trục Z), khoảng cách Euclid tương ứng với mức độ chênh lệch giá trị chuyển vị giữa các điểm đo.

Thuật toán HAC thực hiện phân cụm dựa trên cây phân cấp (dendrogram), trong đó dữ liệu được gom dần từ các điểm riêng lẻ đến các cụm lớn hơn, và có thể "cắt" cây tại mức tương ứng với số cụm k . Trong khi đó, GMM dựa trên mô hình hỗn hợp Gaussian và sử dụng thuật toán kỳ vọng–tối đa (EM) để ước lượng các tham số và gán nhãn cụm cho từng điểm dựa trên xác suất.



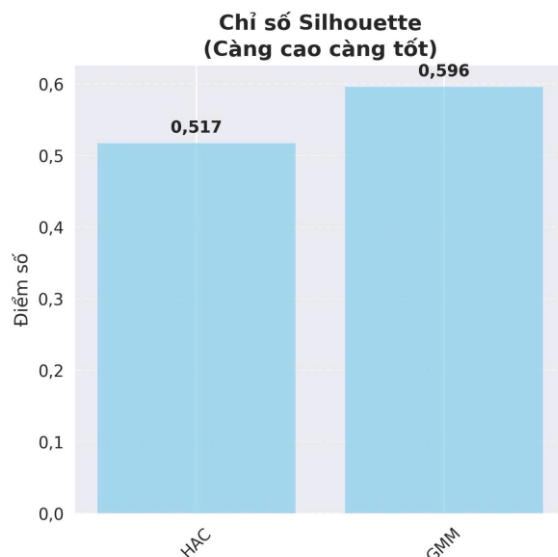
Hình 9. Biểu đồ Silhouette Score theo số cụm k giữa hai thuật toán HAC và GMM.

Hình 9 trình bày biểu đồ chỉ số Silhouette theo số lượng cụm k từ 2 đến 10 đối với cả hai thuật toán HAC và GMM. Kết quả cho thấy khi $k=2$, cả hai thuật toán đều đạt giá trị Silhouette cao (trên 0,50), cho thấy đây là số cụm tối ưu với chất lượng phân cụm tốt nhất. Điều này phản ánh rằng cả HAC và GMM đều phân cụm hiệu quả khi chia dữ liệu thành hai nhóm chính.



Hình 10. Hình ảnh trực quan hóa kết quả phân cụm của HAC và GMM bằng t-SNE với $k=2$.

Hình 10 minh họa kết quả phân cụm của hai thuật toán HAC và GMM với số cụm $k=2$, được trực quan hóa bằng kỹ thuật t-SNE nhằm hỗ trợ quan sát trực tiếp mức độ tách biệt giữa các cụm trong không gian hai chiều. Cả hai thuật toán đều thể hiện cấu trúc phân cụm rõ ràng, phù hợp với kết quả chỉ số Silhouette cao ở mức $k=2$. Trong đó, GMM đạt khoảng 0,596 và HAC đạt khoảng 0,52. Điều này cho thấy cả hai thuật toán đều phân cụm hiệu quả nhất khi chia dữ liệu thành hai nhóm, phản ánh khả năng tồn tại hai trạng thái đặc trưng trong chuỗi chuyển vị GNSS. Sau mốc $k=2$, chỉ số Silhouette của cả hai mô hình đều giảm dần, cho thấy việc tăng số cụm không cải thiện chất lượng phân cụm mà còn làm mờ ranh giới giữa các nhóm dữ liệu. Nhìn chung, GMM duy trì mức Silhouette cao hơn HAC ở hầu hết các giá trị k , chứng tỏ khả năng tách cụm của GMM rõ ràng và ổn định hơn, được thể hiện rõ hơn trong Hình 11.



Hình 11. So sánh chỉ số Silhouette ($k = 2$) giữa HAC và GMM.

4. KẾT LUẬN

- Nghiên cứu cho thấy hai thuật toán phân cụm không giám sát HAC và GMM đều có thể phân tách hiệu quả dữ liệu GNSS-RTK theo phương Z tại cầu Nhật Tân. GMM cho kết quả cụm rõ hơn nhờ mô hình hóa xác suất linh hoạt; HAC đơn giản hơn nhưng vẫn đáng tin cậy sau xử lý nhiễu.

- Cả hai phương pháp đều xác định số cụm tối ưu là $k = 2$, tương ứng với hai trạng thái vận hành của cầu: bình thường (dao động nhỏ) và bất thường (dao động lớn). Kết quả này phản ánh đúng thực tế vận hành của kết cấu cầu dây văng.

- HAC phù hợp khi chưa biết trước số cụm và cần trực quan hóa cấu trúc dữ liệu. GMM thích hợp khi cần đánh giá xác suất hoặc giả định phân phối dữ liệu. Việc kết hợp cả hai có thể tận dụng ưu điểm từng phương pháp.

- Kết quả khẳng định tiềm năng ứng dụng phân cụm không giám sát trong hệ thống SHM, đặc biệt trong việc tự động phát hiện và cảnh báo sớm trạng thái bất thường của công trình cầu.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông Vận tải (ĐH GTVT) trong đề tài mã số T2025-CT-009. Tác giả xin chân thành cảm ơn nhà trường đã tài trợ kinh phí cho nghiên cứu này.

TÀI LIỆU THAM KHẢO

- [1]. A. Masiero, A. Guarnieri, V. Baiocchi, D. Visintini, F. Pirotti, Machine Learning Clustering Techniques to Support Structural Monitoring of the Valgadana Bridge Viaduct (Italy), Remote Sensing, 16 (2024) 3971. <https://doi.org/10.3390/rs16213971>.
- [2]. Lê Văn Hiến, Lê Minh Ngọc, Trần Đức Công, Nghiên cứu phương pháp tiền xử lý dữ liệu quan trắc liên tục gnss của cầu dây văng nhiều trụ thấp, Tạp Chí Khoa Học Giao Thông Vận Tải, 75 (2024) 2345–2355. <https://doi.org/10.47869/tcsj.75.9.9>.
- [3]. A. Guo, A. Jiang, J. Lin, X. Li, Data mining algorithms for bridge health monitoring: Kohonen clustering and LSTM prediction approaches, Journal of Supercomputing, 76 (2020) 932–947. <https://doi.org/10.1007/s11227-019-03045-8>.
- [4]. A. Diez, N.L.D. Khoa, M.M. Alamdari, Y. Wang, F. Chen, P. Runcie, A clustering approach for structural health monitoring on bridges, Journal of Civil Structural Health Monitoring, 6 (2016) 429–

445. <https://doi.org/10.1007/s13349-016-0160-0>.
- [5]. S. Aghabozorgi, A.S. Shirkhorshidi, T.Y. Wah, Time-series clustering – A decade review, Information Systems, 53 (2015) 16–38. <https://doi.org/10.1016/j.is.2015.04.007>.
- [6]. Geo Matching, Which is Better Among Static Survey, RTK or PPK? <https://geo-matching.com/articles/which-is-better-among-static-survey-rtk-or-ppk>, 2022 (truy cập ngày 5 tháng 8 năm 2025).
- [7]. P.N. Tan, M. Steinbach, A. Karpatne, V. Kumar, Introduction to data mining, Second edition, Pearson, New York, 2019.
- [8]. A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31 (2010) 651-666. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- [9]. J.O. Palacio-Niño, F. Berzal, Evaluation Metrics for Unsupervised Learning Algorithms, (2019). <https://doi.org/10.48550/arXiv.1905.05667>.
- [10]. D. Müllner, Modern hierarchical, agglomerative clustering algorithms, 2011. <https://doi.org/10.48550/ARXIV.1109.2378>.
- [11]. L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009.
- [12]. D. Reynolds, Gaussian Mixture Models, in Encyclopedia of biometrics, Springer, Boston, MA, 2015.
- [13] Lê Khánh Giang, Hồ Thị Lan Hương, Đỗ Văn Mạnh, Trần Quang Học, Applying a two-step cluster algorithm in traffic accident data analysis, Transport and Communications Science Journal, 75 (2024) 1673-1687. <https://doi.org/10.47869/tcsj.75.4.16>
- [14] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics, 20 (1987) 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- [15] Cầu Nhật Tân, https://vi.wikipedia.org/w/index.php?title=C%E1%BA%A7u_Nh%E1%BA%ADt_T%C3%A2n&oldid=73707273, 2015 (truy cập ngày 4 tháng 8 năm 2025).
- [16] Bộ GTVT, Tài liệu thiết kế và lắp đặt hệ thống quan trắc cầu Nhật Tân.