ĐẠI HỌC QUỐC GIA TP.HCM
TẠP CHÍ PHÁT TRIỂN KHOA HỌC VÀ CÔNG NGHỆ

Số: 115 /GXN-TCPTKHCN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

Thành phố Hồ Chí Minh, ngày 28 *tháng* 11 *năm* 2022

GIẤY XÁC NHẬN

Tổng Biên tập Tạp chí Phát triển Khoa học và Công nghệ, Đại học Quốc gia TP.HCM xác nhận bản thảo có tên:

Nhận dạng các vỉa than và môi trường trầm tích khu mỏ Núi Béo, Quảng Ninh bằng thuật toán K-means và phương pháp hồi quy

của tác giả/nhóm tác giả:

Khương Thế Hùng, Nguyễn Danh Tuyên

đã được chấp nhận đăng trên Tạp chí:

Tạp chí Phát triển Khoa học và Công nghệ - Khoa học Trái đất và Môi trường (ISSN 2588-1078)

Ngày nhận bài: 8-Feb-2022

Ngày chấp nhận đăng bởi Tạp chí thành viên: 07-Oct-2022

Ngày chấp nhận đăng bởi Tạp chí PTKHCN: 28-Nov-2022

Ngày dự kiến xuất bản:

TẠP CHÍ PHÁT TRIỂN KHOA HỌC VÀ CÔNG NGHỆ
TỔNG BIÊN TẬP

TẠP CHÍ PHÁT TRIỂN KHOA HỌC VÀ CÔNG NGHỆ

Phạm Văn Phúc

VNU-HCM Press
Academic Publisher

Bài nghiên cứu

Nhận dạng các vỉa than và môi trường trầm tích khu mỏ Núi Béo, Quảng Ninh bằng thuật toán K-means và phương pháp hồi quy

Khương Thế Hùng[1,*], Nguyễn Danh Tuyên[2]

[1]Khoa khoa hoc và kỹ thuật Địa chất, Trường Đại học Mỏ - Địa chất, Hà Nội, Việt Nam
[2]Công ty Cổ phần Địa chất Việt Bắc - Vinacomin, Tập đoàn Than và Khoáng sản Việt Nam, Hà Nội, Việt Nam

[*]Tác giả liên hệ: Khương Thế Hùng
Khoa khoa hoc và kỹ thuật Địa chất, Trường Đại học Mỏ - Địa chất, Hà Nội, Việt Nam
Email: khuongthehung@humg.edu.vn

Lịch sử: Ngày nhận: 08-02-2022; Ngày chấp nhận: 29-11-2022; Ngày đăng:

TÓM TẮT

Mỏ than Núi Béo nằm về phía nam tỉnh Quảng Ninh, thuộc dải than Hòn Gai-Cẩm Phả, nơi được đánh giá là một trong những khu vực có tiềm năng lớn về than khoáng của nước ta. Trên cơ sở tổng hợp tài liệu, xử lý số liệu về các thông số vỉa và chất lượng than bằng thuật toán K-means và phương pháp hồi quy cho phép phân khu mỏ Núi Béo thành 3 vùng rõ rệt, vùng thứ nhất có chiều dày ≥18,5 m; vùng thứ hai có chiều dày <18,5 m và góc dốc $<32^0$; vùng ba có chiều dày <18,5 m và góc dốc $\geq 32^0$; kết hợp số lớp kẹp lấy ranh giới là 5 lớp kẹp. Kết quả nhân dạng theo phương pháp K-means và hàm hồi quy cho 2 hoặc 3 thông số vỉa tính cho cả khu mỏ có tỷ lệ nhận dạng rất thấp, chỉ có vỉa 4, 12, 14 là có tỷ lệ nhận dạng cao hơn (trên 50%). Sau khi phân khu, việc nhận dạng các vỉa tăng lên nhiều, đặc biệt các vỉa 5, 7, 11, 12, và 14, trung bình đạt 30,38% so với nhận dạng cho toàn bộ mỏ là 16,83%. Kết quả nhận dạng các vỉa than theo 6 thông số phân tích cho tỷ lệ nhận dạng trung bình đạt 16,64%, theo 9 thông số tỷ lệ nhận dạng đạt 25,93%. Nhìn chung, sau khi phân khu tỷ lệ nhận dạng các vỉa than theo các thông số vỉa và chất lượng than của từng khu vực đã tăng lên đáng kể. Kết hợp các đặc tính về chiều dày, góc dốc, số lớp kẹp và chất lượng than, cho phép phân chia mỏ than Núi Béo thành 07 khu đồng nhất tương đối, trong đó khu A1 mang đặc tính môi trường đầm lầy; khu A2 và A3 mang đặc tính môi trường bãi triều; khu B mang đặc tích môi trường dòng chảy; khu C mang đặc tính bãi triều; khu D mang đặc tính môi trường đầm lầy; và khu E thuộc phần ngoại vi. Kết quả nghiên cứu khẳng định vai trò của các phương pháp toán ứng dụng trong nghiên cứu địa chất, đặc biệt trong việc nhận dạng các vỉa than và môi trường trầm tích hình thành chúng, góp phần phục vụ đắc lực cho việc liên kết, đồng danh các vỉa than được chính xác và phù hợp với môi trường, cũng như cấu trúc khu mỏ.

Từ khóa: Vỉa than, môi trường trầm tích, mỏ Núi Béo, Quảng Ninh

MỞ ĐẦU

Các lớp trầm tích cũng như vỉa than phát triển tương đối liên tục trong quá trình thành tạo cho nên chúng có mối liên hệ không gian gần gũi nhau. Các lớp đá hoặc vỉa than gần gũi nhau về không gian sẽ có chiều dày, góc dốc, số lớp kẹp cũng như đặc điểm, tính chất tương tự nhau, chính các yếu tố này là cơ sở cho việc đồng danh, liên kết vỉa, tập vỉa than[1-3]. Sau khi thành tạo, quá trình hoạt động kiến tạo về sau sẽ làm thay đổi thế nằm, tạo nếp uốn, dịch chuyển các vỉa than gây ra sự gián đoạn, phức tạp hóa trong quá trình nối vỉa. Chính vì vậy, yêu cầu cần phân chia vùng nghiên cứu thành các khu vực có tính đồng nhất tương đối để thực hiện công tác nối, liên kết vỉa than, lớp trầm tích.

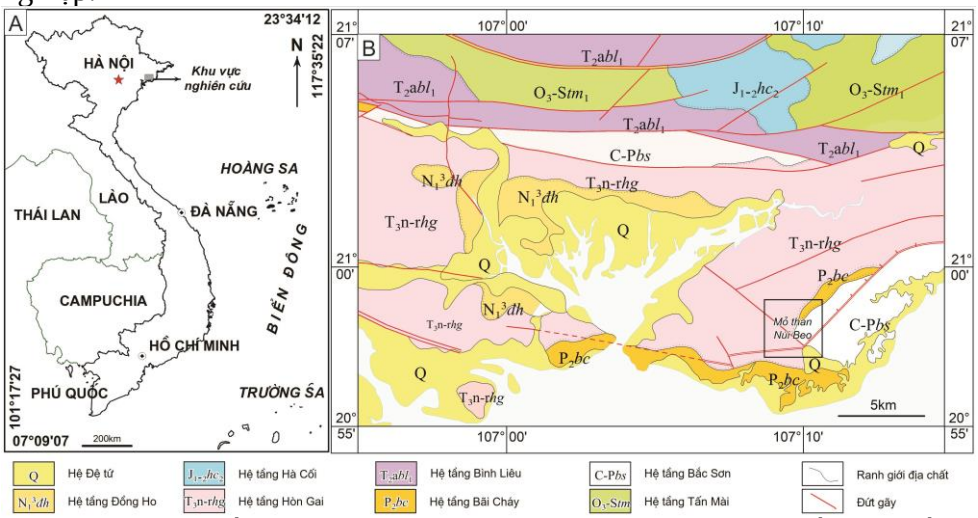
Hầu hết các nghiên cứu trước đây về đồng danh, liên kết vỉa than ít chú ý đến môi trường thành tạo trầm tích. Tuy nhiên, trong cùng thời kỳ thành tạo, trong cùng một bồn trũng, than có thể hình thành trong môi trường đầm lầy (than dày, thế nằm thoải, dạng thấu kính, dạng ổ, chất lượng than tốt), môi trường bãi triều bùn (vỉa than có chiều dày vừa phải, thế nằm thoải, có dạng vỉa, chất lượng than tốt), môi trường bãi bồi ven sông, ven hồ (vỉa than có chiều dày thay đổi, thế nằm hơi dốc, có dạng hình nêm vát, dạng vành khăn, chất lượng than trung bình và kém) và nhiều môi trường khác[1-6]. Như vậy, trong cùng một thời kỳ, cùng một vỉa than nhưng chiều dày vỉa lại thay đổi, hình dạng thay đổi, và thậm chí chất lượng than cũng thay đổi, điều này sẽ dẫn đến phá vỡ tính đồng nhất tương đối gây ra sự phức tạp trong quá trình liên kết, nối vỉa. Các vỉa than ở ven sông, ven hồ của cùng một vỉa nhưng không thể liên kết với nhau vì chúng bị ngăn cách bởi lòng sông cũng như khu vực ngập nước sâu. Tuy nhiên, chu kỳ trầm tích lại phản ánh chúng khá giống nhau về trật tự biến tiến, biển thoái, đây vừa là yếu tố kiến tạo, vừa là môi trường thành tạo trầm tích.

Mục đích của nghiên cứu này là nhận dạng các vỉa than và môi trường trầm tích tại khu vực mỏ Núi Béo, Quảng Ninh bằng thuật toán K-means và phương pháp hồi quy trên cơ sở phân tích các thông số chiều dày, góc dốc và chất lượng vỉa than nhằm phục vụ công tác liên kết vỉa, tính toán tài nguyên, trữ lượng than, định hướng cho công tác thiết kế khai thác được chính xác và phù hợp với điều kiện khu mỏ.

KHÁI QUÁT VỀ ĐẶC ĐIỂM ĐỊA CHẤT MỎ

Mỏ Núi Béo nằm trong dải than Hòn Gai-Cẩm Phả, vị trí khu mỏ thuộc địa phận của 03 phường: phường Hà Phong, phường Hà Tu và phường Hà Trung, thành phố Hạ Long, tỉnh Quảng Ninh. Khu mỏ nằm ở trung tâm thành phố Hạ Long, nằm bên trái đường quốc lộ 18A từ Hạ Long đi Mông Dương (Hình 1A).

Trong khu mỏ có mặt các trầm tích Trias thuộc hệ tầng Hòn Gai, phân hệ tầng giữa và các trầm tích bở rời hệ Đệ tứ[7] (Hình 1B). Thành phần thạch học của phân hệ tầng Hòn Gai giữa bao gồm các lớp cuội kết, cát kết, bột kết, sét kết, sét than và các vỉa than nằm xen kẽ nhau, chiều dày địa tầng khoảng 1.800 m. Phân hệ tầng Hòn Gai giữa là đối tượng chứa các vỉa than công nghiệp.



Hình 1: A-Bản đồ Việt Nam và vị trí vùng Hạ Long, B-Sơ đồ địa chất vùng Hạ Long, Quảng Ninh và vị trí khu mỏ Núi Béo (theo Hùng và nnk, 1996).

Trầm tích hệ Đệ tứ (Q) phủ trực tiếp lên các thành tạo của phân hệ tầng Hòn Gai giữa, chúng được phân bố ở các khu vực thấp, thung lũng xung quanh khu mỏ. Thành phần trầm tích bao gồm cuội, sỏi, cát, sét bở rời, đôi nơi là các tảng lăn, đây là sản phẩm phong hoá từ các đá có trước.

Trong khu mỏ Núi Béo phát triển các nếp uốn và hệ thống các đứt gãy, chúng làm phức tạp và gây khó khăn cho công tác đồng danh vỉa và khai thác than. Theo Anh PT (2009)[8], từ trên bề mặt

địa hình trở xuống, trong phạm vi khu mỏ Núi Béo tồn tại các vỉa than sau V14, V13, V11, V10, V9, V7, V6, V5, và V4.

CƠ SỞ TÀI LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Trên cơ sở thu thập các tài liệu tìm kiếm, thăm dò và khai thác đã tiến hành tại khu vực nghiên cứu, đặc biệt là các tài liệu thăm dò, khai thác triển khai trên khu vực Hà Lầm-Núi Béo và tài liệu mới bổ sung của 11 lỗ khoan thực hiện trong năm 2021 thuộc Dự án thăm dò bổ sung than mỏ Núi Béo do Công ty Cổ phần Địa chất Việt Bắc thực hiện, các số liệu phân tích tại Trung tâm Phân tích Thí nghiệm Địa chất, thuộc Tổng cục Địa chất và Khoáng sản Việt Nam. Những số liệu có được qua các tài liệu là số liệu khoan, phân tích được xử lý, tổng hợp cho từng vỉa, sử dụng để phục vụ công tác nghiên cứu.

Xây dựng cơ sở dữ liệu

Trên cơ sở dữ liệu gốc về các lỗ khoan cắt qua các vỉa than mỏ Núi Béo và kết quả phân tích mẫu, tiến hành đồng bộ hóa dữ liệu.

Đồng bộ dữ liệu theo hệ tọa độ: Toàn bộ dữ liệu tọa độ công trình khai đào được lấy theo hệ tọa độ VN2000 múi 6 độ, kinh tuyến trục $107^0 45'$. Dữ liệu phân tích được đối sánh với dữ liệu khoan để lấy tọa độ (x, y) tương ứng. Kết quả đồng bộ tọa độ được ghi và lưu vào file *_ToadoPT.xls.

Loại trừ các lỗ khoan trùng lặp: Có rất nhiều lỗ khoan trùng lặp khi để các tuyến dọc cùng các tuyến ngang. Điều đó sẽ gây ra sự trùng chập và khí tính chiều dày vỉa và sẽ gây khó khăn. Chính vì vậy, cần phải loại trừ các lỗ khoan và vỉa trùng lặp trước khi đưa số liệu vào tính toán.

Hiệu chỉnh thông số khoan: chiều dày các vỉa được tính tổng cho các vỉa riêng lẻ; góc dốc, số lớp kẹp được tính trung bình của các giá trị. Kết quả hiệu chỉnh thông số khoan được ghi vào file *_DLkhoan.xls.

Hiệu chỉnh kết quả phân tích: Các kết quả phân tích theo mẫu, khi tổng hợp theo vỉa ta tính trung bình gia quyền theo chiều dày. Kết quả phân tích được ghi vào file *_DLPhantich.xls.

Đồng bộ hóa dữ liệu khoan và phân tích: Các lỗ khoan, vỉa có các dữ liệu chiều dày, góc dốc, số lớp kẹp sẽ đồng bộ hóa. Kết quả khoan, phân tích được ghi vào file *_DulieuKPT.xls.

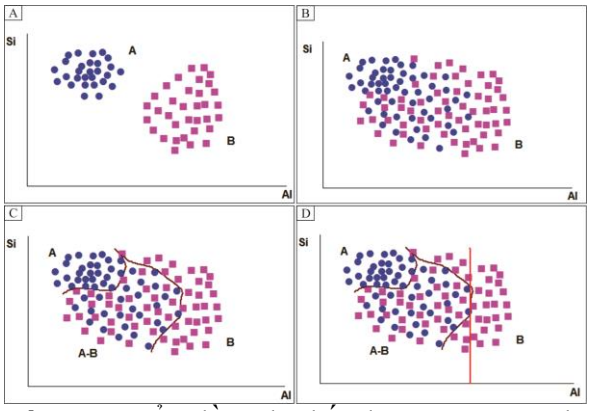
Hiệu chỉnh dữ liệu để phân loại: Hiệu chỉnh dữ liệu khi tính khoảng cách Euclide, giá trị được quy về từ 0 tới 1. Hiệu chỉnh dữ liệu khi phân tích nhận dạng, giá trị được quy về 0 (sai) đến 1 (đúng).

Phương pháp nghiên cứu

Đánh giá đặc trưng phân loại của các thông số

Khi phân loại các thông số người ta thường áp dụng 2 nhóm phương pháp, phương pháp phân loại theo đặc điểm, tính chất đặc trưng của đối tượng và phương pháp phân loại theo quá trình phát triển của đối tượng.

Phần lớn các phương pháp hiện nay đều phân loại theo đặc điểm, tính chất đặc trưng của đối tượng. Để phân loại đối tượng có thể áp dụng phương pháp K-means, phân tích tính đồng nhất, phân tích biệt thức, hồi quy tuyến tính, hồi quy logistic, mạng trí tuệ nhân tạo (ANN),... Việc phân loại theo quá trình phát triển của đối tượng thường rất hạn chế và quá trình phân loại rất phức tạp nên ít được áp dụng. Khi đối tượng là các lớp đá trầm tích, các vỉa than tuân theo quy luật trầm tích, môi trường trầm tích nên đặc điểm, tính chất của vỉa than luôn bị biến động, điều đó có nghĩa là ta không thể lấy đặc điểm, tính chất của vỉa than để phân loại mà cần phân ra các môi trường trầm tích khác nhau.



Hình 2: Biểu đồ tính chất đặc trưng của nhóm A và B: dễ phân loại (A), nhóm A và B bị xóa nhòa (B), khả năng phân loại theo đặc trưng của nhóm (C), phân loại của thông số Al và Si (D).

Nhìn vào biểu đồ (Hình 2A) cho thấy, dễ dàng phân loại nhóm A và B theo thông số Si và Al. Tuy nhiên, thực tế thường phức tạp hơn rất nhiều, việc phân tách giữa nhóm A và B không phải dễ dàng (Hình 2B). Như vậy, căn cứ vào 2 thông số Si và A1 ta không thể phân loại được 2 nhóm A và B (Hình 2C). Do đó, cần bổ sung thêm các thông số nhận dạng khác hoặc chấp nhận một sai số nhất định cho kết quả phân loại.

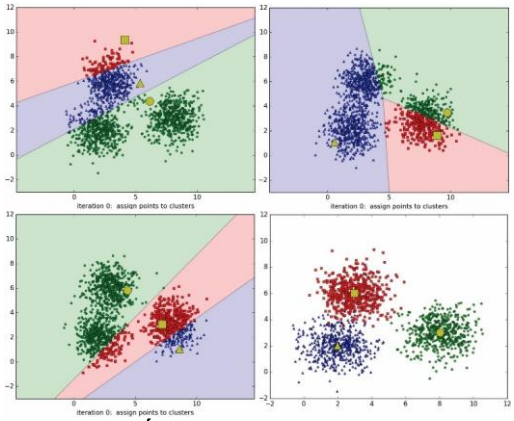
Với 2 thông số Si và A1 ta chỉ có thể phân loại đối tượng A và B đúng một phần nào (Hình 2D). Gọi nA là tổng số điểm nhóm A, mA là số điểm phương pháp nhận dạng được. Tỷ lệ thành công nhận dạng đạt: 100×mA/nA (%).

Nếu chỉ căn cứ thông số Al thì ta có thể phân loại B được còn căn cứ vào thông số Si thì không phân biệt được A và B. Điều đó nói lên mỗi thông số có giá trị đặc trưng phân loại khác nhau. Đối với thông số Al: giá trị đặc trưng phân loại nhóm B đạt 100×21/60=35%, còn đối với thông số Si gần đạt 0%.

Phương pháp phân nhóm theo K-means

Thuật toán K-means clustering (phân cụm K-means) là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm (cluster). Tư tưởng chính của thuật toán K-means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất mà đặc trưng là các giá trị gần với giá trị trung bình (mean) của tâm nhóm đó.

Khi phân cụm K-means thì số cụm kết quả phụ thuộc rất nhiều vào các trung tâm cụm đề ra ban đầu.

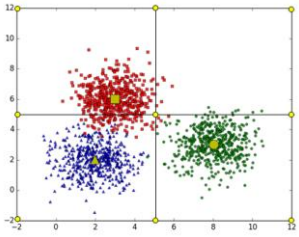


Hình 3: Kết quả của việc xây dựng trung tâm các cụm ban đầu.

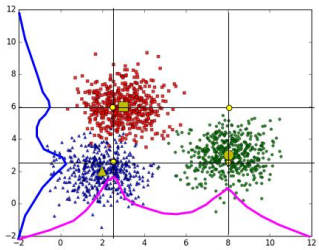
Điều đó cho thấy vị trí các cụm ban đầu đưa ra khác nhau sẽ cho ta kết quả khác nhau. Để tránh tình trạng này ta cần đưa ra tiêu chuẩn xây dựng trung tâm các cụm ban đầu chuẩn.

Có 3 phương pháp tạo trung tâm cụm, đó là: phương pháp chia đều - Chia các yếu tố thành phần các khoảng cách đều nhau; Phương pháp mật độ - Chia các yếu tố thành phần theo mức độ

tập trung các điểm theo mật độ điểm; Phương pháp tự chọn như theo vỉa, theo cấu trúc, theo môi trường,… Số nhóm, các điểm trung tâm tự các nhà nghiên cứu đề xuất theo mục đích.



Phương pháp chia đều



Phương pháp mật độ

Hình 4: Phương pháp tạo trung tâm cụm.

Phương pháp mật độ sẽ tiến gần tới kết quả hơn nhưng thêm các phép tính phức tạp và đôi khi tạo ra các trung tâm giả (Hình 4).

Nếu dùng phương pháp chia đều theo số ngưỡng (k) cho một thông số thì với m thông số sẽ có $\eta=k^m$ nhóm. Giả sử có 3 ngưỡng min, mean, max nếu 2 thông số ta có $3^2 = 9$ nhóm, 3 thông số $3^3=27$ nhóm, 4 thông số $3^4=91$ nhóm,….

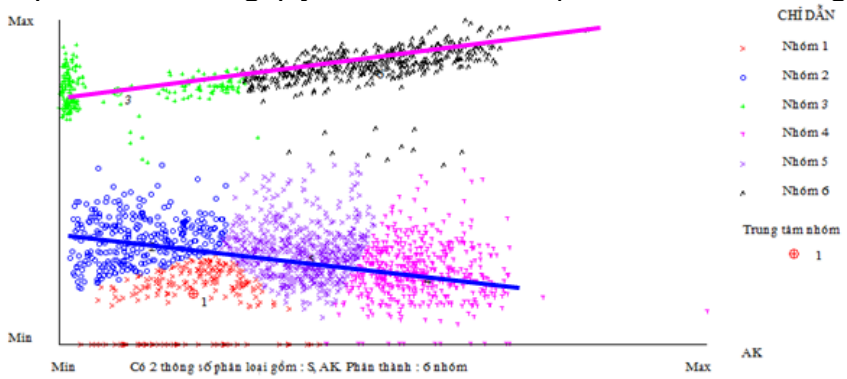
Nếu dùng phương pháp mật độ thì số nhóm là $\eta = \prod_{j=1}^{m} k(j)$

Trong đó: η là số nhóm; $k(j)$ là số ngưỡng của thông số j; m là số thông số.

Giả sử có 2 thông số, thông số 1 chọn 2 ngưỡng, thông số 2 chọn 3 ngưỡng thì số nhóm theo phương pháp mật độ là $2\times3 = 6$ nhóm. Nếu luôn chọn 2 ngưỡng thì số nhóm là $\eta=2m$. Giá trị ngưỡng theo phương pháp mật độ sẽ khác với chia đều mặc dù mỗi thông số có 2 ngưỡng.

Với phương pháp chia đều, 2 ngưỡng ta thường chọn 0,25 và 0,75 trên đoạn [0,1]; 3 ngưỡng ta thường chọn 0, 0,5 và 1 trên đoạn [0,1]; 4 ngưỡng ta thường chọn 0, 1/3, 2/3 và 1 trên đoạn [0,1];…

Thông thường người ta chỉ nên dùng tối đa 3 ngưỡng vì càng nhiều ngưỡng số nhóm sẽ tăng lên quá nhiều gây ra sự quá phức tạp và khó tìm hiểu quy luật. Tuy nhiên, trong nhiều trường hợp các điểm không quy tụ thành điểm mà phát triển thành đường (Hình 5).



Hình 5: Biểu đồ phân nhóm K-means với các điểm tuân theo quy luật đường thẳng tuyến tính.

Như vậy, việc áp dụng K-means sẽ không có tác dụng nếu không xem xét đến yếu tố đường (Line), để bù đắp những thiếu sót này và mong muốn phát triển phương pháp K-means chúng tôi đã tiến hành xây dựng thuật toán L-mean clustering (Line means) hay gọi là phân cụm theo đường tuyến tính L-means. Kết hợp sự cải tiến này với K-means cho ta thuật ngữ chung là KL-means (phân cụm theo điểm, đường).

Tính đồng nhất của các nhóm

Khi xây dựng các nhóm, yêu cầu phải tính mức độ đồng nhất của các nhóm. Mặc dù theo K-means phân thành 2 nhóm nhưng khi so sánh chúng lại giống nhau thì ta phải gộp làm một. Có rất nhiều tiêu chuẩn đánh giá mức độ đồng nhất của các nhóm tập mẫu 1 và tập mẫu 2 có thể theo các tiêu chuẩn như Rodionov (1981)[9], Rao (1967)[10],…

Khi không xét tới mối liên quan giữa các thông số, tiêu chuẩn của Rodionov (1981) thường được sử dụng để đánh giá mức độ đồng nhất giữa 2 tập mẫu[9].

$$V(T_1,T_2)=\frac{n_1+n_2-1}{n_1\times n_2\times(n_1+n_2)}\sum_{j=1}^{m}\frac{(n_2\sum_{t=1}^{n_1}x_{t_j}-n_1\sum_{t=1}^{n_2}x_{t_j})^2}{\sum_{t=1}^{n_1+n_2}x_{t_j}^2-\frac{1}{n_1+n_2}\left(\sum_{t=1}^{n_1+n_2}x_{t_j}\right)^2} \quad (1)$$

Trong đó: $V(T_1,T_2)$ là giá trị Rodionov tính toán giữa 2 tập mẫu T_1, T_2; m là số thông số; n_1, n_2 là số mẫu tập mẫu T_1, T_2; x_{tj} là giá trị ứng với mẫu thứ t, thông số j.

$V(T_1,T_2) \le \chi^2(0.05,m)$ thì 2 tập mẫu là hoàn toàn giống nhau.

Giá trị $\chi^2(0.05,m)$ là giá trị tra bảng tương ứng với giới hạn dưới $V(L)=\chi^2(0.05,m)$.

Nếu các thông số đều đạt mức độ $t(j)=t(0.05,n-2)$ thì giới hạn trên được coi là hoàn toàn khác nhau được xác định là $V(U)$ theo $t(0.05,n-2)$. Nghĩa là.

$V(U)=m\times t_2(0.05,n-2)$ hoặc $V(U)=m\times 1.962$

Nếu giá trị nằm giữa $V(L)$ và $V(U)$ thì 2 tập mẫu đó có thể giống nhau hoặc có thể khác nhau.

hoặc $V=\frac{n_1\times n_2}{n_1+n_2}\sum_{j=1}^{m}\frac{(\bar{x}_1[j]-\bar{x}_2[j])^2}{S[j]^2}\le\chi^2_{\alpha,m}$

Trong đó $S2[j]$ là phương sai chung của hai tập mẫu

$$S2[j]=\frac{1}{n_1+n_2-1}\left\{\sum_{i=1}^{n_1}x_1^2[i,j]+\sum_{i=1}^{n_2}x_2^2[i,j]-\frac{1}{n_1+n_2}(\sum_{i=1}^{n_1}x_1[i,j]+\sum_{i=1}^{n_2}x_2[i,j])^2\right\} \quad (2)$$

Xét tới mối liên quan giữa các thông số, sử dụng tiêu chuẩn Rao (1967)

$$V=-h\times\ln\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right)\le\chi^2_{\alpha,m} \quad (3)$$

Trong đó $h=n_1+n_2-2-\frac{m}{2}, |\Sigma_1|, |\Sigma_2|$ là định thức của ma trận có các phần tử $\sigma_1[i,j]$ và $\sigma_2[i,j]$.

$$\sigma_1[i,j]=\frac{1}{n_1+n_2-2}\left[A+B-\frac{1}{n_1+n_2}(C+D)(E+F)\right];\ \sigma_2[i,j]=\frac{1}{n_1+n_2-2}\left[A+B-\frac{C.E}{n_1}-\frac{D.F}{n_2}\right] \quad (4)$$

$$A=\sum_{t=1}^{n_1}x[t,i]\times x[t,j];\ B=\sum_{t=1}^{n_2}y[t,i]\times y[t,j];C=\sum_{t=1}^{n_1}x[t,i];D=\sum_{t=1}^{n_2}y[t,i];\ E=\sum_{t=1}^{n_1}x[t,j];\ F=\sum_{t=1}^{n_2}y[t,j] \quad (5)$$

Trong nghiên cứu này, chúng tôi sử dụng tiêu chuẩn của Rodionov (1981), đây là bước cải tiến quan trọng đối với phương pháp K-means, gọi là phương pháp K-means cải tiến hay KL-means (K-means và L-means tương ứng cụm điểm và đường).

Các bước thực hiện thuật toán K-means

Thuật toán K-means cải tiến thực hiện qua các bước chính sau:

1) Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster), mỗi cụm được đại diện bằng các tâm của cụm.

2) Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclide).

3) Nhóm các đối tượng vào nhóm gần nhất.

4) Nếu dữ liệu theo quy luật đường thì chuyển sang L-means.

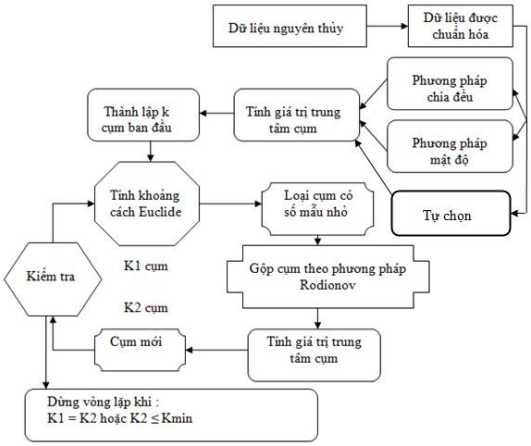
5) Loại trừ các nhóm có ít điểm tham gia, chọn giới hạn 0,1%.

6) Ghép các nhóm theo tiêu chuẩn của Rodionov (1981), gộp các nhóm có chỉ số V thấp.

7) Xác định lại tâm mới cho các nhóm.

8) Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng.

Các bước 1, 2, 3, 5, 7, 8 là dùng cho K-means.



Hình 6: Sơ đồ thuật toán K-means cải tiến (KL-means).

Vòng lặp chỉ dừng lại khi số cụm ổn định K1=K2 (K1 cụm đầu vòng lặp, K2 cụm cuối vòng lặp) hoặc $K2 \leq K_{min}$ (K2 cụm cuối nhỏ hơn K cụm tối thiểu).

Số cụm tối thiểu K_{min} và số cụm tối đa K_{max} tùy theo đối tượng, mục đích, số thông số đầu vào và chương trình. Ta có thể đưa ra các giới hạn này để quy định cho chương trình tiếp tục thực hiện hoặc dừng vòng lặp sao cho kết quả đạt mục đích mong muốn. Ví dụ như ta có 9 vỉa bao gồm các vỉa 4, 5, 6, 7, 9, 10, 11, 13, 14 thì ta cần chọn số cụm tối đa K_{max}=9. Số thông số đầu vào là 3 thì số cụm khởi tạo là $3^3 = 27$. Số cụm tối thiểu thường lấy $K_{min} = 2 \times m\text{-}1 = 2 \times 3\text{-}1 = 5$. Như vậy, yêu cầu chương trình chạy sao cho số cụm cuối cùng có kết quả dừng lân cận 5÷9 cụm. Tuy nhiên, chương trình có thể dừng lại số cụm nhỏ hơn K_{min} khi chưa ổn định. Do vậy, cần điều khiển chương trình ra số nhóm theo mong muốn khi gộp các nhóm có khoảng cách trung tâm nhóm gần nhau nhất.

Đánh giá mức độ phụ thuộc

Mức độ phụ thuộc hay mức độ độc lập nói lên khả năng sử dụng của phương pháp, giả sử có 200 điểm thuộc nhóm A, sau khi phân loại ta nhận được 186 điểm còn 14 điểm bị nhận sang nhóm B thì tỷ số 186/200=0.94 hay 94% nhận diện đúng. Tỷ số này càng nhỏ chứng tỏ các điểm nhóm A phụ thuộc vào B càng lớn và ngược lại tỷ số này càng lớn thì mức độ phụ thuộc của các điểm thuộc nhóm A vào B càng giảm. Khi tỷ số đạt tới 100%, ta khẳng định A độc lập không chịu bất cứ ảnh hưởng nào của B.

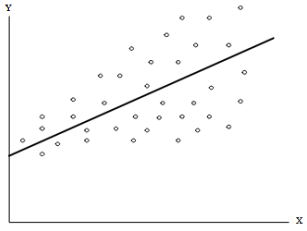
Để đánh giá mức độ phụ thuộc, người ta sử dụng mức xác suất (P):

+ $P < 58\%$ - Hoàn toàn phụ thuộc;
+ $P \geq 58 \div 90\%$ - Phụ thuộc nhiều;
+ $P \geq 90 \div 95\%$ - Phụ thuộc ít;
+ $P \geq 95\%$ - Độc lập.

Phần lớn mức xác suất (P) rơi vào ngưỡng phụ thuộc nhiều (58÷90%) nên ta cần hiệu chỉnh trung tâm nhóm ngay từ ban đầu. Giá trị trung bình độ phụ thuộc của các nhóm cũng góp phần đánh giá việc lựa chọn nhóm sao cho phù hợp nhất.

Phân nhóm theo phương pháp hồi quy tuyến tính (linear regression)

Hồi quy tuyến tính (linear regression) thuộc nhóm supervised learning (Học có giám sát). Hồi quy tuyến tính là một phương pháp rất đơn giản để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập. Hồi quy tuyến tính là một phương pháp để dự đoán biến phụ thuộc (y) dựa trên giá trị của biến độc lập (x).



Hình 7: Hồi quy tuyến tính (linear regression).

Có thể phân ra hồi quy tuyến tính đơn biến (Simple Linear Regression - SLR) và hồi quy tuyến tính đa biến (Multiple Linear Regression - MLR). Đơn biến còn gọi là 2 chiều, đơn giản, đa biến còn gọi là đa chiều hay bội.

Phương trình hồi quy tuyến tính đơn biến: $y = \beta_0 + \beta_1 x + \varepsilon$

Phương trình hồi quy tuyến tính đa biến: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$

Trong đó y là biến phụ thuộc, x là biến độc lập, β là hệ số tương ứng với biến x, ε là sai số hồi quy.

Do hệ số β chưa biết cho nên phải đi xác định, có 2 phương pháp xác định hệ số β cơ bản đó là phương pháp giải hệ phương trình hồi quy tuyến tính (Linear Regression Equation - LRE) và phương pháp luyện phương trình hồi quy tuyến tính (Linear Regression Training - LRT) theo thuật toán học máy (Machine Learning - ML). Cả hai phương pháp này đều áp dụng thuật toán ước lượng bình phương nhỏ nhất (Ordinary Least Squares - OLS).

Đối với các thành tạo trầm tích, các vỉa là tập hợp sự biến đổi có quy luật theo tướng trầm tích nên đặc điểm của vỉa phần lớn có sự biến đổi tuân theo quy luật hồi quy tuyến tính. Một vỉa than có thể rất dày, chất lượng tốt được thành tạo ở môi trường đầm lầy sẽ được nối vỉa với vỉa than ở ven bờ có chiều dày trung bình, chất lượng than kém hơn, thậm chí vỉa than này còn được liên kết với vỉa có chiều dày mỏng, chất lượng kém ở địa hình dốc hơn. Như vậy, khi sử dụng thuật toán K-means để xác định đặc trưng của vỉa có thể dẫn tới sai lầm nên việc ứng dụng kết hợp phương pháp hồi quy tuyến tính sẽ làm giảm thiểu những sai sót đó. Từ các tập hợp dữ liệu ta xây dựng hàm hồi quy tuyến tính đa chiều và phân nhóm theo các đường hồi quy tuyến tính này. Theo phương pháp này ta chọn kiểu L (Line) đường để phân cụm (cluster) tương tự như K-means và gọi tên là phương pháp L-means.

Khoảng cách Euclide tính từ điểm bất kỳ tới đường thẳng $A(0) + \sum_{j=1}^{m} A(j) \times x(j) = 0$ được áp dụng theo công thức sau:

$$d(i) = \frac{|A(0) + \sum_{j=1}^{m} A(j) \times x(i,j)|}{\sqrt{\sum_{j=1}^{m} A(j)^2}} \qquad (6)$$

Trong đó: *d(i)* là khoảng cách Euclide từ điểm i tới đường thẳng; *A(0)* và *A(j)* là hệ số tự do và hệ số theo các thông số j của hàm hồi quy tuyến tính; *x(i,j)* là giá trị điểm i của thông số j.

Các bước tiến hành phân cụm theo đường (L-means).

1) Chọn L đường (Line) cho L cụm (cluster), mỗi cụm được đại diện bằng các hàm hồi quy tuyến tính (Linenear Regression);

2) Tính khoảng cách giữa các đối tượng (objects) đến L đường (thường dùng khoảng cách Euclide);

3) Nhóm các đối tượng vào đường gần nhất;

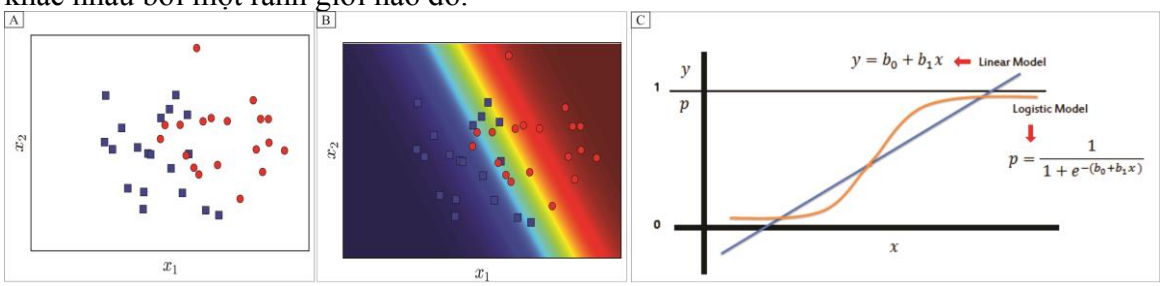
4) Loại trừ các nhóm có ít điểm tham gia, chọn giới hạn 0,1%;

5) Xác định lại đường mới cho các nhóm;

6) Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng.

Phân nhóm theo phương pháp hồi quy logistic (logistic regression)

Hồi quy logistic (logistic regression) là phương pháp phân chia các đối tượng thành các nhóm khác nhau bởi một ranh giới nào đó.



Hình 8: Biểu đồ phân chia các đối tượng theo phương pháp hồi quy logistic: phân bố các nhóm khác nhau (A), xây dựng các đường phân chia các đối tượng (B), phân biệt mô hình hồi quy

tuyến tính và hồi quy logistic (C).

Như vậy, mô hình hồi quy tuyến tính được xác định theo hàm: $y = A(0) + \sum_{j=1}^{m} A(j) \times x(j)$

Mô hình hồi quy logistic được xác định theo hàm: $P = \frac{1}{1+e^{-z}}$

Trong đó: $Z = A(0) + \sum_{j=1}^{m} A(j) \times x(j)$; P là xác suất 0/1 hoặc [0,1]; y - Biến phụ thuộc; $x(j)$ - Biến độc lập của thông số j; m là số thông số.

Các giá trị 0/1 được xác đinh theo nguyên lý lấy vỉa hiện tại so sánh với vỉa chuẩn. Ví dụ ta cần xác định vỉa 6 theo phân loại ban đầu. Nếu vỉa này được xác định là chuẩn của vỉa 6 thì sẽ nhận được giá trị 1. Nếu không xác định là vỉa chuẩn thì ta cần dùng phương pháp phân loại để trả lời. Các vỉa khác không phải vỉa 6 ban đầu và không là vỉa chuẩn thì nhận giá trị 0.

Ngoài ra, chúng ta có thể hiệu chỉnh dựa trên không gian địa tầng gần gũi với các vỉa chuẩn cho các vỉa phân loại một gia số nào đó mà không vượt quá 0,5. Cách xử lý này sẽ giúp làm tăng sự điều chỉnh nhận dạng các vỉa lân cận. Ví dụ, ở mỏ than Núi Béo, vỉa 5, vỉa 7 hiệu chỉnh lấy giá trị 0,5 (so với vỉa 6 là vỉa chuẩn); vỉa 4 lấy 0; vỉa 13 lấy 0; vỉa 14 lấy giá trị 0,…

KẾT QUẢ VÀ THẢO LUẬN

Đánh giá các thông số nhận dạng vỉa than

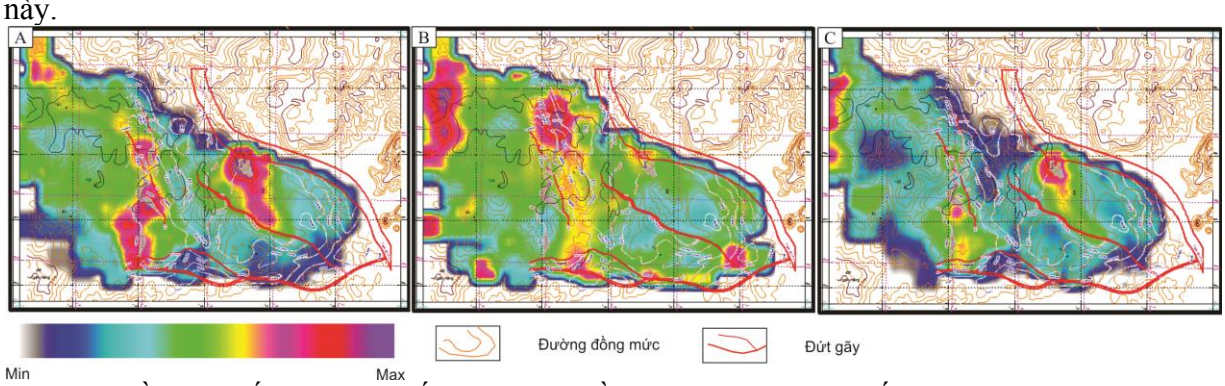
Để phục vụ việc phân chia cấu trúc mỏ than, nhóm nghiên cứu tiến hành đánh giá thống kê cho các yếu tố thành phần cơ bản của vỉa than như chiều dày, góc dốc, lớp kẹp, và các kết quả phân tích W^{pt}, A^{K}, V^{ch}, Q^{ch},…

Trên sơ đồ phân bố các thông số vỉa (hình 9) cho thấy về phía đông khu mỏ Núi Béo các vỉa có chiều dày khá lớn gợi ý đây là vùng trũng dạng đầm lầy là chủ yếu. Kết quả đánh giá độ thông tin 3 thông số là góc dốc, chiều dày, số lớp kẹp như bảng 1.

Bảng 1: Đánh giá tương quan giữa các thông số góc dốc, chiều dày và số lớp kẹp của vỉa than.

Thông số	**Góc dốc**	**chiều dày thật**	**Số lớp kẹp**
Góc dốc	1	-0,051	-0,041
chiều dày thật	-0,051	1	**0,741**
Số lớp kẹp	-0,041	**0,741**	1
Thứ tự	chiều dày thật	Số lớp kẹp	Góc dốc
Tỷ lệ	60,52%	82,84%	100%

Kết quả tính toán cho thấy chiều dày vỉa có độ tin cao nhất chiếm 60,52% lượng thông tin sau đó là số lớp kẹp và cuối cùng là góc dốc. Để đảm bảo nhận dạng 95% trở lên ta dùng cả 3 thông số này.



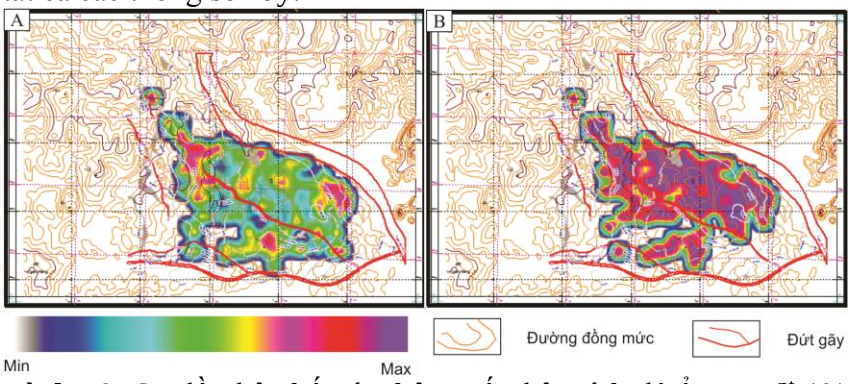
Hình 9: Sơ đồ phân bố các thông số vỉa than, chiều dày thật (A), góc dốc (B), số lớp kẹp (C) khu vực mỏ Núi Béo, Quảng Ninh.

Bảng 2: Đánh giá mối quan hệ tương quan và độ thông tin các thông số phân tích mẫu.

Thông số	**W^{pt}**	**A^{K}**	**V^{ch}**	**Q^{ch}**	**Q^{kh}**	**S**
W^{pt}	1	0,188	-0,460	0,002	-0,150	0,053
A^{K}	0,188	1	0,367	-0,486	-0,972	-0,116
V^{ch}	-0,460	0,3673	1	-0,156	-0,364	-0,071
Q^{ch}	0,002	-0,486	-0,158	1	**0,652**	0,043

Q^{kh}	-0,150	-0,972	-0,364	**0,652**	1	0,108
S	0,053	-0,116	-0,071	0,043	0,108	1
Thứ tự	Q^{ch}	Q^{kh}	W^{pt}	V^{ch}	S	A^{K}
Tỷ lệ	46,54%	65,64%	76,92%	86,28%	94,49%	100%

Từ bảng 2 cho thấy yếu tố nhiệt lượng Q^{ch} và Q^{kh} có mức độ tương quan chặt nhất và nhiệt lượng Q^{ch} chiếm độ thông tin cao nhất, tuy nhiên để phân loại cấu trúc khu mỏ cũng cần sử dụng tất cả các thông số này.



Hình 10: Sơ đồ phân bố các thông số phân tích độ ẩm - W^{pt} (A), nhiệt lượng - Q^{ch} (B) khu mỏ than Núi Béo, Quảng Ninh.

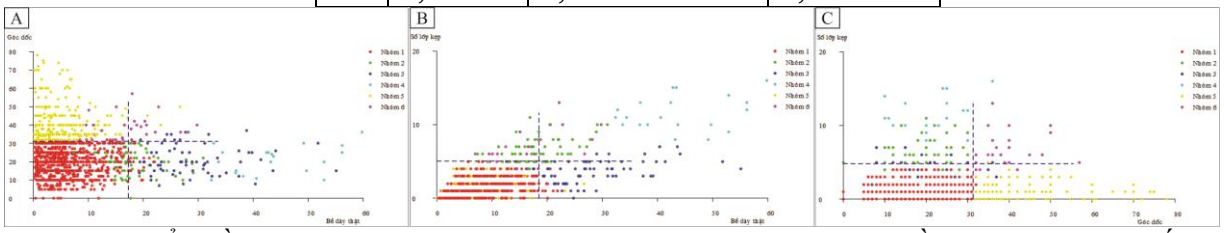
Trên cơ sở phân nhóm theo phương pháp K-means cho 3 đối tượng là Độ dốc - chiều dày - Số lớp kẹp, mỗi nhóm phân thành 3 ngưỡng, thì từ 3^3=27 nhóm, ta thấy chương trình còn 10 nhóm. Mỗi nhóm phân thành 2 ngưỡng, thì từ 2^3 = 8 nhóm ban đầu, chương trình còn 6 nhóm.

Bảng 3: Các trung tâm K-means ban đầu.

Nhóm	Góc dốc	Chiều dày thật	Số lớp kẹp
1	0,25	0,25	0,25
2	0,25	0,25	0,75
3	0,25	0,75	0,25
4	0,25	0,75	0,75
5	0,75	0,25	0,25
6	0,75	0,25	0,75
7	0,75	0,75	0,25
8	0,75	0,75	0,75

Bảng 4: Các trung tâm K-means sau khi ghép nhóm.

TT	Góc dốc	Chiều dày thật	Số lớp kẹp
1	0,256	0,105	0,073
2	0,277	0,329	0,474
3	0,265	0,521	0,280
4	0,269	0,735	0,734
5	0,560	0,088	0,078
6	0,514	0,333	0,466



Hình 11: Biểu đồ phân nhóm K-means khu mỏ than Núi Béo theo chiều dày thật - góc dốc (A), chiều dày thật - số lớp kẹp (B), và góc dốc - số lớp kẹp (C).

+ Kết quả phân nhóm theo K-means (Hình 11A) cho thấy ranh giới rõ nét nhất cho phân chia chiều dày là giữa nhóm 3 so với nhóm 1 và nhóm 5.

Nếu lấy ranh giới giữa nhóm 3 và 1 ta có :

RG=60×(0,5205+0,1045)/2=0,625/2=60×0,3125=18,75

Nếu lấy ranh giới giữa nhóm 3 và 5 ta có :

RG=60×(0,5205+0,088)/2=0,6085/2=60×0,30425=18,255

Ranh giới trung bình là (18,75+18,255)/2=18,5025; do vậy ta lấy ranh giới chiều dày thật làm cơ sở phân loại là 18,5 m.

+ Đối với góc dốc (Hình 11B) ranh giới giữa nhóm 5 và nhóm 1 là rất rõ ràng

RG=80×(0,5599+0,2557)/2=0,8156/2=80×0,4078=32,624

Do vậy, ranh giới góc dốc được chọn là 32^0.

+ Đối với số lớp kẹp (Hình 11C) ranh giới nhóm 2 và nhóm 1 là rõ ràng.

RG=20×(0,0726+0,4741)/2=20×0,5447/2=20×0,27335=5,467

Kết quả cho phép chọn ranh giới số lớp kẹp là 5 lớp kẹp.

Một cách tổng thể, các nhóm có tính phân tán cao nhưng ta có thể chọn ranh giới góc dốc ở 32^0 làm ranh giới phân chia độ dốc, còn chiều dày là 18,5m có khả năng phân chia chiều dày. Kết hợp giữa chiều dày và góc dốc cho phép phân thành 3 vùng rõ rệt, vùng 1 có chiều dày ≥18,5 m; vùng 2 có chiều dày <18,5 m và góc dốc $<32^0$; vùng 3 có chiều dày <18,5 m và góc dốc $\geq 32^0$; kết hợp số lớp kẹp lấy ranh giới là 5 lớp kẹp.

Phân chia các khu vực đồng nhất tương đối và nhận dạng các vỉa than mỏ Núi Béo

Trên cơ sở chiều dày, góc dốc, số lớp kẹp và các thông số phân tích tiến hành công tác phân loại các khu vực đồng nhất tương đối cho các vỉa than mỏ Núi Béo và nhận dạng các vỉa than theo thứ tự cho toàn khu mỏ và sau khi đã phân ra các khu vực đồng nhất tương đối.

Bảng 5: Đánh giá nhận dạng các vỉa theo 3 thông số (góc dốc, chiều dày, lớp kẹp) bằng phương pháp K-means cho toàn khu mỏ.

Tên vỉa	Tổng số	Nhận dạng	Tỷ lệ
14	267	131	***49,064***
13	221	10	4,525
11	321	32	9,969
10	359	37	10,306
9	112	4	3,571
7	159	35	22,013
6	84	1	1,191
5	45	1	2,222
4	21	11	***52,381***
12	15	8	***53,333***
Tổng	**1604**	**270**	**16,830**

Bảng 6: Đánh giá kết quả nhận dạng các vỉa than theo 2 thông số (góc dốc, chiều dày) bằng phương pháp K-means cho toàn khu mỏ.

Tên vỉa	Tổng số	Nhận dạng	Tỷ lệ
14	267	134	***50,187***
13	221	23	10,407
11	321	29	9,034
10	360	34	9,444
9	113	4	3,540
7	160	22	13,750
6	84	1	1,191
5	45	1	2,222
4	21	11	***52,381***
12	15	10	***66,667***
Tổng	**1607**	**269**	**16,740**

Kết quả đánh giá theo bảng 5, 6 cho thấy việc nhận dạng theo phương pháp K-means cho 2 hoặc 3 thông số tính cho khu mỏ có tỷ lệ nhận dạng rất thấp, chỉ có vỉa 4, 12, 14 là có tỷ lệ nhận

dạng cao hơn (trên 50%). Do vậy, không thể giữ nguyên dữ liệu để phân loại cấu trúc cho cả khu mỏ được mà phải phân theo môi trường trầm tích và theo từng phân khu riêng biệt.

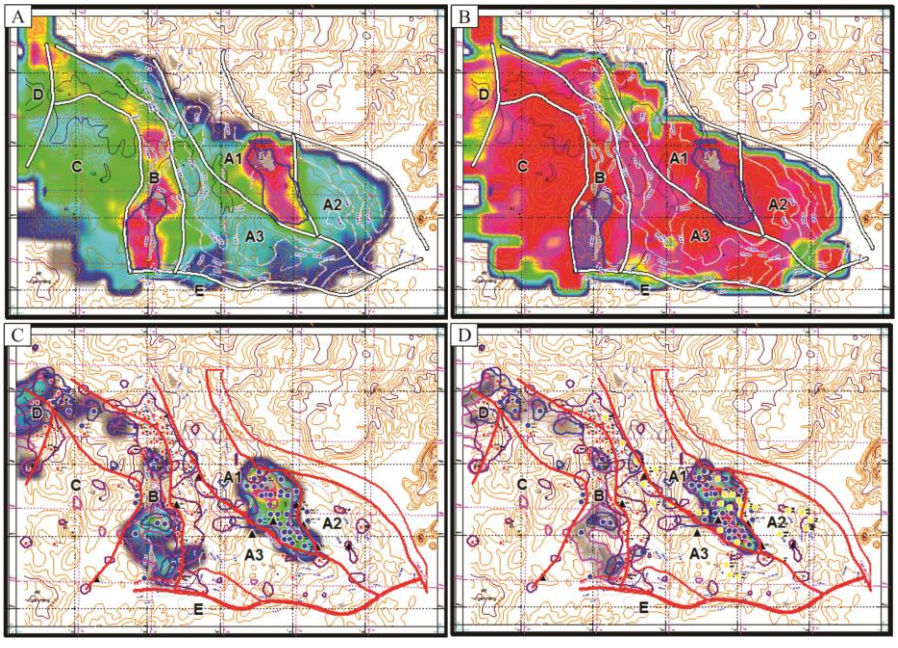
Bảng 7: Đánh giá thông tin theo tương quan các thông số.

Thông số	R	Độ tin cậy	Trọng số
Góc dốc	-0,015	0,150	1
chiều dày thật	***0,178***	***1,778***	***9***
Số lớp kẹp	***0,194***	***1,938***	***10***
W^{pt}	-0,062	0,616	3
A^K	-0,057	0,566	3
V^{ch}	-0,017	0,170	1
Q^{ch}	-0,061	0,609	3
Q^{kh}	-0,021	0,212	1
d	-0,025	0,245	1

Bảng 8: Đánh giá độ tin cao của các thông số.

Thông số	Góc dốc	Chiều dày thật	Số lớp kẹp	W^{pt}	A^K	V^{ch}	Q^{ch}	Q^{kh}	d
Góc dốc	1	0,114	0,140	0,303	-0,090	-0,156	-0,040	-0,003	-0,0004
Chiều dày thật	0,114	1	***0,701***	0,053	-0,176	-0,191	-0,129	-0,005	-0,085
Số lớp kẹp	0,140	***0,701***	1	0,127	-0,052	-0,162	-0,153	-0,091	-0,096
W^{pt}	0,302	0,053	0,127	1	-0,064	-0,263	-0,137	-0,101	-0,135
A^K	-0,090	-0,176	-0,052	-0,064	1	0,388	-0,076	-0,580	0,432
V^{ch}	-0,156	-0,191	-0,162	-0,263	0,388	1	0,389	0,084	***0,504***
Q^{ch}	-0,040	-0,129	-0,153	-0,137	-0,076	0,389	1	***0,839***	0,470
Q^{kh}	-0,003	-0,005	-0,091	-0,101	-0,580	0,084	***0,839***	1	0,181
d	-0,0004	-0,085	-0,096	-0,135	0,432	***0,504***	0,470	0,181	1
Thứ tự	V^{ch}	Q^{ch}	Q^{kh}	W^{pt}	A^K	chiều dày thật	d	Góc dốc	Số lớp kẹp
Tỷ lệ	38,18%	53,21%	64,40%	73,01%	79,82%	85,48%	90,87%	95,63%	100%

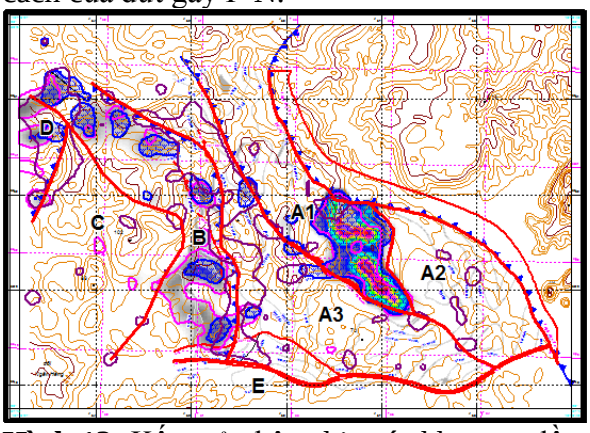
Kết quả đánh giá độ thông tin (bảng 7) và độ tin cao các thông số (bảng 8) cho thấy mối quan hệ với tập vỉa thì chiều dày, số lớp kẹp có độ thông tin và quan hệ tốt, còn các giá trị tin cao được đánh giá cho chất bốc V^{ch} là cao nhất. Trên cơ sở đó chúng tôi tiến hành phân khu dựa trên các tính chất của vỉa than.



Hình 12: Sơ đồ phân khu mỏ than Núi Béo theo chiều dày (A), theo môi trường (B), theo phương pháp hồi quy (C), và theo phương pháp hồi quy có bổ sung thêm yếu tố độ ẩm (W^{pt}) và nhiệt lượng (Q^{ch}).

Trên sơ đồ hình 12 cho thấy khu A1, A2, và A3 có tính chất tương tự nhau, tuy nhiên khu A1 và A2 có mức độ giống nhau hơn. Chính vì vậy, cho phép phân chia mỏ than Núi Béo thành 07 khu sau đây: Khu A1 mang đặc tính môi trường đầm lầy; khu A2, A3 mang đặc tính môi trường bãi triều; khu B mang đặc tích môi trường dòng chảy; khu C mang đặc tính bãi triều; khu D mang đặc tính môi trường đầm lầy; và khu E thuộc phần ngoại vi.

Về ranh giới phân khu: khu A1 và A2 lấy ranh giới bằng đứt gãy Monplane, giữa A1 và A2 lấy ranh giới của các yếu tố phân nhóm. Khu B và khu A ranh giới là đứt gãy F-K và F-C đồng thời là đới có độ dốc lớn. Khu C và khu B được giới hạn bởi sự hình thành biểu hiện dòng chảy qua khu B. Khu D và khu C có ranh giới là đới độ dốc lớn. Khu E được tách biệt bởi sự ngăn cách của đứt gãy F-N.



Hình 13: Kết quả phân chia các khu vực đồng nhất tương đối mỏ than Núi Béo.

Bảng 9: Đánh giá kết quả nhận dạng các vỉa theo 3 thông số (góc dốc, chiều dày và lớp kẹp) khu A1 bằng phương pháp K-means sau khi đã phân khu.

Tên vỉa	Tổng số	Nhận dạng	Tỷ lệ
14	82	53	***64,634***
13	37	0	0
11	47	16	34,043
10	41	2	4,878
9	11	0	0
7	27	6	22,222
6	13	0	0
5	1	1	***100***
12	1	1	***100***
Tổng	**260**	**79**	**30,38**

Khi phân khu, việc nhận dạng các vỉa cao lên nhiều, đặc biệt vỉa 5, 7, 11, 12, 14, trung bình đạt 30,38% so với nhận dạng toàn bộ 16,83%. Phân khu theo môi trường càng chuẩn, càng hẹp thì khả năng nhận dạng chính xác càng cao.

Bảng 10: Đánh giá kết quả nhận dạng các vỉa theo 6 thông số (W^{pt}, A^{k}, V^{ch}, Q^{ch}, Q^{kh}, S) bằng phương pháp K-means sau khi đã phân khu.

Tên vỉa	Tổng số	Nhận dạng	Tỷ lệ
14	23	12	***52,174***
13	21	3	14,286
11	49	11	22,449
10	57	2	3,509
9	24	0	0
7	45	3	6,667

6	14	1	7,143
5	3	1	33,333
4	3	2	***66,667***
Tổng	**239**	**35**	**14,640**

Bảng 11: Đánh giá nhận dạng các vỉa theo 9 thông số (góc dốc, chiều dày thật, lớp kẹp, W^{pt}, A^{k}, V^{ch}, Q^{ch}, Q^{kh}, S) bằng phương pháp K-means sau khi đã phân khu.

Tên vỉa	Tổng số	Nhận dạng	Tỷ lệ
14	21	13	***61,91***
13	18	5	27,78
11	43	5	11,63
10	51	7	13,77
9	21	10	47,62
7	44	11	25,0
6	13	2	15,39
5	3	2	***66,67***
4	2	1	***50,00***
Tổng	**216**	**56**	**25,93**

Qua kết quả đánh giá ở các bảng 9÷11 cho thấy, việc nhận dạng các vỉa than theo 6 thông số phân tích cho tỷ lệ nhận dạng trung bình đạt 16,64% còn theo 9 thông số tỷ lệ nhận dạng đạt 25,93%. Như vậy, sau khi phân khu tỷ lệ nhận dạng các vỉa than theo các thông số vỉa và chất lượng than của từng môi trường đã tăng lên đáng kể.

KẾT LUẬN

Kết quả tổng hợp, phân tích tài liệu, xử lý thống kê số liệu lỗ khoan và kết quả phân tích mẫu bằng thuật toán K-means và phương pháp hồi quy khu mỏ Núi Béo, Quảng Ninh cho phép rút ra một số kết luận như sau:

1) Áp dụng thuật toán K-means và hồi quy cho phép lựa chọn ranh giới góc dốc ở 32^0 làm ranh giới phân chia độ dốc, còn ranh giới phân chia chiều dày là 18,5m. Kết hợp giữa chiều dày và góc dốc cho phép phân thành 3 vùng rõ rệt, vùng 1 có chiều dày ≥18,5 m; vùng 2 có chiều dày <18,5 m và góc dốc $<32^0$; vùng 3 có chiều dày <18,5 m và góc dốc $\geq 32^0$; kết hợp số lớp kẹp lấy ranh giới là 5 lớp kẹp. Kết quả nhân dạng theo phương pháp K-means cho 2 hoặc 3 thông số tính cho khu toàn mỏ có tỷ lệ nhận dạng thấp, chỉ có vỉa 4, 12, 14 là có tỷ lệ nhận dạng cao hơn (trên 50%). Sau khi phân khu, việc nhận dạng các vỉa than được nâng lên nhiều, đặc biệt các vỉa 5, 7, 11, 12, 14, trung bình đạt 30,38% so với nhận dạng vỉa cho toàn bộ khu mỏ là 16,83%. Phân khu theo môi trường càng chuẩn, càng hẹp thì khả năng nhận dạng chính xác càng cao. Kết quả nhận dạng các vỉa than ở từng khu riêng lẻ theo 6 thông số phân tích cho tỷ lệ nhận dạng trung bình đạt 16,64%, còn theo 9 thông số tỷ lệ nhận dạng đạt 25,93%. Nhìn chung, sau khi phân khu tỷ lệ nhận dạng các vỉa than theo các thông số vỉa và chất lượng than của từng môi trường đã tăng lên đáng kể.

2) Kết hợp các đặc tính về chiều dày, góc dốc, số lớp kẹp và chất lượng than, cho phép phân khu mỏ Núi Béo thành 07 khu vực đồng nhất tương đối mang các đặc trưng khác nhau về môi trường trầm tích, trong đó khu A1 mang đặc tính môi trường đầm lầy; khu A2 và A3 mang đặc tính môi trường bãi triều; khu B mang đặc tính môi trường dòng chảy; khu C mang đặc tính bãi triều; khu D mang đặc tính môi trường đầm lầy; và khu E thuộc phần ngoại vi.

3) Kết quả nghiên cứu khẳng định vai trò của các phương pháp toán ứng dụng trong nghiên cứu địa chất, đặc biệt trong việc nhận dạng các vỉa than một cách khoa học và môi trường trầm tích chúng góp phần phục vụ đắc lực cho việc liên kết, đồng danh các vỉa than được chính xác và phù hợp hơn với môi trường, cũng như cấu trúc khu mỏ.

XUNG ĐỘT LỢI ÍCH

Các tác giả đồng ý không có bất kỳ xung đột lợi ích nào liên quan đến các kết quả đã công bố.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Khương Thế Hùng thực hiện nghiên cứu, tổng hợp dữ liệu, tính toán và hoàn thành bài báo.
Nguyễn Danh Tuyên thực hiện công tác thực địa, lấy mẫu, phân tích và chỉnh sửa hình ảnh.

TÀI LIỆU THAM KHẢO

1. Meng X, Ge M, Tucker ME. Sequence stratigraphy, sea-level changes and depositional systems in the Cambro-Ordovician of the North China carbonate platform. Sedimentary Geology, 114 (1-4), 1997;p. 189–222. Available from: https://doi.org/10.1016/S0037-0738(97)00073-0.
2. Hùng KT, Phương N, Cúc NT, Sang PN, Tuyên ND. Identifying Correlation of Coal Seams in the Tien Hai Area, Northern Vietnam by Using Multivariate Statistic Methods. Inżynieria Mineralna – Journal of the Polish Mineral Engineering Society, 2(46), 2021;p. 129–148. Available from: http://doi.org/10.29227/IM-2021-02-11.
3. Duan H, Xie W, Zhao J, Jia T. Sequence stratigraphy and coal accumulation model of the Taiyuan Formation in the Tashan Mine, Datong Basin, China. Open Geosciences, 13 (1), 2021;p. 1259–1272. Available from: https://doi.org/10.1515/geo-2020-0303.
4. Nalendra S, Kuncoro B, Burhanudin A. Thickness Variation of Coal Seams in Loa Janan Anticline: Implications for Exploration and Mining Activities. Proceedings joint convention Malang 2017, HAGI-IAGI-IAFMI-IATMI (JCM 2017) Ijen Suites Hotel, Malang, September 25-28, 2017;.
5. Einsele G. Sedimentary Basins, Evolution, Facies, and Sediment Budget. Springer-Verlag Berlin Heidelberg, 2020;. Available from: https://doi.org/10.1007/978-3-662-04029-4.
6. Hou H, Shao L, Tang Y, Li Y, Liang G, Xin Y, et al. Coal seam correlation in terrestrial basins by sequence stratigraphy and its implications for palaeoclimate and palaeoenvironment evolution. J Earth Sci. 2021;p. 1–24. Available from: http://kns.cnki.net/kcms/detail/42.1788.P.20200914.1528.002.html.
7. Hùng L (chủ biên). Bản đồ địa chất và khoáng sản nhóm tờ Hòn Gai - Cẩm Phả, tỷ lệ 1:50.000. Lưu trữ Địa chất - Tổng cục Địa chất và Khoáng sản Việt Nam, Hà Nội. 1996;.
8. Anh PT (chủ biên). Báo cáo chuyển đổi cấp trữ lượng và cấp tài nguyên than khu mỏ Hà Lầm, Lưu trữ tại Tập đoàn Công nghiệp Than - Khoáng sản Việt Nam, 2009;.
9. Rodionov DA. Statisticeskie resenija v geologii.: Izd. "Nedra", Moskva, 1981;pp. 231.
10. Rao CR. Linear Statistical Inference and Its Applications The Annals of Mathematical Statistics, 38 (1), 1967; p. 281–284.

Research article

Identifying the coal seams and their sedimentary environments in the Nui Beo mine, Quang Ninh province using K-means and regression methods

Khuong The Hung[1,*], Nguyen Danh Tuyen[2]

[1]*Faculty of Geosciences and Geoengineering, Hanoi University of Mining and Geology, Hanoi, Vietnam*
[2]*Vietbac Joint Stock Company, Vietnam National Coal and Mineral Industries Group, Hanoi, Vietnam*
[*]Correspondence: Khuong The Hung
Faculty of Geosciences and Geoengineering, Hanoi University of Mining and Geology, Hanoi, Vietnam
Email: khuongthehung@humg.edu.vn

History: Received: 08-02-2022; Accepted: 29-11-2022; Published:

Abstract

The Nui Beo mine is located in the southern part of the Quang Ninh province. It belongs to the Hon Gai-Cam Pha coal zone, where many coal resources have been estimated in Vietnam. Based on synthesizing and processing data of the coal seam parameters and coal quality by using K-means and regression methods, results allow dividing the Nui Beo mine into three areas, the area 1 bearing coal thickness greater than or equal to 18.5m, the area 2 containing coal thickness less than 18.5m, and coal seam dipping angle less than 32.0 degrees; and the area 3 obtaining coal thickness less than 18.5m, and coal seam dipping angle greater than or equal to 32.0 degrees, combining with five interlayers as the scared boundary. Identifying results using K-means and regression methods for 2 or 3 coal seam parameters is very low; only coal seams of V4, V12, and V14 are higher (over 50%). After dividing the mining area, the identifying rate is higher, especially the coal seams of V5, V7, V11, V12, and V14 reach 30.38% on average compared to 16.83% for the whole mining area. Identifying coal seams based on six analyzed parameters shows 16.64% on average; the nine analyzed parameters show 25.93% on average. Generally, each environment has increased significantly after dividing the identifying rates of coal seams according to seam parameters and coal quality. Using the seam characteristics as thickness, dip angle, the number of interlayers, and coal quality, they help divide the Nui Beo mine into 07 relatively homogenous blocks, namely A1, A2, A3, B, C, D, and E, respectively. In which block A1 is formed in a swamp environment; blocks A2 and A3 are characterized by a mudflat environment; block B is suggested a flow environment; block C is characterized by mudflat; block D features a swampy and block E in the periphery. The results confirm that applied mathematical methods in geological science are effective, especially in identifying coal seams and sedimentary environments.

Keywords: Coal seams, sedimentary environment, Nui Beo mine, Quang Ninh province