Ứng dụng phân tích thống kê đa biến trong đánh giá chất lượng nước dưới đất huyện Tân Thành, tỉnh Bà Rịa – Vũng Tàu

Nguyễn Hải Âu, Phan Thị Khánh Ngân, Hoàng Thị Thanh Thủy, Phan Nguyễn Hồng Ngọc

Tóm tắt **- Ở nghiên cứu này, các phương pháp phân tích thống kê đa biến (MSA) như phân tích thành phần chính (PCA) và phân tích cụm (CA) được ứng dụng cho việc xác định sự biến thiên về không gian và thời gian của chất lượng nước dưới đất huyện Tân Thành, tỉnh Bà Rịa – Vũng Tàu. Các mẫu nước dưới đất được thu thập từ 18 giếng quan trắc vào tháng 4 (mùa khô) và tháng 10 (mùa mưa) trong năm 2012. Mười lăm thông số chất lượng nước (pH, độ cứng, TDS, Cl^-, F^-, NO_3^-, SO_4^{2-}, Cr^{6+}, Cu^{2+}, Ca^{2+}, Mg^{2+}, Na^+, K^+, HCO_3^- và Fe^{2+}) được lựa chọn để tiến hành phân tích thống kê đa biến.**

PCA xác định được ba thành phần chính ảnh hưởng đến chất lượng nước dưới đất. Ba thành phần chính gồm yếu tố nhiễm mặn, sự tương tác của các thành phần thạch học và nhân sinh đã giải thích được 70,5% (mùa khô) và 71,28 % (mùa mưa) biến thiên phương sai của tập mẫu. Kết quả phân tích cụm (CA) chỉ ra 2 nhóm khác nhau với sự đồng nhất trong nội bộ từng cụm.

Kết quả nghiên cứu đã cho thấy tính cần thiết của phân tích thống kê đa biến trong xử lý bộ dữ liệu quan trắc để trích rút ra những thông tin cần thiết phục vụ quản lý tài nguyên nước dưới đất.

Từ khoá**- phân tích thống kê đa biến, phân tích thành phần chính, phân tích cụm, nước dưới đất, quan trắc môi trường.**

Bài nhận ngày 30 tháng 05 năm 2017, chấp nhận đăng ngày 28 tháng 11 năm 2017

Nguyễn Hải Âu, Viện Môi trường và Tài nguyên, ĐHQG-TP.HCM, (Email: haiauvtn@gmail.com)

Phan Thị Khánh Ngân, Viện Môi trường và Tài nguyên, ĐHQG,TP.HCM, (Email: khanhngan2109@gmail.com)

Hoàng Thị Thanh Thủy, Đại học Tài nguyên và Môi trường TP.HCM, (Email: httthuy@hcmunre.edu.vn)

Phan Nguyễn Hồng Ngọc, Đại học Tài nguyên và Môi trường TP.HCM (Email: ngocphan1201@gmail.com)

1 GIỚI THIỆU

Phân tích thống kê đa biến (MSA - Multivariate Statistics Analysis) bao gồm các kỹ thuật thống kê khác nhau, bao gồm phân tích cụm (CA - Cluster Analysis) và phân tích biệt số (DA - Discriminant Analysis), phân tích nhân tố (FA - Factor Analysis), phân tích thành phần chính (PCA - Principal Component Analysis), phân tích phương sai đa biến (MANOVA),…trong đó, PCA và CA là 2 phương pháp được sử dụng phổ biến nhất [1].

PCA được áp dụng để giảm số chiều của một tập dữ liệu bao gồm một số lượng lớn của các biến liên quan. Những cắt giảm được thực hiện bằng cách chuyển đổi các dữ liệu vào một tập mới của các biến, các thành phần chủ yếu (PCs), đó là trực giao (không tương quan) và được sắp xếp theo thứ tự giảm dần tầm quan trọng. CA là phương pháp phân loại các đối tượng hay các biến sao cho các đối tượng trong cùng một cụm xét theo các đặc tính được chọn để phân tích [2].

Trong những năm gần đây, các phương pháp PCA và CA đã được sử dụng khá rộng rãi trong các ứng dụng môi trường, bao gồm các đánh giá quan trắc diễn biến chất lượng nước ngầm, nước mặt, kiểm tra kết quả các mô hình mô phỏng chất lượng nước theo không gian và thời gian, xác định các yếu tố hóa học liên quan đến các điều kiện thủy văn, và đánh giá các chỉ thị chất lượng môi trường [3, 4]. Ở Mỹ và các nước Châu Âu như Pháp, Thổ Nhĩ Kỳ [5] và các quốc gia ở Châu Á như Malaysia [6], Trung Quốc [7], Nhật Bản [4], Ấn Độ [8-10], các nghiên cứu này đã ứng dụng các phương pháp MSA đánh giá chất lượng nước mặt, nước dưới đất ở các lưu vực sông dựa vào mối quan hệ giữa các thông số quan trắc với các đặc điểm các tầng chứa nước, từ đó đề xuất được các thông số đặc trưng chất lượng nước để giám sát và quản lí hiệu quả.

Ở Việt Nam, các kĩ thuật thống kê đa biến cũng được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau như tâm lý, kinh tế, xã hội, kỹ thuật trong đó có lĩnh vực môi trường (chủ yếu là sử dụng phương pháp phân tích hồi quy tuyến tính để xử lí các số

liệu phân tích thí nghiệm, kiểm định thông số mô hình mô phỏng. Phân tích đánh giá chất lượng nước chỉ dựa vào phương pháp so sánh với qui chuẩn cho phép, phân vùng dựa vào chỉ số chất lượng nước (WQI) và mô hình toán), chưa có nghiên cứu nào sử dụng các kĩ thuật phân tích thống kê đa biến đánh giá chất lượng nước một cách đầy đủ, riêng biệt [11].

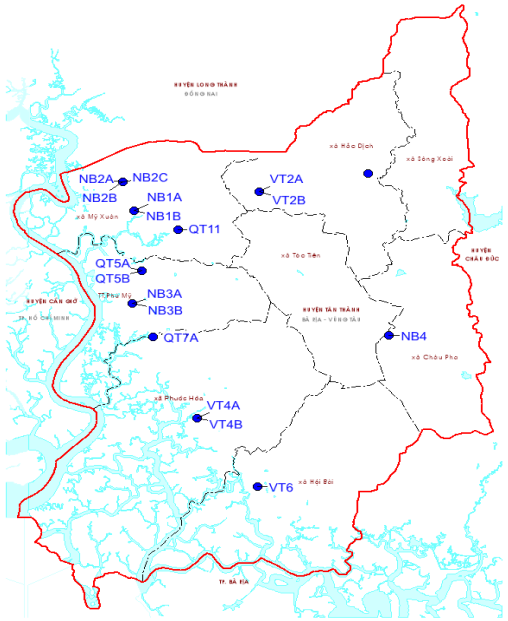
Trong nghiên cứu này, tập trung phân tích các thông số lí hóa của chất lượng nước dưới đất từ số liệu quan trắc các tầng chứa nước Pleistocen bằng cách sử dụng các kỹ thuật đa biến PCA, CA kết hợp với đặc điểm tầng chứa nước và phân bố nguồn thải trên địa bàn huyện Tân Thành, tỉnh Bà Rịa - Vũng Tàu.

2 TÀI LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Mô tả vùng nghiên cứu

Huyện Tân Thành đã và đang trở thành một trong ba địa phương có nền kinh tế phát triển bậc nhất tỉnh Bà Rịa – Vũng Tàu, là nơi tập trung nhiều khu công nghiệp nhất của tỉnh. Nhu cầu sử dụng nước trên địa bàn huyện là khá lớn để phục vụ phát triển kinh tế, trong khi đó các nguồn khai thác nước mặt từ các sông, hồ trên địa bàn không đáp ứng được nhu cầu sử dụng. Theo số liệu điều tra năm 2012 của Sở Tài Nguyên và Môi trường tỉnh Bà Rịa - Vũng Tàu thì tổng lưu lượng khai thác nước dưới đất trên địa bàn huyện trung bình 18.608.430 m^3/năm (chủ yếu từ trạm cấp nước Phú Mỹ - Mỹ Xuân và Nhà máy nước Tóc Tiên), trong đó cấp nước sinh hoạt là 13.174.310m^3/năm và cho sản xuất là 5.434.120 m^3/năm, nước được khai thác chủ yếu trong tầng chứa nước lỗ hổng các trầm tích Pleistocen. Tầng chứa nước lỗ hổng Pleistocen được tạo thành từ đất đá hạt thô nằm dưới hệ tầng Củ Chi, hệ tầng Thủ Đức và hệ tầng Trảng Bom với các khoáng vật chính: Fluorit-apatit, felspat, thạch cao, turmalin, montmorilonit, ilmenit và một số tạp chất khác [12, 13].

Huyện Tân Thành có diện tích tự nhiên là 33.825 ha, dân số trung bình là 137.334 người, phía Đông giáp huyện Châu Đức, phía Tây giáp huyện Cần Giờ và thành phố Vũng Tàu, phía Nam giáp thành phố Bà Rịa và phía Bắc giáp huyện Long Thành, tỉnh Đồng Nai. Huyện Tân Thành nằm trong vùng khí hậu đồng bằng Nam Bộ có khí hậu nhiệt đới gần xích đạo, chịu ảnh hưởng chủ yếu từ gió mùa Đông Bắc và Tây Nam. Một năm có hai mùa rõ rệt là mùa khô và mùa mưa. Mùa khô kéo dài từ tháng 12 đến tháng 4 năm sau, có nền nhiệt độ trung bình hàng năm cao (26,3^oC) và hầu như không thay đổi nhiều trong năm. Mùa mưa bắt đầu từ tháng 5 đến tháng 11, lượng mưa trung bình năm là 1356.5 mm/năm.



Hình 1. Sơ đồ vị trí lấy mẫu quan trắc nước dưới đất huyện Tân Thành, tỉnh Bà Rịa – Vũng Tàu

2.2 Tài liệu nghiên cứu

Trong bài báo này, 15 thông số chất lượng nước (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^{-}, pH, độ cứng, TDS, Cl^{-}, F^{-}, NO_3^{-}, SO_4^{2-}, Cr^{6+}, Cu^{2+} và Fe^{2+}) từ 18 giếng quan trắc chất lượng nước dưới đất trên địa bàn huyện Tân Thành được Sở Tài nguyên và Môi trường tỉnh Bà Rịa – Vũng Tàu thực hiện vào mùa khô năm 2012 (NB2C, NB2A, NB3A, NB1B, NB4, VT2B, NB1A, VT2A, VT6, QT5B, NB2B, QT11, QT7B, NB3B, QT5A, VT4B, QT7A, VT4A) được lựa chọn xử lý và đánh giá. Vị trí các giếng quan trắc được trình bày trong sơ đồ vị trí quan trắc ở Hình 1.

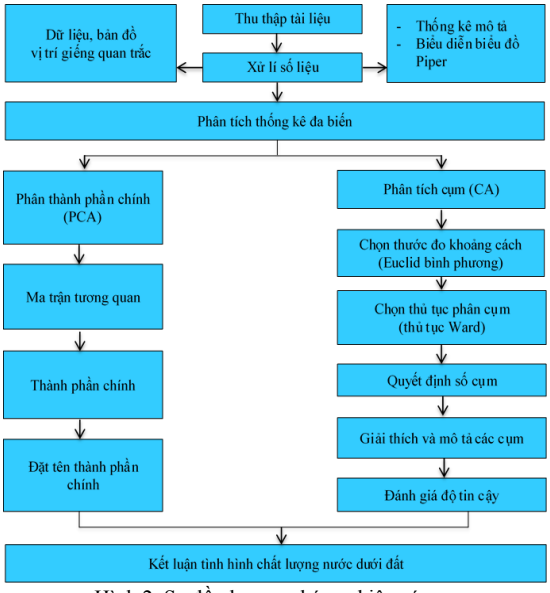
3 PHƯƠNG PHÁP PHÂN TÍCH THỐNG KÊ ĐA BIẾN VÀ XỬ LÝ SỐ LIỆU

Tất cả các tính toán toán học và thống kê được thực hiện bằng cách sử dụng phần mềm EXCEL 2013 (Microsoft Office). Các phân tích CA, DA được xử lí bằng phần mềm SPSS. Sơ đồ phương pháp nghiên cứu được trình bày ở Hình 2.

3.1 Phân tích thành phần chính (PCA)

Kỹ thuật PCA rút ra các giá trị riêng và phương sai từ ma trận tương quan của các biến ban đầu. Các thành phần chính là các biến không tương quan, thu được bằng cách nhân các biến tương quan ban đầu với hệ số tải nhân tố. Vì vậy, các thành phần chính

được kết hợp tuyến tính của các biến ban đầu. PC cung cấp thông tin về các thông số có ý nghĩa nhất, trong đó mô tả toàn bộ dữ liệu thiết lập dựng hình dữ liệu giảm với sự giảm tối thiểu các thông tin ban đầu. Nó là một kỹ thuật mạnh mẽ cho mô hình giải thích sự thay đổi của một tập lớn các tương quan biến và chuyển đổi thành một tập hợp nhỏ hơn của các biến độc lập (thành phần chính) [3].



Hình 2. Sơ đồ phương pháp nghiên cứu

FA tiếp tục làm giảm sự đóng góp của ít biến quan trọng thu được từ PCAvà nhóm mới của các biến được rút ra thông qua việc quay trên trục được xác định bởi PCA. Trục đồ thị xác định bởi PCA quay để giảm sự liên kết các biến ít quan trọng. FA có thể được biểu diễn như sau:

$$F_i = a_1y_{1j} + a_2y_{2j} + \ldots + a_my_{mj} \quad (1)$$

Khi Fi là nhân tố, a là hệ số tải nhân tố, y là giá trị đo của biến, i là số nhân tố, j là số mẫu và m là tổng số biến. Và các nhân số (các điểm số tổng số ước lượng được cho từng quan sát trên các nhân tố được rút ra) có thể được biểu thị như sau:

$$Z_{ij} = a_1f_{1j} + a_2f_{2j} + \ldots\ldots + a_mf_{mj} + e_{ij} \quad (2)$$

3.2 Phân tích cụm (CA)

Trong nghiên cứu này, phương pháp phân tích CA được lựa chọn là phương pháp phân tích cụm tích tụ dựa vào phương sai là "thủ tục Ward" trong loại thủ tục phân cụm thứ bậc (Hierarchical clustering). Theo thủ tục Ward thì ta sẽ tính giá trị trung bình tất cả các biến cho từng cụm một. Sau đó tính khoảng cách Euclid bình phương (Squared Euclidean distance) giữa các phần tử trong cụm với giá trị trung bình của cụm, rồi lấy tổng tất cả các khoảng cách bình phương này. Ở mỗi giai đoạn tích tụ thì hai cụm có phần tăng trong tổng các khoảng cách bình phương trong nội bộ cụm nếu kết hợp với nhau là nhỏ nhất sẽ được kết hợp. Cụ thể hơn, trong phương pháp này khoảng cách hoặc sự giống nhau giữa 2 nhóm A và B được xem là khoảng cách nhỏ nhất giữa 2 điểm A và B [6]:

$$D(A,B) = \text{Range}\{d(x_i,x_j), \text{ với } x_i \in A \text{ và } x_j \in B\} \quad (3)$$

Khi d(xi,xj) là khoảng cách Euclid bình phương trong công thức (3). Tại mỗi bước khoảng cách được tìm thấy từng cặp nhóm và 2 nhóm có khoảng cách nhỏ nhất (giống nhau nhiều nhất) được gộp lại. Sau khi 2 nhóm được gộp lại, tiếp tục lặp lại các bước tiếp theo: khoảng cách giữa tất cả các cặp nhóm được tính lại lần nữa, và cặp có khoảng cách ngắn nhất được gộp vào nhóm đơn. Kết quả của việc phân nhóm cấu trúc được biểu diễn bằng đồ thị - biểu đồ hình cây.

4 KẾT QUẢ VÀ THẢO LUẬN

Thống kê mô tả về bộ dữ liệu thông số chất lượng nước dưới đất được thể hiện trong Bảng 1. Sự phân bố các thông số chất lượng nước dưới đất được đánh giá bằng cách xác định giá trị lớn nhất, giá trị nhỏ nhất, giá trị trung vị, độ lệch chuẩn của tập dữ liệu quan trắc gồm 15 thông số. Kết quả thấy được xu hướng biến động của các thông số chất lượng nước được lấy ở 18 giếng quan trắc tầng chứa nước Pleistocen khu vực nghiên cứu.

4.1 Phân tích PCA

PCA được sử dụng phân tích 15 thông số chất lượng nước. Bảng 2 thể hiện kết quả phân tích nhân tố được thực hiện bởi phương pháp rút trích nhân tố (phương pháp mặc định là rút các thành phần chính). Vòng xoay nhân tố chính được thực hiện theo phương pháp xoay nguyên gốc các nhân tố để tối thiểu hoá số lượng biến có hệ số lớn tại cùng một nhân tố (Vanimax với Kaiser bình thường). Trong bài báo này, tất cả các liên hệ giữa các thông số có hệ số tương quan lớn hơn 50% đều được góp phần xác định thông số chất lượng nước đặc trưng của khu vực nghiên cứu.

Đối với dữ liệu mùa khô, ba thành phần chính được rút trích giải thích được 70,50% tổng phương sai bộ dữ liệu thủy hoá. Với giá trị tổng phương sai đạt 38,71% của thành phần 1, có liên quan đến các biến gồm Cl^-, độ cứng, Fe^{2+}, Na^+, K^+, Mg^{2+}, Ca^{2+} và TDS. Các thông số tương quan trong thành phần 1 được giải thích cho việc chất lượng nước khu vực nghiên cứu chịu ảnh hưởng khá lớn từ thành phần hóa học có trong trầm tích sông-sông biển hiện hữu

hoặc có xu hướng bị nhiễm mặn từ biển. Sự nhiễm mặn thể hiện ở hàm lượng cao của TDS cũng như xu hướng tập trung cao các ion Cl^-, Na^+, K^+, Mg^{2+} và một số muối sắt hiện hữu. Trong tính chất nước dưới đất, khi có sự tương quan chặt giữa Cl^- thường đi cùng với các ion Na^+, K^+, Mg^{2+} đây là các thành phần có trong nước biển chôn vùi hoặc nước biển xâm nhập vào tầng chứa nước thông qua các cửa sông. Thành phần 2 với tổng giá trị phương sai đạt 16,59%, có sự tương quan các biến HCO_3^-, F^-, NO_3^- và SO_4^{2-}. Các thông số tương quan trong thành phần 2 được giải thích cho việc chất lượng nước khu vực chịu ảnh hưởng bởi đặc điểm đất đá của tầng chứa nước, chủ yếu là do sự rửa lủa đá vôi ($CaCO_3$), dolomit ($CaMg[CO_3]_2$) và sét vôi với sự có mặt của khí CO_2^-. Thành phần 3 được giải thích với giá trị phương sai đạt 15,206%, bao gồm các thông số Cu, Cr^{6+} và pH, trong đó mối tương quan giữa các thông số khác với thông số pH thường ở dạng thuận nghịch. Tuy nhiên, mối tương quan của các thông số trong thành phần 3 chưa thật sự mạnh mẽ và chưa cho thấy xu hướng chung về các đặc trưng của chất lượng nước.

Đối với dữ liệu mùa mưa, ba thành phần chính được rút trích giải thích được 71,28% tổng phương sai bộ dữ liệu thủy hoá. So với 3 thành phần ở mùa khô thì 3 thành phần rút trích ở mùa mưa có sự thay đổi về thành phần các thông số tương quan đại diện cho mùa mưa, chủ yếu là nước nhạt với tổng khoáng hóa giảm. Với giá trị tổng phương sai đạt 41,86% của thành phần 1, các thông số tương quan gồm Cl^-, độ cứng, F^-, Fe^{2+}, Na^+, K^+, Mg^{2+} và Ca^{2+}. So với mùa khô, hai thông số đặc trưng là Na^+ và Cl^- đã giảm do lượng bổ cập nước nhạt lớn vào tầng chứa nước, tuy nhiên sự tương quan trong thành phần 1 cũng được giải thích cho việc chất lượng nước khu vực nghiên cứu chịu ảnh hưởng khá lớn từ biển. Thành phần 2 với tổng giá trị phương sai đạt 18,703 % có sự tương quan các biến TDS, độ cứng, Cu^{2+}, SO_4^{2-} và HCO_3^-.

Các thông số tương quan trong thành phần 2 được giải thích cho việc chất lượng nước khu vực ngoài chịu ảnh hưởng bởi đặc điểm đất đá của tầng chứa nước (chủ yếu là đá vôi, sét vôi và khoáng dolomit) còn nhiều tác động lớn chất lượng nước mặt, nước mưa chảy tràn bề mặt bổ cập vào các giếng. Tương tự như mùa khô, thành phần 3 ở mùa mưa được giải thích với giá trị phương sai đạt 10,88%, bao gồm các thông số Fe^{2+}, Cu^{2+}, NO_3^-, Cr^{6+} và pH. Vai trò giải thích phần trăm phương sai ở thành phần này không đáng kể đối với tập dữ liệu mô hình cũng chưa thể hiện được xu hướng chung của chất lượng nước.

Nhìn chung, qua kết quả phân tích chất lượng nước dưới đất huyện Tân Thành bằng phương pháp PCA cho thấy chất lượng nước có hai đặc trưng là xu hướng nhiễm mặn từ biển và đặc điểm thủy địa hóa tầng chứa nước. Kết quả phân tích này sẽ được chỉ ra rõ hơn trong phương pháp phân tích nhóm (CA) ở phần tiếp theo.

Bảng 1. Thống kê mô tả các thông số chất lượng nước dưới đất khu vực nghiên cứu

Thông số	Đơn vị	Mùa khô					Mùa mưa				
		Max	Min	Median	Mean	Std	Max	Min	Median	Mean	Std
Ca	mg/l	430,86	1,40	7,76	40,34	99,34	100,20	1,40	16,03	20,39	24,00
Mg	mg/l	145,92	0	1,22	17,52	37,60	118,56	0,24	0,79	11,30	28,32
Na	mg/l	1223	2,57	7,67	133,71	304,29	644,44	3,57	7,12	72,92	161,32
K	mg/l	105	0,41	2,58	11,42	24,83	57,90	1,22	3,26	7,31	13,24
HCO_3^-	mg/l	289,85	0	30,51	51,17	74,13	494,26	0	51,87	67,12	111,22
pH	-	7,28	4,1	6,13	6,14	0,90	7,33	4,16	6,17	6,21	0,92
TH	mg/l	657,5	8,5	46,25	104,65	162,07	308,5	8	22	73,06	103,49
TDS	mg/l	1357	38	105,5	261,53	379,68	1368	43	87,5	218,33	365,76
Cl	mg/l	953,61	7,09	17,73	121,47	258,69	475,03	7,09	13,29	65,92	126,11
F	mg/l	1,63	0	0,14	0,27	0,37	1,70	0	0,11	0,21	0,38

NO_3^-	mg/l	2,44	0,11	0,60	0,80	0,65	6,09	0,17	0,48	1,27	1,83
SO_4^{2-}	mg/l	95,58	5,76	47,59	50,04	24,41	59,08	2,88	13,21	17,18	13,03
Cr^{6+}	mg/l	0,07	0	0	0,01	0,02	0	0	0	0	0
Cu	mg/l	0,11	0,01	0,02	0,03	0,03	0,02	0	0	0	0
Fe	mg/l	69,36	0,16	2,70	7,30	15,79	46,45	0,18	1,42	7,45	12,27

Ghi chú: Max – Giá trị lớn nhất; Min – Giá trị nhỏ nhất; Median – Giá trị trung vị; TH: Độ cứng
Mean – Giá trị trung bình; Std - Độ lệch chuẩn;

Bảng 2. Ma trận rút trích thành phần chính mùa khô và mùa mưa năm 2012

	Mùa khô			Mùa mưa		
Thông số	Thành phần 1	Thành phần 2	Thành phần 3	Thành phần 1	Thành phần 2	Thành phần 3
Na	0,944			0,796		
Mg	0,908			0,764		
TDS	0,883				0,822	
Cl	0,873			0,703		
K	0,872			0,850		
Ca	0,794			0,658		
Fe	0,648			0,529		-0,745
Độ cứng (TH)	0,618			0,713		
HCO_3^-		0,771			-0,527	
F		0,763		0,878		
NO_3^-		0,743				0,885
SO_4^{2-}		0,521			0,678	
Cu			0,844		0,615	
Cr^{6+}			0,838			0,741
pH			-0,635			0,579
Eigenvalues	5,807	2,488	2,281	6,255	2,805	1,633
% Phương sai	38,710	16,588	15,206	41,698	18,703	10,884
% Tích lũy	38,710	55,299	70,505	41,698	60,401	71,286

Ghi chú: Phương pháp phân tích: Phân tích thành phần chính (PCA - Principal Component Analysis)
Phương pháp xoay: Varimax with Kaiser Normalization

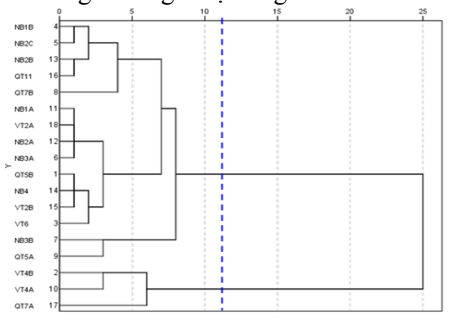
Bảng 3. Ma trận rút trích thành phần chính mùa khô và mùa mưa năm 2012

		Mùa khô		Mùa mưa	
Thông số quan trắc	Đơn vị	Cụm 1	Cụm 2	Cụm 1	Cụm 2
		NB2C, NB2A, NB3A, NB1B, NB4, VT2B, NB1A, VT2A, VT6, QT5B, NB2B, QT11, QT7B, NB3B, QT5A	VT4B, QT7A, VT4A	NB2C, NB2A, NB3A, NB1B, NB4, VT2B, NB1A, VT2A, VT6, QT5B, NB2B, QT11, QT7B, NB3B, QT5A	VT4B, QT7A, VT4A
		Giá trị trung bình		Giá trị trung bình	
Ca^{2+}	mg/l	12,54	179,36	12,45	60,12
Mg^{2+}	mg/l	3,21	89,07	2,08	57,36
Na^+	mg/l	24,08	681,88	20,51	334,94
K^+	mg/l	3,07	53,18	3,35	27,09
HCO_3^-	mg/l	42,08	96,62	41,9	193,23
pH	-	6,23	5,66	6,31	5,7
Độ cứng	mg/l	44,01	407,83	29,27	292
TDS	mg/l	105,64	1041	97,07	824,67
Cl^-	mg/l	24,05	608,56	15,41	318,46
F^-	mg/l	0,2	0,6	0,12	0,67
NO_3^-	mg/l	0,89	0,36	1,27	1,27
SO_4^{2-}	mg/l	45,47	72,85	14,47	30,74
Cr^{6+}	mg/l	0,01	0,01	0	0
Cu^{2+}	mg/l	0,03	0,04	0	0,01
Fe^{2+}	mg/l	2,86	29,52	6,75	10,91

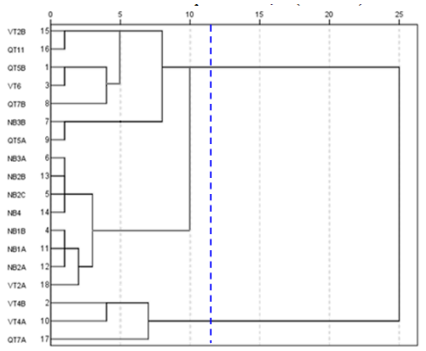
4.2 Phân tích CA

Phân tích cụm đã được áp dụng để kết hợp các giếng trong khu vực nghiên cứu vào các nhóm đồng nhất do chất lượng nước ngầm. Trong nghiên cứu này, phương pháp liên kết Ward với khoảng cách Euclide bình phương đã được sử dụng để nhóm giếng khảo sát vào các cụm. Phân tích cụm cho thấy hai nhóm nước ngầm (Hình 3, Hình 4 và Bảng 3). Dựa vào Bảng 3 cho thấy kết quả phân cụm khá tương đồng cho cả mùa mưa và mùa khô, cụm 1 gồm 15 giếng (NB2C, NB2A, NB3A, NB1B, NB4, VT2B, NB1A, VT2A, VT6, QT5B, NB2B, QT11, QT7B, NB3B, QT5A) đại diện cho các giếng nước

nhạt và cụm 2 gồm 3 giếng (VT4B, QT7A, VT4A) đại diện cho các giếng nước bị nhiễm mặn thông qua các thông số có giá trị trung bình cao như TDS (mùa khô 1041 mg/l, mùa mưa 824 mg/l); Clorua (mùa khô 608 mg/l, mùa mưa 318 mg/l).



Hình. 3. Biểu đồ phân tích cụm (mùa khô)



Hình. 4. Biểu đồ phân tích cụm (Mùa mưa)

KẾT LUẬN

Kỹ thuật thống kê đa biến được ứng dụng trong nghiên cứu này như một công cụ phân tích rõ hơn về chất lượng nước dưới đất, giúp các nhà quản lý hiểu rõ hơn về sự biến đổi theo không gian của chất lượng nước dưới đất, từ đó đưa ra được các giải pháp nhằm quản lý bền vững nguồn tài nguyên nước. Kết quả phân tích thành phần chính (PCA) đã chỉ ra được ba nhân tố chính của chất lượng nước dưới đất và kết quả phân tích CA cũng chỉ ra được có 2 nhóm các giếng có chất lượng nước tương đồng. Kết quả phân tích cho thấy chất lượng nước có sự thay đổi về phân bố các ion ở các thành phần chính theo mùa, tuy nhiên với chuỗi số liệu còn hạn chế nên các thay đổi này chưa được giải thích rõ về mối tương quan của nó trong nước dưới đất tầng pleistocen khu vực. Riêng các ion đại diện cho xu hướng nhiễm mặn như TDS, Cl^-, Na^+, K^+, SO_4^{2-} thì kết quả phân tích đã được giải thích là khá thỏa đáng.

LỜI CẢM ƠN

Để hoàn thành nghiên cứu này, nhóm tác giả trân trọng cảm ơn Viện Môi trường và Tài Nguyên, ĐHQG Tp.HCM đã hỗ trợ kinh phí cho nghiên cứu này. Các tác giả cũng chân thành cảm ơn sự hỗ trợ của Sở Tài nguyên và Môi trường tỉnh Bà Rịa – Vũng Tàu trong việc cung cấp các dữ liệu quan trắc chất lượng nước dưới đất năm 2012 tại huyện Tân Thành, tỉnh Bà Rịa – Vũng Tàu.

TÀI LIỆU THAM KHẢO

[1] N. V. Tuấn, "Phân tích dữ liệu với R" tại *NXB Tổng hợp TP. HCM* . Thành phố Hồ Chí Minh , 2016.

[2] F. Akbal, L. Gürel, T. Bahadır, İ. Güler, G. Bakan, and H. Büyükgüngör, "Multivariate statistical techniques for the assessment of surface water quality at the Mid-Black Sea Coast of Turkey," *Water, Air, & Soil Pollution*, vol. 216, pp. 21-37, 2011.

[3] E. M. de Andrade, H. A. Q. Palácio, I. H. Souza, R. A. de Oliveira Leão, and M. J. Guerreiro, "Land use effects in groundwater composition of an alluvial aquifer (Trussu River, Brazil) by multivariate techniques," *Environmental Research*, vol. 106, pp. 170-177, 2008.

[4] S. Shrestha and F. Kazama, "Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan," *Environmental Modelling & Software*, vol. 22, pp. 464-475, 2007.

[5] M. Varol and B. Sen, "Assessment of surface water quality using multivariate statistical techniques: a case study of Behrimaz Stream, Turkey," *Environ Monit Assess*, vol. 159, pp. 543-53, Dec 2009.

[6] A. F. Alkarkhi, A. Ahmad, and A. M. Easa, "Assessment of surface water quality of selected estuaries of Malaysia: multivariate statistical techniques," *The Environmentalist*, vol. 29, pp. 255-262, 2009.

[7] H. Bu, X. Tan, S. Li, and Q. Zhang, "Water quality assessment of the Jinshui River (China) using multivariate statistical techniques," *Environmental Earth Sciences*, vol. 60, pp. 1631-1639, 2010.

[8] T. A. Khan, "Groundwater Quality Evaluation Using Multivariate Methods, in Parts of Ganga Sot Sub-Basin, Ganga Basin, India," *Journal of Water Resource and Protection*, vol. 7, p. 769, 2015.

[9] S. A. Romshoo, R. A. Dar, K. O. Murtaza, I. Rashid, and F. A. Dar, "Hydrochemical characterization and pollution assessment of groundwater in Jammu Siwaliks, India," *Environmental monitoring and assessment*, vol. 189, p. 122, 2017.

[10] S. Gholami and S. Srikantaswamy, "Statistical multivariate analysis in the assessment of river water quality in the vicinity of KRS Dam, Karnataka, India," *Natural resources research*, vol. 18, pp. 235-247, 2009.

[11] N. H. Âu, Vũ Văn Nghị, Lê Thanh Hải, "Bước đầu áp dụng kỹ thuật phân tích thống kê đa biến phân tích số liệu chất lượng nước lưu vực sông Thị Tính, Tỉnh Bình Dương," *Tạp chí Phát triển khoa học và công nghệ của Viện Hàn Lâm Khoa học và Công nghệ Việt Nam*, vol. 52 (2B), p. 9, 2014.

[12] Sở Tài Nguyên và Môi trường tỉnh Bà Rịa - Vũng Tàu, "Nghiên cứu điều tra bổ sung, quy hoạch quản lý khai thác, bảo vệ bền vững tài nguyên nước dưới đất tỉnh Bà Rịa - Vũng Tàu", 2012.

[13] Sở Tài Nguyên và Môi trường tỉnh Bà Rịa - Vũng Tàu, "Vận hành mạng quan trắc nước dưới đất tỉnh Bà Rịa-Vũng Tàu", 2015.

Application of multivariate statistical analysis in the assessment of groundwater quality of Tan Thanh district, Ba Ria – Vung Tau province

Nguyen Hai Au, Phan Thi Khanh Ngan, Hoang Thi Thanh Thuy, Phan Nguyen Hong Ngoc

Abstract **- In the present study, Multivariate Statistical Analysis (MSA) such as Principle Component Analysis (PCA) and Cluster Analysis (CA) were applied to determine the temporal and spatial variations of groundwater quality in Tan Thanh district, Ba Ria – Vung Tau province. Groundwater samples were collected from 18 monitoring wells in April (dry season) and October (wet season) during the year 2012. Fifteen parameters (pH, TH, TDS, Cl^-, F^-, NO_3^-, SO_4^{2-}, Cr^{6+}, Cu^{2+}, Ca^{2+}, Mg^{2+}, Na^+, K^+, HCO_3^- and Fe^{2+}) were selected for MSA.**

PCA identified a reduced number of mean three latent factors of groundwater quality. Three factors called salinization, water-rock interaction and anthropogenic pollution explanined 70,5% (dry season) and 71.28% (wet season) of the variances. Cluster analysis revealed two main different groups of similarities between the sampling sites.

This study presents the necessity of MSA in order to extract more precise information from a huge minitoring data, which will be usefull to groundwater quality management.

Keywords **- cluster analysis, environmental monitoring, groundwater, multivariate statistic analysis, principal component analysis.**