Journal of Japan Society of Civil Engineers, Ser.
B1 (Hydraulic Engineering), Vol. 70, No. 4, I_241-I_246, 2014.

SPATIAL CLASSIFICATION OF GROUNDWATER CHEMISTRY MONITORING DATA IN THE RED RIVER DELTA, VIETNAM USING SELF-ORGANIZING MAPS

Thuy Thanh NGUYEN[1], Akira KAWAMURA[2], Thanh Ngoc TONG[3],
Naoko NAKAGAWA[4], Hideo AMAGUCHI[5] and Romeo GILBUENA[1]

[1]Student Member of JSCE, Dept. of Civil and Environmental Engineering, Tokyo Metropolitan University (1-1 Minami-Ohsawa,Hachioji, Tokyo 192-0397, Japan)

[2]Member of JSCE, Dr. of Eng, Professor, Dept. of Civil and Environmental Engineering, Tokyo Metropolitan University (1-1 Minami-Ohsawa,Hachioji, Tokyo 192-0397, Japan)

[3]Dr. of Eng, National Center for Water Resources Planning and Investigation (93/95 Vu Xuan Thieu Street – Sai Dong Ward, Long Bien, District, Hanoi, Vietnam)

[4]Member of JSCE, Associate Research Professor, Dept. of Civil and Environmental Engineering, Tokyo Metropolitan University (1-1 Minami-Ohsawa,Hachioji, Tokyo 192-0397, Japan)

[5]Member of JSCE, Assistant Professor, Dept. of Civil and Environmental Engineering, Tokyo Metropolitan University (1-1 Minami-Ohsawa,Hachioji, Tokyo 192-0397, Japan)

The groundwater in the Pleistocene confined aquifer (PCA) of the Red River Delta (RRD) was examined using self-organizing map (SOM) and Gibbs diagrams to determine, for the first time, its spatial classifications in terms of its hydrogeochemical characteristics. In this study, the groundwater chemistry dataset used in the analysis is composed of 8 major dissolved ions (i.e. Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^{-}, Cl^{-}, SO_4^{2-} and CO_3^{2-}) that are consistently found in 52 groundwater monitoring wells within the study area. Based on the results, the groundwater in the PCA monitoring wells of the delta can be classified into 2 major water types: high and low salinity. Each water type is composed of cluster-types that have similar hydrogeochemical characteristics. The high salinity water type has 2 clusters (or sub-types), while the low salinity (or fresh water) has 4. From the Gibbs diagrams, results indicate that the high salinity in the groundwater is mostly influenced by either (or both) anthropogenic activities and salt water intrusion.

Key Words: *Hydrogeochemical characteristics, Self-Organizing Map, Gibbs diagrams, Pleistocence confined aquifer, the Red River Delta, major cations and anions*

1. INTRODUCTION

The Red River Delta (RRD) is the second largest delta in Vietnam with an area of about 14,000 km^2 and a population of around 18 million people (23% of Vietnam's total population), which makes it the country's most densely populated region[1)]. In the RRD, all its residents depend entirely on groundwater for their domestic water supply. Until recently, very few studies have been carried out to examine the RRD's groundwater-related issues, which are focused mainly on only a small portion of the delta, specifically in Hanoi (Vietnam's capital) and its surrounding areas (e.g. studies on land subsidence[2)], groundwater pollution[3)] and heavy metal contamination[4)]). There is no reference, however, on the study of hydrogeochemical characteristics of the groundwater anywhere in Vietnam, as far as the authors know.

Through the initiative of the national government (National Hydrogeological Database Project), hydrogeochemical data of the Pleistocene confined aquifer (PCA) in the RRD were collected in 2011 during rainy season. Taking advantage of this unique and comprehensive database, this paper is the first attempt to investigate the spatial characteristic of the hydrogeochemistry of groundwater in the RRD.

The hydrogeochemistry of groundwater is influenced by many factors, including the mineralogy of aquifers, the chemical composition of

rainfall and surface water, climate, topography, and anthropogenic activities[5]. Thus, the hydrogeochemical interpretation of the groundwater quality analyses of representative water samples can provide useful information on the geochemical processes hydrodynamics, origin of groundwater, and interaction of the groundwater with the aquifer materials.

In order to clearly understand the spatial characteristics of multivariate hydrogeochemical data, it is meaningful to obtain and extract information from the classification of multivariate data into several water types, which would represent the different characteristics of the groundwater. Recently, the Self-Organizing Map (SOM) has been widely used as a powerful and effective data analysis tool in the exploration of data properties using pattern classification and visualization on two-dimensional arrays. This study is the first attempt to apply SOM in combination with a hierarchical cluster analysis for classification of groundwater chemistry monitoring data in the RRD. Many researchers have already demonstrated the usefulness of the SOM in other research fields[6), 7)]. Chemical diagrams proposed by Gibbs[8)] are widely uses to infer the mechanism controlling the chemistry of surface and groundwater. In this study, Gibbs diagrams were also aptly used to elucidate the cause and significance of the hydrogeochemical characteristic defined by SOM.

2. STUDY AREA AND DATA USED

Fig.1 shows the geographical locations of the study area (the whole RRD) and 52 groundwater sampling wells in the Pleistocene confined aquifer (PCA). The RRD is composed of 10 provinces and 2 cities, which is situated in the tropical monsoonal region with two distinct seasons: the rainy (May to October) and the dry (November to April) seasons. The tidal range along the coast is around 4m[9)].

In our previous studies[9)], from the data of 778 boreholes, we found that the RRD is composed of Quaternary-aged unconsolidated sediments with a maximum thickness of 100 m. The groundwater mostly exists as porous water that forms the PCA. The thickness of the PCA fluctuates over a large range with an average of about 80m, and gradually increases from the northwest to southeast of the delta.

Recently, the national government of Vietnam has implemented the National Hydrogeological Database Project under the supervision of the Vietnam Department of Geology and Minerals, in order to construct the GIS-based hydrogeological database. To take advantage of the water quality

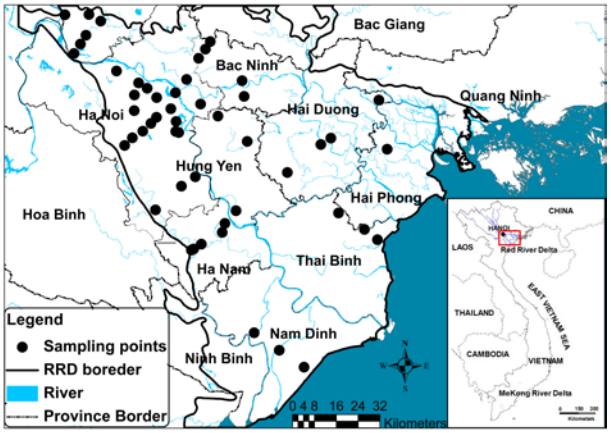

Fig.1 Study area and distribution of sampling points.

database in the National Hydrogeological Database Project, the authors of this study used the most recent groundwater chemical data, which were collected from 52 Pleistocene observation wells in August (rainy season) 2011 (which means 52 samples), in order to investigate the hydrogeochemical characteristics of groundwater in the RRD. The elevations that samples were taken vary according to the depth of the PCA.

The hydrogeochemical data used in this study consist of major dissolved ions found in the Pleistocene confined aquifer (Cations: Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}; Anions: HCO_3^{-}, Cl^{-}, SO_4^{2-} and CO_3^{2-}). Normalization of the data is necessary prior to the application of SOM to ensure that all values of the chemical parameters are given the same or similar importance. The results of the SOM application are highly sensitive to the data pre-processing method used, since the SOM is trained to be organized according to the Euclidean distances between input data. In this study, the range of the normalized values of the hydrogeochemical data, for all parameters, is 0 to 1.

3. METHODOLOGY

Self-organizing maps (SOMs), developed by Kohonen[10)], is a type of Artificial Neural Network that is characterized by unsupervised training. It can project high-dimensional, complex target data onto two-dimensional regularly-arranged units in proportion to the degree of similarity[6)]. Therefore, it is an effective tool to visualize and explore the properties of the data. In this study, the objective of the SOM application is to obtain useful and informative reference vectors. These vectors can be acquired after iterative updates through the training of the SOMs, which is composed of three main procedures: competition between nodes, selection of a winner node and updating of the reference vector.

Design of the SOM structure (calculation of the total number of nodes and side lengths), selection of

proper initialization and data transformation methods are very important features in SOM application. According to the properties of the SOM, the bigger the map size is, the higher the resolution for pattern recognition, while the topographical adjacency is farther among the clusters. A reasonable optimum solution of the compromise among the accuracy of pattern classification and topographical proximity of clusters to determine the number of the SOM nodes is the heuristic rule of $m = 5\sqrt{n}$, with m denoting the number of the SOM nodes, and n representing the number of input data[6),7),11),12)]. In this study, this heuristic formula was used to determine the total number of nodes in the SOMs. The ratio of the number of rows and columns is calculated by the square root of the ratio between the two biggest eigenvalues of the transformed data[13)].

After establishing the SOM structure, reference vectors for the SOMs, with the commonly used hexagonal array, are initially set using the linear initialization method. In this study, due to limited data, the linear initialization method was used, since it is more suitable for the pattern classification than the random initialization. The linear initialization approach can use eigenvalues and eigen vectors of the input data to set the initial reference vectors on the structured SOM. This means that the initial reference vectors already include prior information about the input data, resulting in an acceleration of the training phase[6),12)]. In this study, each reference vector is updated through the training process of the SOM using the batch mode. The reference vectors obtained at the end of the training process can be fine-tuned using cluster analysis methods.

Various clustering algorithms are available in the literature. These algorithms are generally classified into two types: hierarchical clustering and partitional clustering. For partitional clustering methods, the k-means algorithm is used most frequently for SOMs[6),7),13),14)]. The optimal number of cluster is selected by the Davies-Bouldin Index (DBI) using the k-means algorithm. The DBI values are calculated from a minimum of two clusters to the total number of nodes. The calculation is based on the "similarity within a cluster" and "dissimilarity between clusters". Therefore, the number of clusters showing the minimum DBI is optimal for the trained SOM. For hierarchical method, the Ward's linkage method is the most commonly used approach[6),7)]. In this study, the final fine-tuning cluster analysis was carried out by Ward's method using the optimal number of clusters. To investigate the mechanisms governing the chemistry of the groundwater in the RRD, the Gibbs diagrams were used to further evaluate the clustered data (wells).

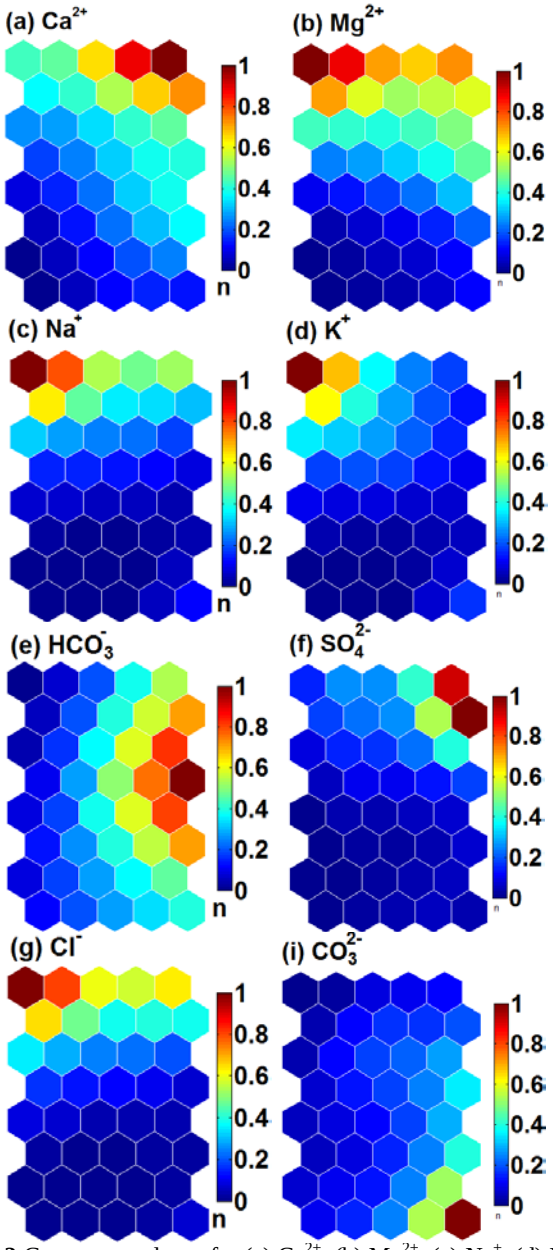

Fig.2 Component planes for (a) Ca^{2+}, (b) Mg^{2+}, (c) Na^{+}, (d) K^{+}, (e) HCO_3^{-}, (f) SO_4^{2-}, (g) Cl^{-}, (i) CO_3^{2-}.

4. RESULTS AND DISCUSSION

Based on the methodology described above, the number of SOM nodes was calculated as 40 nodes and the numbers of rows and columns are 8 and 5, respectively. This SOM was used for the cluster analysis of the standardized groundwater chemistry monitoring data from 52 Pleistocene wells in the RRD.

Fig. 2 shows the component planes of the 40 reference vectors (nodes) of the eight water quality parameters, which were standardized to a range 0 to 1. Each component plane shows the standardized value of each parameter of 40 reference vectors (nodes) using a color bar as code. A comparison between the component planes, by means of a color gradient, can indicate informative and qualitative

relationship (or correlation) among the studied parameters. For example, in **Fig. 2**, Mg^{2+}, Na^{+}, K^{+} and Cl^{-} have similar color gradients in their component planes. This means that there is strong positive correlation among these four parameters. Oppositely, CO_3^{2-} shows a negative correlation with those parameters by the inverse color gradients of the component planes.

In order to select the optimal number of clusters, the DBI values based on the k-means clustering algorithm were calculated for the possible minimum number of clusters (i.e. 2 clusters) to the maximum (40 clusters). **Fig. 3** shows the variation of DBI values after being applied to the data. The most appropriate number of clusters corresponding to the minimum DBI is six. After selecting the best clustering technique, the hierarchical clustering algorithm by Ward's method was carried out for the six clusters to fine-tune the pattern classification.

Fig. 4 shows the hierarchical cluster tree of the SOM nodes. Based on this figure, the nodes are classified into six different clusters. **Fig. 5** shows the pattern classification map of the six clusters, in which the numbers of data classified into each node are also given. The results of the component planes (**Fig. 2**) and the pattern classification (**Fig. 5**) indicate the kind of data the respective clusters include. For example, cluster-1 (upper left part of **Fig. 5**) is associated with high water salinity characterized by high Na^{+}, K^{+}, Mg^{2+}, Cl^{-}. This pattern is observed in the same location of the respective component planes as shown in **Fig. 2**. On the other hand, the groundwater samples in nodes with extremely low concentrations of Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, SO_4^{2-}, Cl^{-}, and relative low concentrations of HCO_3^{-} and CO_3^{2-}, are located at the lower left part of each component plane (associated with cluster-5), as shown in **Fig. 2**.

Fig. 6 displays the radar charts of the six clusters with eight parameters. The plots on the charts show the first quartile, median and third quartile of the reference vectors of each parameter. Based on this figure, Cluster-1 shows high values of Na^{+}, K^{+}, Mg^{2+}, Cl^{-}, with extremely low values for HCO_3^{-} and CO_3^{2-}. Cluster 2 is characterized by significantly high Ca^{2+}, SO_4^{2-} and relatively high Mg^{2+}, Na^{+}, Cl^{-}. Both Clusters-1 and -2 are characterized by the high concentrations of the major ions leading to high salinities. Cluster-3 is associated with moderate concentrations of most of the major ions. Clusters-4 and -5 have the lowest concentrations in all the major ions, which can be assumed that wells in these clusters are of freshwater groundwater types. Cluster-4 thus, is similar to Cluster-5 in terms of water type, however, Cluster-4 exhibits higher concentrations of all major ions than Cluster-5.

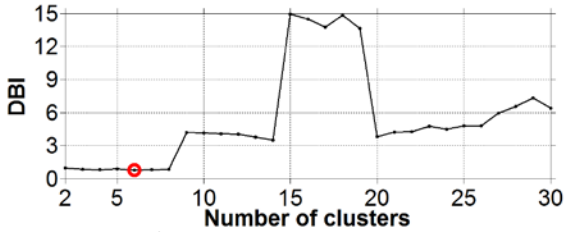

Fig.3 Variation of DBI values with the optimal number of clusters marked by the circle on the figure.

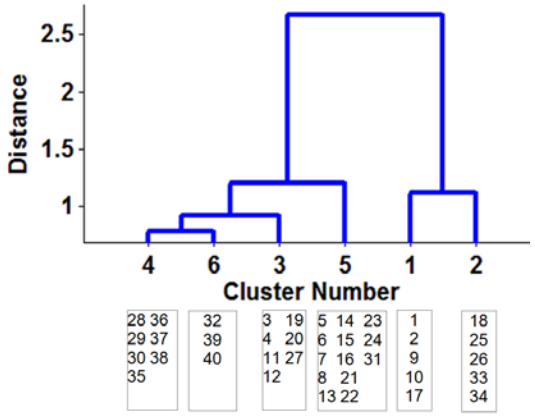

Fig.4 Dendrogram with node numbers classified into the respective clusters.

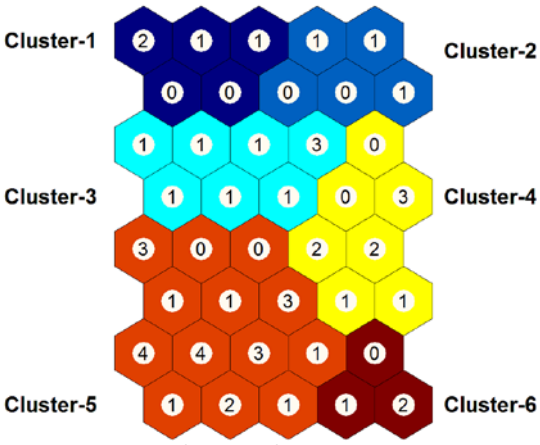

Fig.5 Pattern classification of the six clusters by the SOM.

Cluster-6 shows a pattern at which CO_3^{2-} has the highest value, while Na^{+}, K^{+}, Mg^{2+}, Ca^{2+}, Cl^{-} are very low.

The classified six clusters could be divided into two main water types. Clusters-3, -4 , -5 and -6 can also be associated to low salinity water type (fresh water type) due to the low values in all parameters as indicated in the left part of **Fig. 4** and lower part of **Fig. 5**. The other group included clusters-1 and 2, representing high salinity water type as seen in the right part of **Fig. 4** corresponding to the upper part of **Fig. 5**.

Fig. 7 shows the Gibbs diagrams for the six clusters defined by the SOM. The weight ratios Na/(Na+Ca) and Cl/(Cl+HCO3) were plotted against the total dissolved solids (TDS) separately on a logarithmic axis to represent the Gibbs cation and anion diagrams, respectively. Clusters-1 and-2 show very high TDS and high weight ratios (Na/Ca+Na

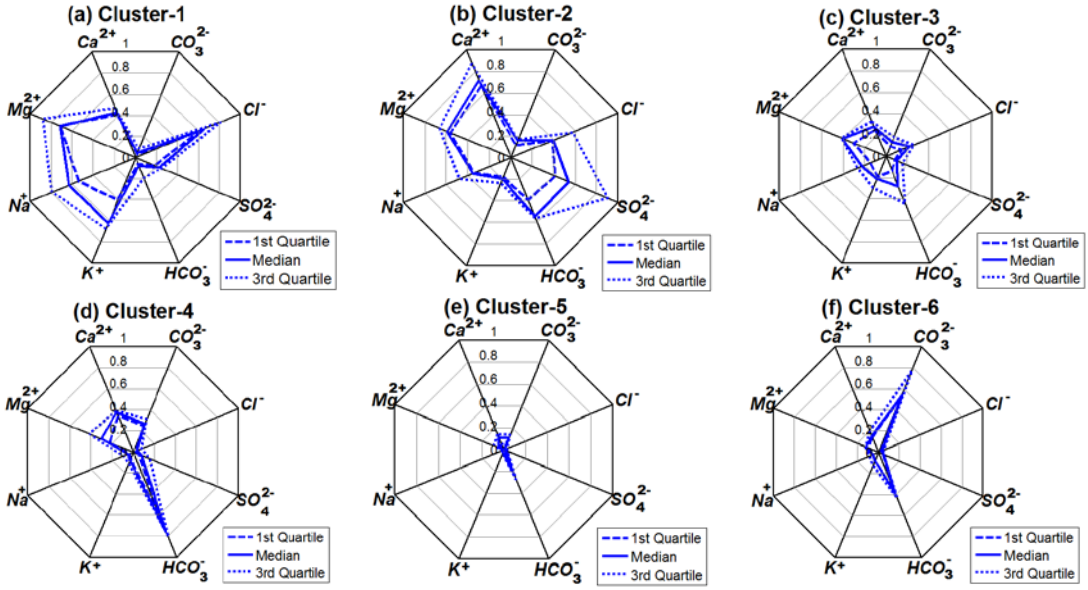

Fig.6 Radar charts for the respective clusters with the first quartile (dashed lines), median (solid lines) and the third quartile (dotted lines) by obtained reference vectors.

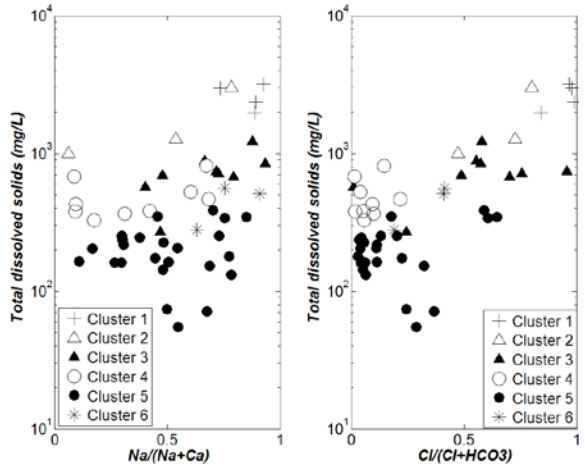

Fig.7 Gibbs diagrams for the classified data into the respective cluster.

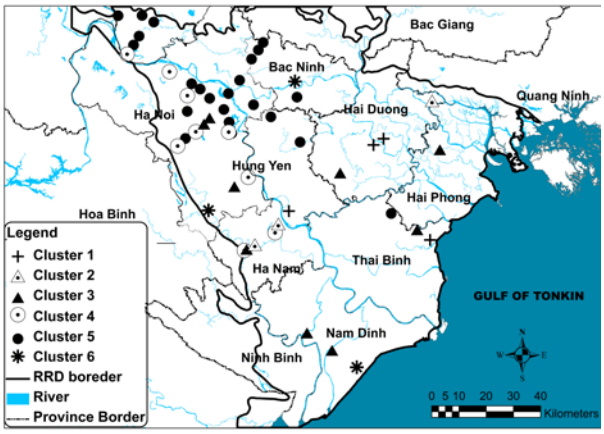

Fig.8 Distribution of the respective clusters.

and Cl/Cl+HCO_3). This is apparently due to elevated TDS, sodium and chloride concentrations arising from salt water intrusion. Cluster-3 shows relative high TDS, Na/Ca+Na and Cl/Cl+HCO3, which may be due to anthropogenic and marine activities. Clusters-4 and -5 were found to have very low TDS and relatively low Na/Ca+Na and Cl/Cl+HCO3. This suggests that Clusters-4 and -5 are dominated by the processes of mineral dissolution. This explains why the water types of these clusters are the freshest in the area. Cluster-6 shows relatively high Na/Ca+Na but relative low Cl/Cl+HCO3 and TDS, which suggests that rock-water interaction is the major source of dissolved ions.

Fig. 8 shows the distribution of the six clusters in the RRD. As shown in **Fig. 8**, all samples belonging to Clusters-1 and -2, representing the high salinity groundwater type, are located in the downstream area of the RRD. This suggests that salt water intrusion affects the groundwater quality in the Pleistocene confined aquifer up to a distance of at least 60km from the coastal line. It is also observed that almost all of the groundwater samples of Clusters-4 and -5 are distributed in the upstream area of the RRD. This implies that groundwater in the area is typically of low salinity water (freshwater) type. Comparing the Na^+ and Cl^- median values of Cluster-3 with those of the other Clusters in Fig.6, Cluster-3 is typical of an intermediate saline groundwater type, in which groundwater samples are found dispersedly in both the upstream and downstream areas. With the closer inspection of land use, these sampling points are located in agricultural area of intensive irrigation or high density of population[15]. This suggests that there are localized influences of domestic waste discharge and agricultural activities on the hydrochemistry of groundwater in the area. Cluster-6, which is characteristic of high carbonate, is distributed near the boundary of the RRD. The delta is surrounded by carbonate rock formations consisting of marble, limestone and dolomite[16]. This suggests that dissolution of theses minerals will add amounts of

CO_3^{2-} to the groundwater near the RRD boundaries by recharge from the mountains.

5. CONCLUSION

In this study, groundwater chemistry monitoring data from 52 sampling wells, obtained during the rainy season in 2011, were classified using the Self-Organizing Map (SOM) in combination with a hierarchical cluster analysis in order to investigate the spatial hydrogeochemical characteristics of groundwater in the Pleistocene confined aquifer of the RRD. The SOM has provided the readily understandable and visualized results for classifying the groundwater chemistry monitoring data into exclusively distinguishable hydrogeochemical types. The first, second and third quartiles of the reference vectors were plotted on radar charts to display the fundamental characteristics of each cluster. In addition, Gibbs diagrams and the distribution map of the respective clusters were also created in order to elucidate the spatial variability of the hydrogeochemical characteristics determined using the SOM.

From the results of the SOM application, the groundwater chemistry data could be divided into six clusters, which basically reveal two representative water types characterized by the high salinity (Clusters-1 and -2) and low salinity (fresh) water types (Clusters-3, -4, -5 and -6). The distribution of Clusters-1 and -2 in the RRD implies that salt water intrusion affects the Pleistocene groundwater quality up to a distance of at least 60km from the coastal line. The results of the Gibbs diagrams suggest that rock weathering is the main process in the evolution of chemical composition of groundwater in the upstream area of the delta, whereas salt water intrusion is the main factor affecting the groundwater chemistry in the downstream area. In addition, domestic waste discharge and agricultural activities could be the reason for increasing salinity of groundwater in some places in the RRD.

ACKNOWLEDGEMENT: This study was carried out as a part of the research project, "Solutions for the water related problems in Asian Metropolitan areas" supported by the Tokyo Metropolitan Government, Japan. Field data were provided by the project, "National Hydrogeological Database Project", financed by the Ministry of Natural Resources and Environment of Vietnam.

REFERENCES

1) Luu, T.N.M et al.: Hydrological regime and water budget of the Red River Delta (Northern Vietnam), *Journal of Asian Earth Sciences*, Vol.37, No.3, pp. 219-228, 2010.
2) Trinh, T.N. and Fredlund, D.G.: Modeling subsidence in the Hanoi City area, Vietnam, *Canadian Geotechincal Journal*, Vol.37, pp.621-637, 2000.
3) Duong, H.A. et al.: Trihalomethane formation by chlorination of ammonium- and bromide- containing groundwater in water supplies of Hanoi, Vietnam, *Water Resources*, Vol.37, No.13, pp.3242-3252, 2003.
4) Berg, M. et al.: Hydrological and sedimentary controls leading to arsenic contamination of groundwater in the Hanoi area, Vietnam: the impact of iron-arsenic ratios, peat, river bank deposits, and excessive groundwater abstraction, *Chemical geology*, Vol.249, pp.91-112, 2008.
5) Edmunds, W.M., Bath, A.H. and Miles, D.L.: Hydrochemical evolution of the East Midlands Triassic sandstone aquifers, England, *Geochimica et Cosmochinica Acta*, Vol.46, No.11, pp.2069-2081, 1982.
6) Jin, Y.H., Kawamura, A., Park, S.C., Nakagawa, N., Amaguchi, H. and Olsson, J.: Spatiotemoral classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps, *Journal of Environmental Monitoring*, Vol.13, pp.2886-2894, 2011.
7) Hentati, A., Kawamura, A., Amaguchi, H. and Iseri, Y.: Evaluation of sedimentation vulnerability at small hillside reservoirs in the semi-arid region of Tunisia using the Self-Organizing Map, *Geomorphology*, Vol.122, pp.56-64, 2010.
8) Gibbs, R.J.: Mechanisms controlling world water chemistry. *Science*, Vol. 17, pp. 1088-1090, 1970.
9) Bui, D.D., Kawamura, A., Tong. T.N., Amaguchi, H., Nakagawa, N. and Iseri, Y.: Identification of aquifer system in the whole Red River Delta, Vietnam, *Geosciences Journal, Springer*, Vol.15, No.3, pp.323-338, 2011.
10) Kohonen T.: *Self-Organizing Maps*, 3rd, Springer, 2001.
11) Jeong, K.S. et al.: Stream modification patterns in a river basin: Field survey and self-organizing map (SOM) application, *Ecological Informatics*, Vol.5, pp.293-303, 2010.
12) Vesanto, J., Himberg, J., Alhoniemi, E. and Parahankangas, J.: *SOM toolbox for Matlab 5*, Helsinki Univeriity Report A57, 2000.
13) Hilario, L.G. and Ivan, M.G.: Self-organizing map and clustering for wastewater treatment monitoring, *Engineering Applications of Artificial Intelligence*, Vol.17, pp.215-225, 2004.
14) Hentati, A., Kawamura, A., Amaguchi, H. and Nakagawa, N.: Erosion assessment at small hillside river basisns in semiarid region of Tunisia, *Annual Journal of Hydraulic Engineering*, Vol.54, pp.145-150, 2010.
15) Asian Development Bank: Final Report on Management Study on Land Use and Water Management, Red River Basin Water Resources Management Project, 2000.
16) Drogue, C., Cat, N.N. and Dazy, J.: Geological factors affecting the chemical characteristics of the thermal waters of the carbonate karstified aquifers of Northern Vietnam, *Hydrology and Earth System Sciences*, Vol.4, No.2, pp.332-340, 2000.

(Received September 30, 2013)