Volume 45 Number 1 2014
ISSN 0029-1277
Hydrology Research
An International Journal
NHF
British Hydrological Society
DHG
IWA PUBLISHING

Classification of groundwater chemistry in Shimabara, using self-organizing maps

Kei Nakagawa, Hiroki Amano, Akira Kawamura and Ronny Berndtsson

ABSTRACT

Shimabara City in Nagasaki Prefecture, Japan, is located on a volcanic peninsula that has abundant groundwater. Almost all public water supplies use groundwater in this region. For this reason, understanding groundwater characteristics is a pre-requisite for proper water supply management. Thus, we investigated the groundwater chemistry characteristics in Shimabara by use of self-organizing maps (SOMs). The input to SOM was concentrations of eight major groundwater chemical components, namely Cl^-, NO_3^-, SO_4^{2-}, HCO_3^-, Na^+, K^+, Mg^{2+}, and Ca^{2+} collected at 36 sampling locations. The locations constituted private and public water supply wells, springs, and a river sampled from April 2012 to May 2015. Results showed that depending on the chemistry, surface water and groundwater could be classified into five main clusters displaying unique patterns. Further, the five clusters could be divided into two major water types, namely, nitrate- and non-polluted water. According to Stiff and Piper trilinear diagrams, the nitrate-polluted water represented Ca-($SO_4 + NO_3$) (calcium sulfate nitrate) type, while the non-polluted water was classified as Ca-HCO_3 (calcium bicarbonate) type. This indicates that recharging rain water in the upstream areas is polluted by agricultural activities in the mid-slope areas of Shimabara.

Key words | cluster analysis, groundwater, major ions, self-organizing map, water chemistry

Kei Nakagawa (corresponding author)
Hiroki Amano
Graduate School of Fisheries and Environmental Sciences,
Nagasaki University,
1-14 Bunkyo-machi,
Nagasaki 852-8521,
Japan
E-mail: *kei-naka@nagasaki-u.ac.jp*

Akira Kawamura
Graduate School of Urban Environmental Sciences,
Tokyo Metropolitan University,
1–1 Minami-Oshawa,
Hachioji,
Tokyo 192-0397,
Japan

Ronny Berndtsson
Division of Water Resources Engineering & Center for Middle Eastern Studies,
Lund University,
Box 118 SE-221 00,
Lund,
Sweden

INTRODUCTION

Groundwater is used for various purposes, such as water supply, agriculture, and industry. During recent decades, groundwater has been polluted by increasing fertilizer applications to meet the demand of food supply due to population growth. Monitoring and protection of groundwater are essential to meet the demand for safe groundwater. To understand the effects of hydrogeological processes and anthropogenic activities on regional groundwater, it is important to study the chemical characteristics. The hydrogeochemistry of groundwater is influenced by many factors, such as climate, mineralogy of aquifers, chemical composition of rainfall and surface water, topography, and anthropogenic activities. Thus, a hydrogeochemical interpretation of groundwater quality from representative water samples can provide useful information on the geochemical processes, hydrodynamics, origin, and interaction of the groundwater with aquifer materials.

Shimabara City is known as a region that, to a great extent, relies on groundwater for the public water supply (Committee on Nitrate Reduction in Shimabara Peninsula 2011). However, Shimabara groundwater has been increasingly polluted by nitrate since 1988. We analyzed the present situation of groundwater pollution by nitrate in Shimabara and showed that agricultural activities are the main polluter of the groundwater (Nakagawa *et al.* 2016). To better understand the characteristics of the water chemistry, multivariate analysis such as principal component analysis (PCA), which can reduce data dimensionality and extract synthetic indexes with minimum information loss, is often used (e.g., Aiuppa *et al.* 2003; Cloutier *et al.*

doi: 10.2166/nh.2016.072

2008; Banoeng-Yakubo *et al.* 2009; Sonkamble *et al.* 2012; Nadiri *et al.* 2013; Omonona *et al.* 2014; Singaraja *et al.* 2014; Ghesquière *et al.* 2015; Marghade *et al.* 2015; Matiatos 2016). Using groundwater chemistry, we classified Shimabara water by use of principal component and cluster analysis (Nakagawa *et al.* 2016). The results showed that groundwater could be classified into four clusters, where one cluster expressed nitrate pollution and the other clusters showed ion dissolution from the aquifer matrix. However, it is sometimes difficult to decipher PCA results due to bias resulting from the complexity and nonlinearity of large data (Choi *et al.* 2014). Recently, multivariate analysis using self-organizing maps (SOMs) has been applied to various research fields, such as ecology (Céréghino *et al.* 2001; Bedoya *et al.* 2009), geomorphology (Hentati *et al.* 2010), hydrology (Kalteh & Berndtsson 2007), meteorology (Nishiyama *et al.* 2007), and wastewater treatment (García & González 2004). SOM has also been used to classify the water chemistry of rivers and groundwater (Hong & Rosen 2001; Jin *et al.* 2011; Choi *et al.* 2014; Nguyen *et al.* 2015). Thus, SOM is a powerful and effective tool for detection and interpretation of spatially varying phenomena. Especially, SOM has a better ability to handle the nonlinearities, noisy or irregular data, and multivariate data without mechanistic understanding of the system. SOM is also easily and quickly updated when adding new data (Hong & Rosen 2001; Kalteh *et al.* 2008). The similarity of extracted pattern classification can be visually compared using color gradients (Jin *et al.* 2011).

In the previous study (Nakagawa *et al.* 2016), we used field observed data from August 2011 to November 2013. We continued to collect data, and available data were extended to May 2015. Therefore, in this study, we confirmed our previous results by using a more informative method, SOM, together with an extended database. Using SOM, visual representation of groundwater characteristics is easy, and more detailed clustering with better analyses results is possible as compared to conventional PCA. To improve the understanding of groundwater characteristics in Shimabara we applied SOM combined with hierarchical cluster analysis using water chemistry as input. According to the results obtained by SOM analysis, we discuss the spatial trends of groundwater characteristics in Shimabara and the practical application of SOM for future water use.

STUDY AREA AND DATA USED

Figure 1 shows the study area and the sampling locations in Shimabara, Nagasaki Prefecture, Japan. Shimabara has an area of 82.8 km^2 and is located in the northeastern part of Shimabara Peninsula. In the center of the peninsula, the active volcano Unzen (Mt Fugendake) is located. The geology of the Shimabara area is thus formed by volcanic deposits composed of dacite, andecite, volcanic ash, and lapilli. Average annual precipitation is about 2,100 mm (1967–2013). The mean annual temperature is 16.9 °C, and the average monthly temperature ranges from 4.2 (January) to 29.0 °C (in August) (Japan Meteorological Agency 2015).

Figure 2 shows altitude and land use in Shimabara. According to the figure, the land use can generally be divided into forest, agriculture, and urban areas. Areas above an altitude of 200 m are generally occupied by forest. According to the estimated regional groundwater flow, the forest areas, which comprise 36.5% of Shimabara, may be recognized as groundwater recharge zones. Upland and paddy fields are concentrated into the northern parts of the area, occupying 23.6% and 7.5% of Shimabara, respectively. Buildings are usually located at altitudes below 100 m along the coast and represent 14.9% of Shimabara. Other land use is 17.5%.

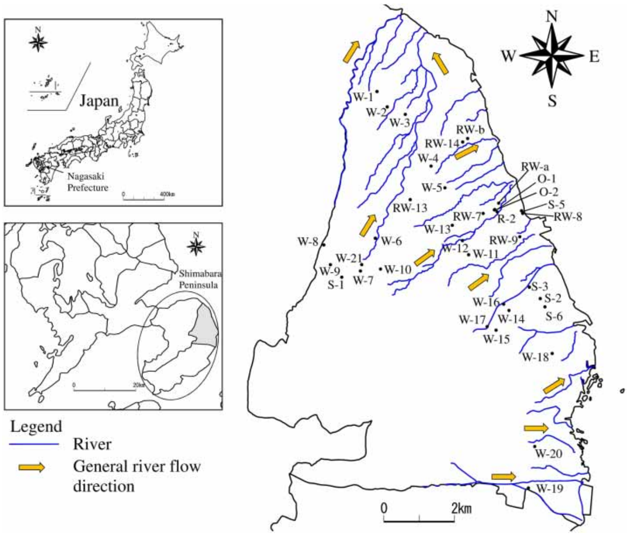

Figure 1 | Study area and sampling locations in Shimabara, Nagasaki Prefecture, Japan (RW: residential well, W: public water supply well, O: observation well, S: spring, and R: river).

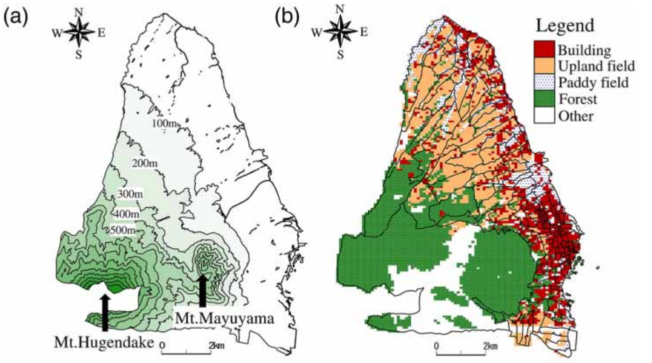

Figure 2 | Altitude and land use map of Shimabara; (a) altitude and (b) land use.

In total, 353 water samples were collected from April 2012 to May 2015. Sampling was performed at seven resident wells (RW), 21 public water supply wells (W), two observation wells (O), five springs (S), and one river (R) (Figure 1). To ensure spatially representative groundwater conditions, sampling sites covering the whole area of Shimabara except for forest and other land use (Figures 1 and 2) were used. Sampling was done four times annually with 2–4 month intervals to ensure temporally varying groundwater conditions. Sampling at specific locations (RW-14, b, W-21, O-2, S-2, 3, 5, and R-2) was done with less frequency. The hydrogeochemical data used in this study consist of major dissolved ion concentrations for Cl^-, NO_3^-, SO_4^{2-}, HCO_3^-, Na^+, K^+, Mg^{2+}, and Ca^{2+}. Mean and standard deviation of 36 sampling sites using averaged temporal ion concentrations for each of the sampling sites are summarized in Table 1. It is necessary to normalize the data prior to application of SOM to ensure that all parameters are given the same importance. SOM results are highly sensitive to data pre-processing method due to the fact that the Euclidean distance between input data is used (e.g., Jin *et al.* 2011). To solve this problem, the range between minimum and maximum ion concentrations was standardized into [0, 1] (Nishiyama *et al.* 2007; Jin *et al.* 2011) as preprocessing in this study.

Table 1 | Mean and standard deviations of 36 sampling sites using averaged temporal ion concentrations for each sampling site used in the SOM

Major ion (mg L^{-1})	Mean	SD
Cl^-	12.4	1.4
NO_3^-	38.4	5.0
SO_4^{2-}	21.9	3.2
HCO_3^-	55.7	6.6
Na^+	12.1	2.4
K^+	6.4	1.2
Mg^{2+}	8.7	1.1
Ca^{2+}	22.4	2.9

METHODOLOGY

The SOM is a modified artificial neural network characterized by unsupervised training that can project high-dimensional information onto a low-dimensional array (e.g., Vesanto *et al.* 2000). Many researchers have chosen a two-dimensional array (e.g., Jiang *et al.* 2014). The result is a readily understandable and visual pattern classification. The objective here of the SOM application was to obtain physically explainable reference vectors using input vectors. Thus, the input vectors were composed of, in total, 353 hydrogeochemical data points (approximately quarterly sampling at the 36 sampling locations) with eight variables (major dissolved ion concentrations: Cl^-, NO_3^-, SO_4^{2-}, HCO_3^-, Na^+, K^+, Mg^{2+}, and Ca^{2+}). Reference vectors were obtained after iterative updates through a training phase that comprised three main procedures: competition between nodes, selection of a winner node, and updating of the reference vectors (e.g., Vesanto *et al.* 2000). Selection of proper initialization and data transformation methods are important factors when designing a relevant SOM methodology. In SOM applications, in general, a larger map size gives a higher resolution for pattern recognition. The optimum number of SOM nodes is determined by applying the heuristic rule $m = 5\sqrt{n}$, where m denotes the number of SOM nodes and n represents the number of input data (García & González 2004; Hentati *et al.* 2010; Jin *et al.* 2011). Herein, this heuristic rule was used to determine the total number of nodes in the SOM. The ratio of the number of rows and columns is determined by the square root of the ratio between the two largest eigenvalues of the correlation matrix of input data. The eigenvalues are obtained from PCA. In a previous study using the sampled data from August 2011 to November 2013, two principal components (Factor 1 and Factor 2) explained 86.5% of the total variance (Nakagawa *et al.* 2016).

After organizing the SOM structure with the above rule, a linear initialization technique made each node set with a reference vector. A linear initialization technique increases the speed of the training phase and proper abstracting pattern for limited data (Jeong *et al.* 2010). Further, when only limited data are available, the linear initialization is more suitable for the pattern classification as compared to random initialization, because of small data sets and boundary effects (Nguyen *et al.* 2015). The linear initialization used eigenvalues and eigenvectors of input data to set initial reference vectors on the structured SOM. This means that the initial reference vectors already include prior information about the input data, resulting in a quicker and more efficient training phase (Vesanto *et al.* 2000). In this study, each reference vector was updated through the SOM training process using a batch mode with neighborhood function taking a Gaussian form. Although some issues on the implementation of the batch SOM are discussed in some detail in Jiang *et al.* (2014), the results of the SOM analysis supported previous clustering results (Nakagawa *et al.* 2016; shown below). The reference vectors obtained at the end of the training process were fine-tuned using cluster analysis.

There are various clustering algorithms available in the literature (e.g., García & González 2004; Jin *et al.* 2011). In this study, partitioned algorithms and hierarchical algorithms, which are k-means and Ward's algorithms, respectively, were applied for appropriate clustering of reference vectors. For partitioned clustering methods, the k-means algorithm is most frequently used for SOM (e.g., Jin *et al.* 2011). The Davies–Bouldin Index (DBI) applying k-means algorithm determines the optimal number of clusters (García & González 2004; Jin *et al.* 2011). The DBI values, based on similarity within a cluster and dissimilarity between clusters, were calculated from a minimum of two clusters to the total number of nodes. Therefore, the smaller DBI value appears as the dissimilarity to each cluster becomes larger. In other words, a minimum DBI represents the optimal number of clusters for the trained SOM. The Ward's linkage method, which is one of the hierarchical techniques, is the most commonly used clustering method (Faggiano *et al.* 2010; Hentai *et al.* 2010; Jin *et al.* 2011). In this study, the final fine-tuning cluster analysis was carried out using Ward's method. The above calculation processes were carried out using a modified version of SOM Toolbox 2.0 (Vesanto *et al.* 2000). The output SOM clusters were plotted on Piper trilinear and Stiff diagrams to explain the main features of each cluster. Furthermore, the SOM clusters were mapped spatially to clarify influence from land use.

RESULTS AND DISCUSSION

Based on the methodology described above, the number of SOM nodes was determined to be equal to 91. The number of rows and columns was 7 and 13, respectively. Thus, this SOM design was used for the cluster analysis of standardized water chemistry data from the 36 locations in Shimabara.

Figure 3 shows the obtained component planes for the 91 reference vectors (nodes) of the eight ion component concentrations (standardized to a range between 0 and 1). Each component plane shows the standardized value of each parameter (concentration) of the 91 reference vectors (nodes) using a color gradient. Comparison between the component planes shows relationships (or correlation) among the parameters. For example, a similar color gradient can be observed for Cl^- (Figure 3(a)) and NO_3^- (Figure 3(b)). The same trend can be seen for Na^+ (Figure 3(e)) and Mg^{2+} (Figure 3(g)) in their respective component planes. This means that there is high positive correlation between these variables. A great advantage of SOM is that relationships between nodes on the component plane are clearly visualized. For example, the node located at the uppermost left end shows lower normalized concentrations for all ions (Cl^-:0.00, NO_3^-:0.00, SO_4^{2-}:0.00, HCO_3^-:0.11, Na^+:0.00, K^+:0.00, Mg^{2+}:0.00, and Ca^{2+}:0.00). The node, located at the uppermost right end, shows moderately higher normalized concentrations for HCO_3^-, Mg^{2+}, and Ca^{2+} (Cl^-:0.13, NO_3^-:0.09, SO_4^{2-}:0.15, HCO_3^-:0.46, Na^+:0.09, K^+:0.18, Mg^{2+}:0.33, Ca^{2+}:0.40). The node located at the lowermost left shows relatively higher normalized ion concentrations except for HCO_3^- (Cl^-:0.85, NO_3^-:0.83, SO_4^{2-}:0.78, HCO_3^-:0.04, Na^+:0.30, K^+:1.00, Mg^{2+}:0.43, Ca^{2+}:0.90). On the other hand, the node located at the lowermost right shows higher normalized ion concentrations except for Cl^- and NO_3^- (Cl^-:0.29, NO_3^-:0.17, SO_4^{2-}:0.95, HCO_3^-:0.95, Na^+:1.00, K^+:0.65, Mg^{2+}:0.98, Ca^{2+}:1.00).

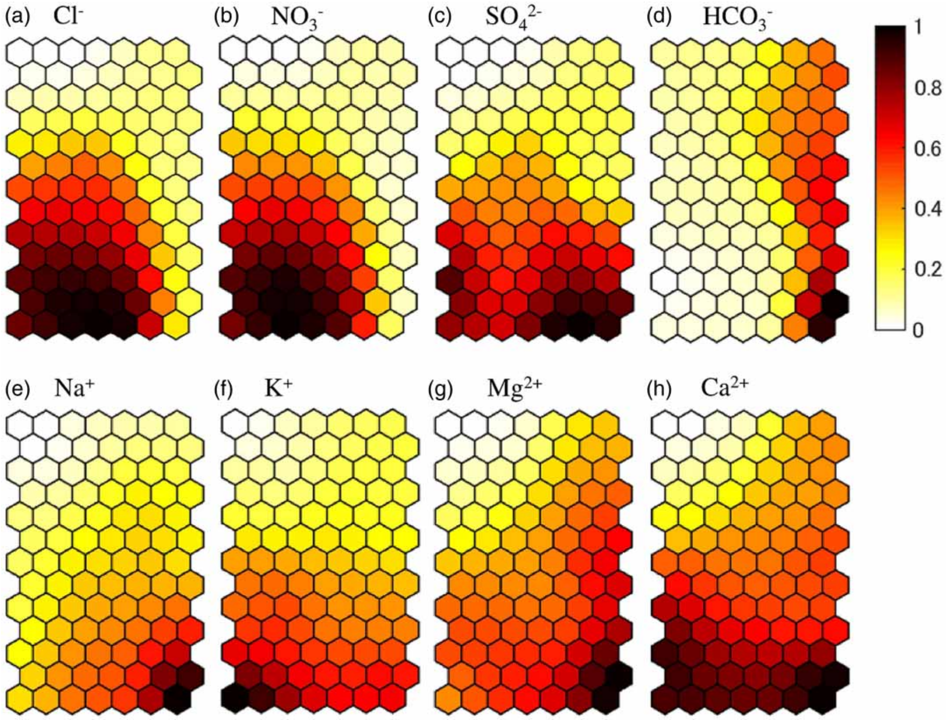

Figure 3 | Component plane for (a) Cl^-, (b) NO_3^-, (c) SO_4^{2-}, (d) HCO_3^-, (e) Na^+, (f) K^+, (g) Mg^{2+}, and (h) Ca^{2+}.

To confirm quantitative relationships, as mentioned above, correlation coefficients between reference vectors for each parameter were calculated (Table 2). There is a high correlation ($r = 0.99$) between Cl^- and NO_3^-. There is also a high correlation between Na^+ and Mg^{2+} ($r = 0.92$). Similarly, the color gradient for the relationship between SO_4^{2-} and Ca^{2+} indicates a high correlation coefficient ($r =$ 0.94). The relation between each ion indicates factors affecting groundwater chemistry. For example, a high co-variation ($R^2 = 0.72$) between higher concentrations of NO_3^- and Cl^- was observed, indicating that they originate from common sources, such as human and animal waste (e.g., Diédhiou *et al.* 2012). Moreover, the same result can be observed between SO_4^{2-} and Ca^2 ($r = 0.79$). The high correlation implies that the dissolution of gypsum may be one of the key factors controlling the geochemical evolution of groundwater (Liu *et al.* 2015).

Figure 4 shows the variation of DBI with a magnified front between 2 and 14 clusters. The minimum DBI is shown for five clusters, meaning that this number should

Table 2 | Correlation between reference vectors for each parameter

	NO_3^-	SO_4^{2-}	HCO_3^-	Na^+	K^+	Mg^{2+}	Ca^{2+}
Cl^-	0.99*	0.82*	−0.51*	0.47*	0.86*	0.46*	0.78*
NO_3^-		0.75*	−0.60*	0.38*	0.82*	0.36*	0.71*
SO_4^{2-}			−0.03	0.84*	0.92*	0.79*	0.94*
HCO_3^-				0.43*	−0.11	0.52*	0.11
Na^+					0.71*	0.92*	0.82*
K^+						0.72*	0.94*
Mg^{2+}							0.88*

*Correlations significant at $p = 0.01$.

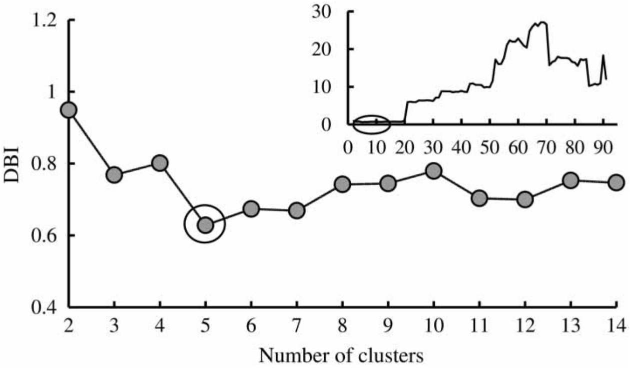

Figure 4 | Variation of DBI values with the optimal number of clusters marked by the circle on the figure.

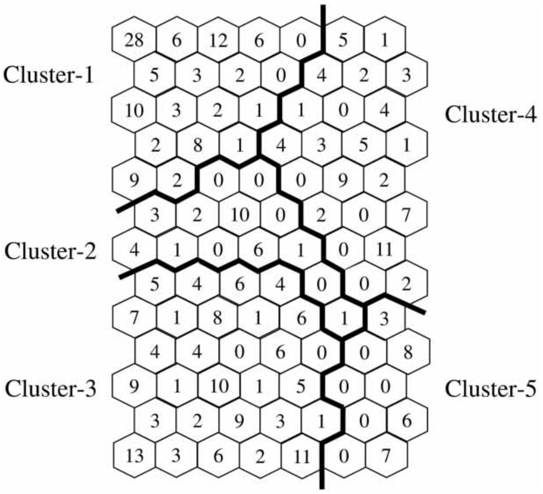

Figure 6 | Pattern classification map of the five clusters by the SOM. The numbers on the hexagons of the map represent the number of data classified into each node.

be used as an optimal value. After determining the number of clusters, the hierarchical clustering algorithm by Ward was carried out for the five clusters to fine-tune pattern classification. Figure 5 shows the hierarchical cluster dendrogram. The 91 nodes of the SOM were classified into five different clusters. Figure 6 shows the pattern classification map for these five clusters. The number for each node represents the raw data classified into each node. Simultaneous analysis of the component planes (Figure 3) and the pattern classification result (Figure 6) indicates what kind of data the respective clusters include. For example, cluster-3 (the lower left part of Figure 6) is associated with a high content of Cl^- and NO_3^-. This pattern is observed in the same part of the respective component planes for each parameter, as shown in Figure 3. On the other hand, groundwater samples in nodes with an extremely low concentration of all ions are located at the upper left part of each component plane (associated with cluster-1), as shown in Figure 3.

More quantitative information than the visualized pattern classification can be extracted and interpreted from the obtained reference vectors. Stiff diagrams for the respective clusters were represented by mean and upper and lower limits of one standard deviation using reference vectors of each cluster to characterize the clustered data. For example, the Stiff diagram for cluster-1 is represented by reference vectors of 18 nodes classified into the cluster. Figure 7 shows Stiff diagrams for the five clusters, with eight parameters containing mean values and standard deviations. Cluster-1 (Figure 7(a)) shows low values for all ions compared to other clusters. The visible patterns of cluster-2 (Figure 7(b)) and cluster-3 (Figure 7(c)) are not similar, as shown in the figure. However, they are characterized by high concentrations of NO_3^-. Cluster-2 represents lower concentrations than that of cluster-3 for all ions except HCO_3^-. The pattern with the highest Ca^{2+} in cations and HCO_3^- in anions is associated with cluster-4 (Figure 7(d)). In this cluster, the concentration of Na^+, K^+, and Mg^{2+} is slightly lower than that for Ca^{2+}. For anions, the concentration of HCO_3^- is significantly higher than other anions. This pattern is also shown in cluster-5 (Figure 7(e)). It is clear that all ion

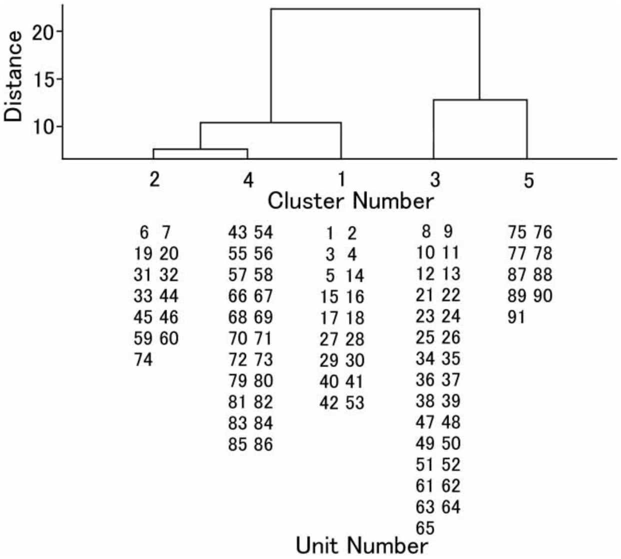

Figure 5 | Dendrogram with node number classified into clusters.

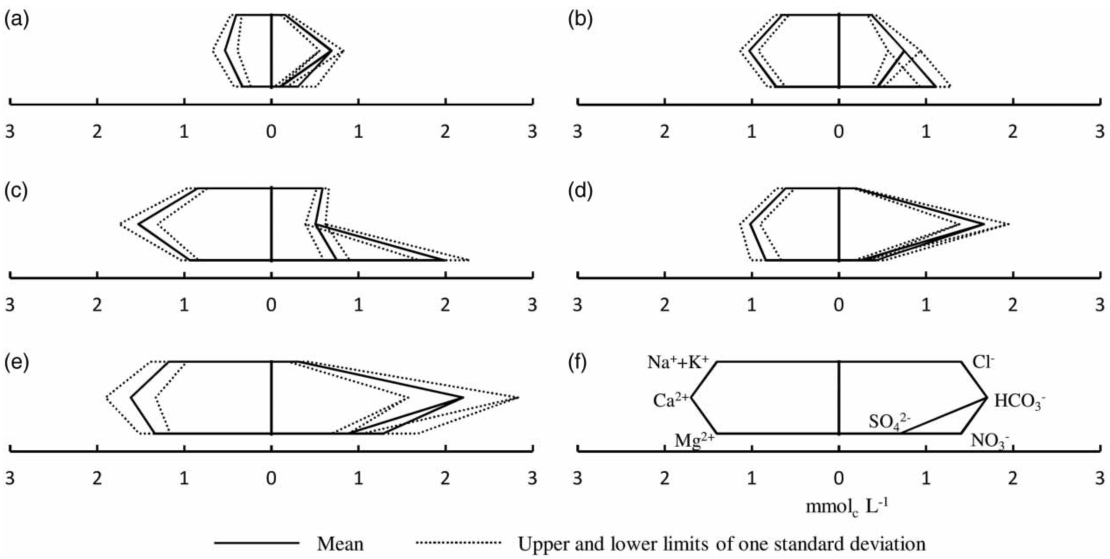

Figure 7 | Stiff diagrams for the respective clusters with mean value and upper and lower limits of one standard deviation by obtained reference vectors: (a) cluster-1, (b) cluster-2, (c) cluster-3, (d) cluster-4, (e) cluster-5, and (f) legend.

concentrations except for Cl^- and NO_3^- of cluster-5 are higher than that of cluster-4.

The five classified clusters can generally be divided into two water quality types. Cluster-2 and -3 can be characterized as polluted water due to the high concentration of NO_3^-. The other group includes cluster-1, -4, and -5, representing non-polluted water (pristine water type).

Table 3 shows mean ion concentrations calculated from raw data and classified into the respective cluster. The NO_3^- for cluster-3 indicates a higher mean value than 50 mg L^{-1} which is the maximum contamination level recommended by the World Health Organization (WHO 2011) for drinking water. The NO_3^- for cluster-2 meets the WHO standard. However, it exceeds 13 mg L^{-1} which is the maximum nitrate concentration unaffected by human activities (Eckhardt & Stackelberg 1995). It confirms that the two clusters include polluted water as mentioned above. Cluster-1, -4, and -5 display much lower mean NO_3^- concentrations. An NO_3^- concentration exceeding the maximum concentration level recommended by the WHO has also been reported in other studies (e.g., Diédhiou *et al.* 2012; Hansen *et al.* 2012; Liu *et al.* 2015; Dragon *et al.* 2016; Matiatos 2016). In these investigations, the maximum NO_3^- concentration ranged from 91 to 855 mg L^{-1}.

Figure 8 shows Piper trilinear diagrams for all reference vectors (91) and the respective cluster. With respect to cations, most vectors of all clusters are located in zone B in the lower left delta-shaped region, indicating a non-typical water. However, a part of the reference vectors for cluster-3 is located in zone A, indicating a calcium-type water. For anions, reference vectors are mostly located in zone B, E, and F in the lower right delta-shaped region, suggesting that the reference vectors of cluster-1, -4, and -5 are bicarbonate-type water and the reference vectors of cluster-2 and -3 are

Table 3 | Mean ion concentrations calculated from raw data and classified into clusters

	Cl^- (mg L^{-1})	NO_3^- (mg L^{-1})	SO_4^{2-} (mg L^{-1})	HCO_3^- (mg L^{-1})	Na^+ (mg L^{-1})	K^+ (mg L^{-1})	Mg^{2+} (mg L^{-1})	Ca^{2+} (mg L^{-1})
Cluster-1	5.1	9.9	3.2	37.7	6.5	3.4	3.2	8.7
Cluster-2	14.3	42.1	22.5	39.0	11.2	6.2	8.1	20.5
Cluster-3	21.3	78.8	37.7	27.5	14.4	8.6	11.2	31.5
Cluster-4	6.4	9.9	10.5	108.5	11.1	4.9	10.6	21.0
Cluster-5	6.8	6.2	41.3	175.4	25.1	7.9	17.6	33.5

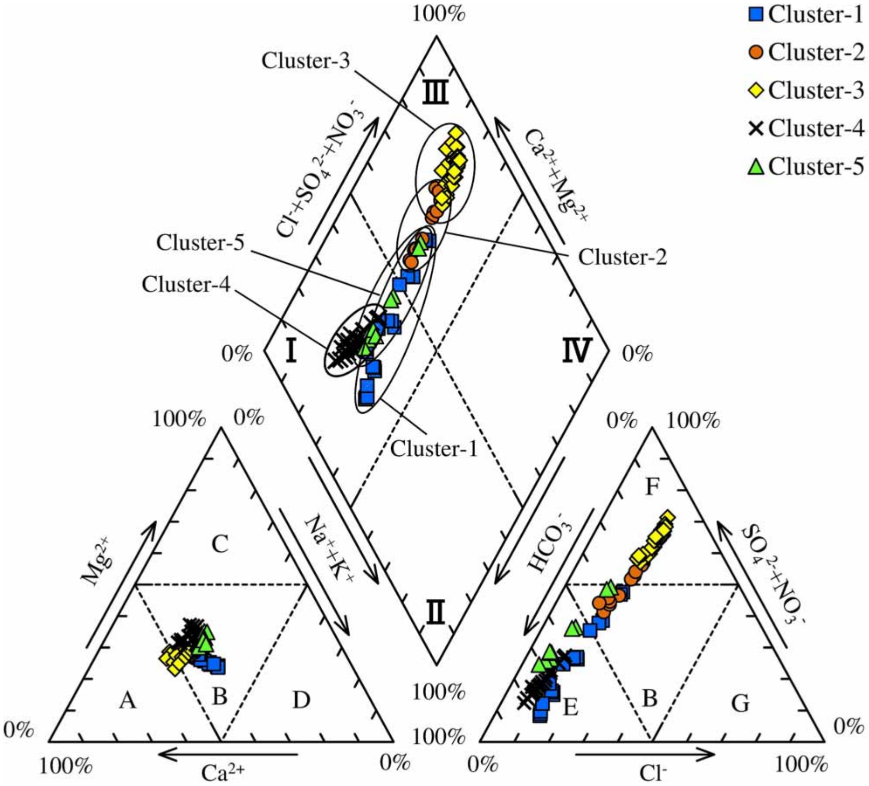

Figure 8 | Trilinear diagram for clusters obtained by reference vectors.

sulfate and nitrate-type water or non-typical water. Thus, in the Piper trilinear diagram, two main water types are revealed. These are calcium-magnesium bicarbonate type (zone I) including cluster-1, 4, and 5 (non-polluted water type) and calcium-magnesium chloride-sulfate-nitrate type (zone III) including cluster-2 and -3 (polluted water type).

Based on the Stiff and Piper trilinear diagrams, the polluted water type is represented as Ca-($SO_4 + NO_3$) (calcium sulfate nitrate type), while the non-polluted water type is classified as Ca-HCO_3 (calcium bicarbonate type). Similar results were reported by Shin *et al.* (2013). According to the study, water samples collected from the upper reaches of Korean rivers were of Ca-HCO_3 type, whereas water samples collected from lower reaches and with relative high nitrate concentrations were classified as Na-Cl-NO_3 type. This indicates that water samples are affected by anthropogenic factors such as fertilizer, manure, and septic waste.

Figure 9 shows the spatial distribution of the five clusters in Shimabara. All sampling locations belonging to cluster-2 and -3, representing the polluted water type, are located in the northern part of Shimabara encompassing a concentration of agricultural fields. In order to investigate the interaction between groundwater and river water, one

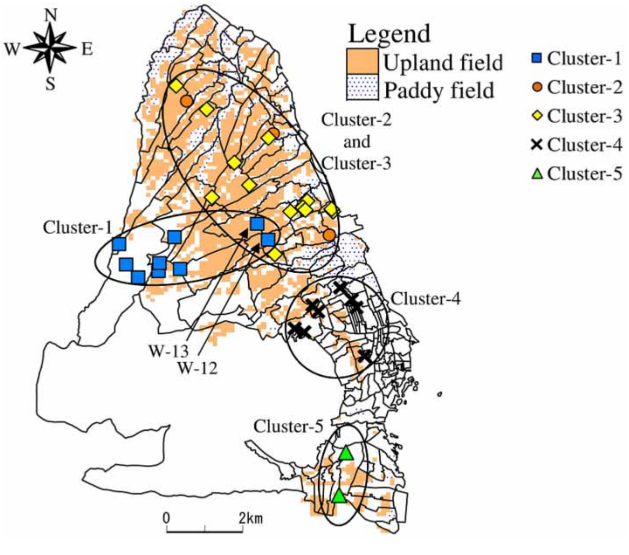

Figure 9 | Spatial distribution of clusters.

sample was taken from the river (R-2) and included into the SOM analysis. The results showed that R-2 also is classified into cluster-3 as O-1 and 2. This revealed that they are connected and exchange water with each other. Samples with high nitrate concentrations often correspond with agricultural land use (Babiker *et al.* 2004; Esmaeili *et al.* 2014). This confirms that agricultural activities are related to high nitrate concentrations in groundwater. Ishihara *et al.* (2002) reported that fecal coliforms were detected in the northern part of Shimabara. This means that the groundwater in this area is affected by livestock waste. It is observed that most sampling locations for cluster-1 are distributed in the mountainside forest area upstream of the heavily polluted areas. This shows that groundwater is recharged in the area and typically is of pristine water type. The average NO_3^- concentration of cluster-1 is slightly lower than that of cluster-4 according to Table 3. Sampling points such as W-12 and 13 located in the agricultural area are thus affected by agricultural activities belonging to cluster-1. This suggests that cluster-1 shows a transition of water chemistry from pristine to polluted water type. The sampling locations for cluster-4 and -5, characterized by high ion concentrations, are located in the urban area at a lower altitude (below 100 m). This suggests that dissolution of ions from the aquifer matrix during groundwater flow from the mountainside to the urbanized area may increase ion concentrations. Mayuyama avalanche debris deposits are distributed in the eastern area of Mt Mayuyama (Ozeki *et al.* 2005). This area corresponds to sampling locations for cluster-5. The pattern of cluster-5 has high concentration for all ions, as shown in Figure 7. This is due to the effect of volcanic deposits on the groundwater chemistry in the area.

SUMMARY AND CONCLUSION

In this study, water chemistry data from 36 sampling locations, obtained from April 2012 to May 2015, were classified using SOM in combination with hierarchical cluster analysis to clarify groundwater characteristics in Shimabara, Japan. The SOM provided readily understandable results for classifying the water chemistry data into distinguishable hydrogeochemical types. The Piper trilinear and Stiff diagrams for the reference vectors were plotted to display fundamental characteristics of each cluster. In addition, the spatial distribution of the respective clusters explained the spatial variability of the hydrogeochemical characteristics determined by the SOM. Based on the SOM results, the water chemistry data could be divided into five clusters that revealed two representative water types characterized by nitrate pollution (cluster-2 and -3) and non-polluted (cluster-1, -4, and -5) water. The spatial distribution of cluster-2 and -3 shows that agricultural activities are causing groundwater pollution in the northern part of Shimabara. The Stiff and Piper trilinear diagrams based on the reference vectors for each cluster showed that non-polluted water and polluted water are characterized by Ca-HCO_3 type and Ca-($SO_4 + NO_3$) type, respectively. This indicates that nitrate pollution is a product from agricultural activities and classified into cluster-2 and -3.

The SOM analysis showed that mountainside recharged pristine groundwater is classified into cluster-1. Some groundwater in cluster-1 is also located close to the mid-slope hills. This means that non-polluted water can be used from this agricultural area. For other purposes, water quality evaluation methods such as the Wilcox classification diagram (Wilcox 1955), can be used to evaluate whether water in cluster-2 or -3 can be used for, e.g., irrigation. The clusters from the SOM analysis are useful for further groundwater remediation alternatives.

The application and results of the SOM support our previous conclusion (Nakagawa *et al.* 2016) regarding the spatial distribution of nitrate pollution in the study area and its causes. Data that display a scattered distribution in the Piper trilinear diagram can be difficult to analyze by PCA. However, in this case, SOM can be an alternative method (Choi *et al.* 2014). In this study, both PCA and SOM successfully classified groundwater chemistry in the study area. However, SOM gives more robust and explainable results that can be used to characterize groundwater chemistry. More detailed characteristics along this line will be described in a new paper (Amano *et al.* in press).

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 24360194 and 15KT0120.

REFERENCES

Aiuppa, A., Bellomo, S., Brusca, L., D'Alessandro, W. & Federico, C. 2003 Natural and anthropogenic factors affecting groundwater quality of an active volcano (Mt. Etna, Italy). *Applied Geochemistry* **18** (6), 863–882.

Amano, H., Nakagawa, K. & Kawamura, A. Classification characteristics of multivariate analyses for groundwater chemistry in the nitrate contaminated area. In press (in Japanese with English abstract).

Babiker, I. S., Mohamed, M. A. A., Terao, H., Kato, K. & Ohta, K. 2004 Assessment of groundwater contamination by nitrate leaching from intensive vegetable cultivation using geographical information system. *Environment International* **29** (8), 1009–1017.

Banoeng-Yakubo, B., Yidana, S. M. & Nti, E. 2009 Hydrochemical analysis of groundwater using multivariate statistical methods–The Volta Region, Ghana. *KSCE Journal of Civil Engineering* **13** (1), 55–63.

Bedoya, D., Novotny, V. & Manolakos, E. S. 2009 Instream and offstream environmental conditions and stream biotic integrity importance of scale and site similarities for learning and prediction. *Ecological Modelling* **220** (19), 2393–2406.

Céréghino, R., Giraudel, J. L. & Compin, A. 2001 Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. *Ecological Modelling* **146** (1–3), 167–180.

Choi, B. Y., Yun, S. T., Kim, K. H., Kim, J. W., Kim, H. M. & Koh, Y. K. 2014 Hydrogeochemical interpretation of South Korean groundwater monitoring data using self-organizing maps. *Journal of Geochemical Exploration* **137**, 73–84.

Cloutier, V., Lefebvre, R., Therrien, R. & Savard, M. M. 2008 Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. *Journal of Hydrology* **353** (3–4), 294–313.

Committee on Nitrate Reduction in Shimabara Peninsula 2011 The second term of Shimabara peninsula nitrate load reduction project (in Japanese).

Diédhiou, M., Cissé Faye, S., Diouf, O. C., Faye, S., Faye, A., Re, V., Wohnlich, S., Wisotzky, F., Schulte, U. & Maloszewski, P. 2012 Tracing groundwater nitrate sources in the Dakar suburban area: an isotopic multi-tracer approach. *Hydrological Processes* **26** (5), 760–770.

Dragon, K., Kasztelan, D., Gorski, J. & Najman, J. 2016 Influence of subsurface drainage systems on nitrate pollution of water supply aquifer (Tursko well-field, Poland). *Environmental Earth Sciences* **75**, 100.

Eckhardt, D. A. V. & Stackelberg, P. E. 1995 Relation of ground-water quality to land use on Long Island, New York. *Groundwater* **33** (6), 1019–1033.

Esmaeili, A., Moore, F. & Keshavarzi, B. 2014 Nitrate contamination in irrigation groundwater, Isfahan, Iran. *Environmental Earth Sciences* **72** (7), 2511–2522.

Faggiano, L., Zwart, D., García-Berthou, E., Lek, S. & Gevrey, M. 2010 Patterning ecological risk of pesticide contamination at the river basin scale. *Science of the Total Environment* **408** (11), 2319–2326.

García, H. L. & González, I. M. 2004 Self-organizing map and clustering for wastewater treatment monitoring. *Engineering Applications of Artificial Intelligence* **17** (3), 215–225.

Ghesquière, O., Walter, J., Chesnaux, R. & Rouleau, A. 2015 Scenarios of groundwater chemical evolution in a region of the Canadian Shield based on multivariate statistical analysis. *Journal of Hydrology: Regional Studies* **4** (B), 246–266.

Hansen, B., Dalgaard, T., Thorling, L., Sørensen, B. & Erlandsen, M. 2012 Regional analysis of groundwater nitrate concentrations and trends in Denmark in regard to agricultural influence. *Biogeosciences* **9**, 3277–3286.

Hentati, A., Kawamura, A., Amaguchi, H. & Iseri, Y. 2010 Evaluation of sedimentation vulnerability at small hillside reservoirs in the semi-arid region of Tunisia using the self-organizing map. *Geomorphology* **122** (1–2), 56–64.

Hong, Y. S. & Rosen, M. R. 2001 Intelligent characterisation and diagnosis of the groundwater quality in an urban fractured-rock aquifer using an artificial neural network. *Urban Water* **3** (3), 193–204.

Ishihara, T., Ura, N. & Hamabe, M. 2002 Investigation of ground water contaminated by nitrate-nitrogen. *Annual Report of Nagasaki Prefectural Institute of Public Health and Environmental Sciences* **48**, 106–109 (in Japanese).

Japan Meteorological Agency 2015 Weather observation data. Japan Meteorological Agency Web. Available at: http://www.jma.go.jp/jma/index.html (accessed 28 January 2015).

Jeong, K. S., Hong, D. G., Byeon, M. S., Jeong, J. C., Kim, H. G., Kim, D. K. & Joo, G. J. 2010 Stream modification patterns in a river basin: field survey and self-organizing map (SOM) application. *Ecological Informatics* **5** (4), 293–303.

Jiang, N., Luo, K., Beggs, P. J., Cheung, K. & Scorgie, Y. 2014 Insights into the implementation of synoptic weather-type classification using self-organizing maps: an Australian case study. *International Journal of Climatology* **35** (12), 3471–3485.

Jin, Y. H., Kawamura, A., Park, S. C., Nakagawa, N., Amaguchi, H. & Olsson, J. 2011 Spatiotemporal classification of environmental monitoring data in the Yeongsan River basin, Korea, using self-organizing maps. *Journal of Environmental Monitoring* **13** (10), 2886–2894.

Kalteh, A. M. & Berndtsson, R. 2007 Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). *Hydrological Sciences Journal* **52**, 305–317.

Kalteh, A. M., Hjorth, P. & Berndtsson, R. 2008 Review of the self-organizing map (SOM) approach in water resources: analysis, modeling and application. *Environmental Modelling & Software* **23** (7), 835–845.

Liu, F., Song, X., Yang, L., Han, D., Zhang, Y., Ma, Y. & Bu, H. 2015 The role of anthropogenic and natural factors in shaping the geochemical evolution of groundwater in the Subei Lake basin, Ordos energy base, Northwestern China. *Science of the Total Environment* **538**, 327–340.

Marghade, D., Malpe, D. B. & Subba Rao, N. 2015 Identification of controlling processes of groundwater quality in a developing urban area using principal component analysis. *Environmental Earth Sciences* **74** (7), 5919–5933.

Matiatos, I. 2016 Nitrate source identification in groundwater of multiple land-use areas by combining isotopes and multivariate statistical analysis: a case study of Asopos basin (Central Greece). *Science of the Total Environment* **541**, 802–814.

Nadiri, A. A., Moghaddam, A. A., Tsai, F. T. C. & Fijani, E. 2013 Hydrogeochemical analysis for Tasuj plain aquifer, Iran. *Journal of Earth System Science* **122** (4), 1091–1105.

Nakagawa, K., Amano, H., Asakura, H. & Berndtsson, R. 2016 Spatial trends of nitrate pollution and groundwater chemistry in Shimabara, Nagasaki, Japan. *Environmental Earth Sciences* **75**, 234.

Nguyen, T. T., Kawamura, A., Tong, T. N., Nakagawa, N., Amaguchi, H. & Gilbuena Jr, R. 2015 Clustering spatio-seasonal hydrogeochemical data using self organizing maps for groundwater quality assessment in the Red River Delta, Vietnam. *Journal of Hydrology* **522**, 661–673.

Nishiyama, K., Endo, S., Jinno, K., Uvo, C. B., Olsson, J. & Berndtsson, R. 2007 Identification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a self-organizing map. *Atmospheric Research* **83** (2–4), 185–200.

Omonona, O. V., Onwuka, O. S. & Okogbue, C. O. 2014 Characterization of groundwater quality in three settlement areas of Enugu metropolis, southeastern Nigeria, using multivariate analysis. *Environmental Monitoring and Assessment* **186** (2), 651–664.

Ozeki, N., Okuno, M. & Kobayashi, T. 2005 Growth history of Mayuyam, Unzen, Kyushu, Southwest Japan. *Bulletin of the Volcanological Society of Japan* **50** (6), 441–454 (in Japanese with English abstract).

Shin, W. J., Ryu, J. S., Lee, K. S. & Chung, G. S. 2013 Seasonal and spatial variations in water chemistry and nitrate sources in six major Korean rivers. *Environmental Earth Sciences* **68** (8), 2271–2279.

Singaraja, C., Chidambaram, S., Prasanna, M. V., Thivya, C. & Thilagavathi, R. 2014 Statistical analysis of the hydrogeochemical evolution of groundwater in hard rock coastal aquifers of Thoothukudi district in Tamil Nadu, India. *Environmental Earth Sciences* **71** (1), 451–464.

Sonkamble, S., Sahya, A., Mondal, N. C. & Harikumar, P. 2012 Appraisal and evolution of hydrochemical processes from proximity basalt and granite areas of Deccan Volcanic Province (DVP) in India. *Journal of Hydrology* **438–439**, 181–193.

Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. 2000 *SOM Toolbox for Matlab 5*. Helsinki University of Technology Report A57.

WHO 2011 *Guidelines for Drinking-Water Quality*. 4th edn. World Health Organization, Geneva, Switzerland.

Wilcox, L. V. 1955 *Classification and use of irrigation water*. Circular No. 969. United States Department of Agriculture, Washington, DC, USA.

First received 23 February 2016; accepted in revised form 6 June 2016. Available online 14 July 2016