Master's Thesis of Natural Science

Statistical approaches to characterize relations between environmental conditions and groundwater quality

지하수 수질과 환경적인 조건의 상관성에 대한 통계적 접근

February 2019

Graduate School of Earth and Environmental Sciences

Seoul National University

Yu, Hakyeong E.

Abstract

Groundwater moves very slowly and has plenty of minerals through water-rock interaction. Because groundwater reflects geological features, distribution of water properties would be significantly different each region. The main purpose of this study is to specify geological conditions producing the groundwater that is suitable for drinking. First of all, multivariate analysis was conducted to classify the groundwater into groups in Hoengseong area. Eight major ions of Na, K, Ca, Mg, SO_4, NO_3, Cl, and HCO_3 were used for analysis. Result of principal component analysis for reducing variable, 3 components were enough for the analysis of whole data of 8 ions. First factor represented artificial contamination, second factor was abundance of carbonate minerals, and third factor was rich in sodium and bicarbonate. Using these 3 factors, the data were classified into 4 clusters by k-means cluster analysis. Properties of each cluster were explained by comparing with 3 factors. Cluster 1, 3, and 4 were less polluted groups and cluster 2 was significantly polluted. Cluster 1 was rich in carbonate minerals and cluster 3 was the lowest concentration of all ions and most suitable for drinking. Clustering analysis is proper to understand specific features of groundwater itself, but difficult to find out influence of environmental variables. And multivariate analysis would be a valid way to find clean water from the existing wells. To understand relations between water quality and environmental conditions, association rules were used. Association rule method could be a data mining method to find a new location where clean water would be produced from target area. Five variables of water quality, depth of the well, geological rock, land use and slope gradient of the location were used for association rules in Hoengseong area. There are total 1186 rules, and 42 rules related

with water quality that is suitable for drinking as result (right hand side of the rule). By three measures, there was highest probability on gneiss, forest, and steep slope area as conditions for suitable for drinking water as results. To verify the result of association rules in Hoengseong, association rule analysis was conducted again with national level of 1269 data. There were 3 variables of water quality, geology, and depth. There were total 90 rules, and 8 rules related with water quality of suitable for drinking as result (right hand side of the rule). The highest probability was on gneiss and shallow depth as conditions and suitable for drinking water as results.

Key words : Groundwater, Geochemical analysis, Statistical analysis, Multivariate analysis, Association rule, Water quality

TABLE OF CONTENTS

Abstract……...…….. i

TABLE OF CONTENTS……... iii

LIST OF FIGURES.....……...… v

LIST OF TALBES……...…… vii

1. Introduction 1

2. METERIALS AND METHODS.....4

2.1. Data 4

2.1.1. Standards of water quality 4

2.1.2. Groundwater data 7

2.2. Site Description 8

2.3. Multivariate statistical analysis 11

2.3.1. Principal components analysis 11

2.3.2. Factor analysis 11

2.3.3. k-means Clustering Analysis 12

2.3.4. Association Rule Analysis 13

3. RESULTS AND DISCUSSION.....15

3.1. Multivariate Analysis (Hoengseong) 15

3.1.1. Principal Component Analysis (Hoengseong) .15

3.1.2. Factor Analysis 17

3.1.3. Cluster Analysis 20

3.2. Association Rules (Hoengseong).....26

3.2.1. Discretization of data 26

3.2.2. Association Rules Analysis (Hoengseong) 34

3.3. Association Rules (National Data).....39

3.3.1. Discretization of data 39

3.3.2. Association Rules Analysis 40

4. CONCLUSIONS 45

5. REFERENCES.....47

LIST OF FIGURES

Figure 1. Location and geological map of Hoengseong province, South Korea. 9

Figure 2. Annual precipitation distribution in 2014 (South Korea) 10

Figure 3. The graph of variances related to each principal component in Hoengseong area. 16

Figure 4. Scatter-plot of 3 factors (Hoengseong).....19

Figure 5. Scatter-plot of 4 clusters (Hoengseong) 22

Figure 6. Histogram of well depth distribution (Hoengseong).....31

Figure 7. Network graph of total 67 rules (Hoengseong) 36

Figure 8. Network graph of top 10 rules for "Suitable for drinking" as result. Size of a circle means support values and color of a circle represents lift values. 38

Figure 9. Network graph of total 90 rules (national data). ...42

Figure 10. Network graph of only for "suitable for drinking". There are 9 rules related with "suitable for drinking". Size of a circle means support values and color of a circle represents

lift values. 44

LIST OF TABLES

Table 1. Water quality analysis from six bottled water companies in Gangwon Province (mg/L). 5

Table 2. Regulations for Drinking Water Management, Water Quality Standards and Inspection Rules for tap water and drinking water (Water quality standards in this study). 6

Table 3. The loading matrix from factor analysis (Hoengseong) 18

Table 4. Number of wells located on single type of geological rock. 29

Table 5. Seven wells located on multiple types of geological rock. Part of total data in Hoengseong. 30

Table 6. Part of categorized data of geology, depth and water quality. 32

Table 7. Categorized data in Hoengseong area (Total 227 wells). 33

Table 8. Top 10 rules in descending order by confidence. Part of total 1186 rules. (Hoengseong) 35

Table 9. Top 10 rules about “suitable for drinking” (Total 42

rules, Hoengseong) 37

Table 10. Top 10 rules in descending order by confidence. Part of total 90 rules. (National-wide data).....41

Table 11. Fourteen rules related with “Suitable for drinking”. 43

1. Introduction

Groundwater is important water resource that occupies 11 percent of available water resource of south Korea (National Groundwater Information Center of K-water). Dependence on groundwater gets higher especially drought season. Since the season when rainfall is concentrated is distinct in Korea, groundwater is helpful resource during drought season. Groundwater is not only stable water source of supply in all seasons, but also able to develop with small scale in inaccessible area.

Groundwater can be used for various purposes such as bottled water, hot springs, industrial use, and agriculture. Geochemical properties of groundwater depend on its use. For instance, temperature is the most important factor for hot springs. In the case of industrial groundwater, economic efficiency and specific ion concentration, such as iron and manganese, are important (National Groundwater Information Center of K-water). The water quality of groundwater is determined by various variables such as quantity, temperature, pH, various major ions and trace elements. And these water quality variables are affected by a number of environmental variables such as geology, slope, land use, depth of the well, precipitation, evapotranspiration, and distance from pollution sources. This study intends to find out the relations of water quality and environmental variables with statistical methods. Especially this study is focused on finding environmental variables for groundwater that is suitable for drinking.

The water quality variables are affected by environmental conditions. For example, ion concentrations of groundwater are produced by water-rock interaction and reflect geological features. Ion concentrations determine high marketability in the bottled water market. The water that is suitable for drinking should be fit into conditions

specified in the relevant laws, and of course, the lower the pollution, the better. Concentrations of major ions (Na, K, Ca, Mg, Cl, NO_3, SO_4, and HCO_3) are important elements in determining water quality. However, determining the suitability can be subjective. Therefore, in this study, ions concentration standards for drinking water were based on the water quality data from six companies in the spring water market and relevant laws.

There are several data set from national groundwater observation network, rural groundwater net, national groundwater background water quality observation network, contamination groundwater observation network and so on. And a massive amount of groundwater data base is being built with basic survey of groundwater across the country by K-water (National Groundwater Information Center of K-water). For water resource management, big data set of groundwater needs to be statistically analyzed, identified and managed.

There have been many studies using statistical methods to analyze the water quality of groundwater (Ko et al., 2005; Moon et al., 2002; Kim et al., 2016). However, these studies have been analyses of small data set or analyzed only the hydrological characteristics inherent in groundwater, and were not related to the environment in which groundwater is produced. Statistical approach is needed because there is a lack of statistical analysis of the environmental variables affecting the water quality of groundwater.

Data mining, on the other hand, refers to the methodology for extracting hidden patterns from big data set. Association rule analysis is one of powerful data mining techniques (Park et al.,1998). Association rules are statistical techniques that reveal strong associations between variables (also called as items), and finding useful patterns or associations in big data set can help decision making systems. (Han, J. et

al., 2011). The purpose of this study is to identify relations between the quality of groundwater and environmental variables and to discover rules about environmental conditions of groundwater suitable for drinking. The results of this study could be used to manage big data set of groundwater when the groundwater fundamental survey is completed in the future. Furthermore, this will enable the determination of environmental conditions for various uses of groundwater and help effective development of groundwater.

2. METERIALS AND METHODS

2.1. Data

2.1.1. Standards of water quality

Water with high ionic concentration of 4 ions (Na, K, Ca and Mg) is not suitable for drinking, but water with too low concentrations is also not good water. For pH, the suitable range is also needed. However, it can be too subjective to determine the appropriate range of ion concentrations and pH. Therefore, in case of 4 ions and pH, the target value between the minimum value and the maximum value of each variable was determined from water quality data of 6 kinds of bottled water. Water within this target value was classified as good water and the rest as bad water.

The remaining ion concentrations followed the legal standards for drinking water quality. Water that met all of these legal standards was classified as good water, and water that did not satisfy any of them was classified as bad water.

Table 1. Water quality analysis from six bottled water companies in Gangwon Province (mg/L).

Company	K	Na	Ca	Mg
A	1.1	4.7	11.4	3.3
B	8.89	21.14	34.74	4.22
C	0.6	7.09	13.43	1.97
D	0.53	4.8	20	1.16
E	0.38	5.16	10.2	1.47
F	0.5	19.8	18.7	6.1
	pH	**Latitude**	**Longitude**	**Geology**
A	7.21	38° 9'56.72"N	128°15'44.55"E	Gneiss
B	7.45	37°13'55.96"N	127°53'53.24"E	Granite
C	7.68	37°33'4.71"N	128°19'37.03"E	Granite
D	8.43	38°12'40.04"N	127°27'59.62"E	Granite
E	7.3	37°42'16.98"N	127°58'30.73"E	Granite
F	8.48	37°39'10.11"N	128° 9'43.57"E	Granite

Table 2. Regulations for Drinking Water Management, Water Quality Standards and Inspection Rules for tap water and drinking water (Water quality standards in this study).

	Legal standards		Legal standards
pH	4.5 ~ 9.5	**Pb**	0.01mg/L or less
Cl	250mg/L or less	**Cr**	0.05mg/L or less
SO_4	250mg/L or less	**TCE**	0.03mg/L or less
Zn	3mg/L or less	**PCE**	0.01mg/L or less
Fe	0.3mg/L or less	**1.1.1-TCE**	0.1mg/L or less
Mn	0.3mg/L or less	**Benzene**	0.01mg/L or less
Al	0.2mg/L or less	**Toluene**	0.7mg/L or less
NO_3	10mg/L or less	**Ethylbenzene**	0.3mg/L or less
Cd	0.005mg/L or less	**Xylene**	0.5mg/L or less
Ar	0.01mg/L or less	**Bacteria**	Undetected (250mL)
Phenol	0.005mg/L or less	**Turbidity**	1NTU or less
Hg	0.001mg/L or less		

2.1.2. Groundwater data

There are two sets of data used in this study. The first set is 227 data collected from basic survey of groundwater by K-water in the Hoengseong area. The second set is 1269 data of 106 areas all over South Korea collected from national groundwater background water quality network by national institute of environmental research.

Eight major ions (Na, K, Ca, Mg, Cl, NO_3, SO_4, and HCO_3), and pH are used as water quality variables of groundwater. And geology of the well location and the depth of the well are used as environmental variables of groundwater. Geological rock types are divide into granite, banded gneiss, granite gneiss, schist, amphibole, andesite, volcanic rock, etc. 97 percent of the wells are located at granite or gneiss.

The second set is 1269 data of 106 areas all over South Korea (17 regions of Gangwon province, 15 regions of Gyeonggi province, 10 regions of Gyeongnam province, 18 regions of Gyeongbuk province, 17 regions of Jeonnam province, 7 regions of Jeonbuk province, 9 regions of Chungnam province, 9 regions of Chungbuk, and Gwangju, Daegu, Incheon). The geological condition is divided into 8 types of rocks (Granite, Gneiss, Limestone, Clastic sedimentary rock, Semi-consolidated sedimentary rock, Unconsolidated sedimentary rock, Unconsolidated sediments, and Vesicular volcanic rock).

2.2. Site Description

Hoengseong is a county where is located in the middle area of South Korea and South-west of Gangwon Province. Hoengseong is located at 37°32′ to 36°41′ north degrees of latitude and 127°46′ to 128°20′east degrees of longitude. Hoengseong is 997.7 square kilometers in area. Because it is located inland mountains and is not affected by the ocean at all, it is a continental climate and a wide difference in temperature. Average annual temperature is 12.6 ℃, January average temperature is -3.2 ℃, August average temperature is 25.3 ℃, and annual precipitation is 1425.3 ㎜.

South Korea lies between latitudes 33° and 39°N, and longitudes 124° and 130°E and its total area is 100,032 square kilometers. South Korea has a humid continental climate and a humid subtropical climate, and is affected by the East Asian monsoon. Average annual temperature is 10 to 15 ℃, August average temperature is 23 to 26 ℃, January average temperature is -6 to 3 ℃. The average annual precipitation varies from 1200 to 1500 mm in middle area of the country, and 1000 to 1800 mm in south area. The average annual precipitation varies from 1370 mm in Seoul to 1470 mm in Busan. 50 to 60 percent of annual precipitation is concentrated in summer months.

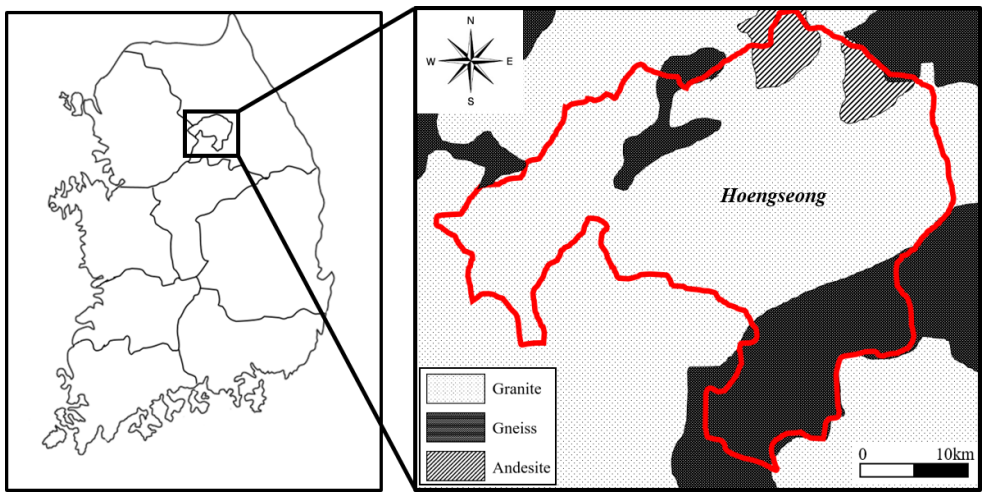

Figure 1. Location and geological map of Hoengseong province, South Korea.

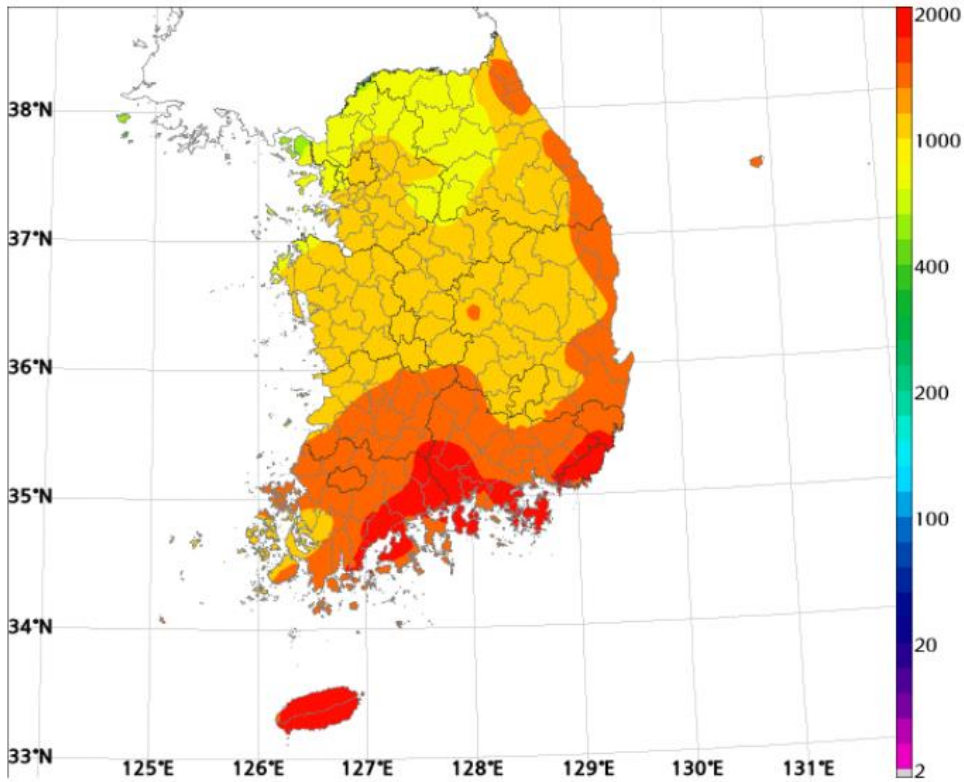

Figure 2. Annual precipitation distribution in 2014 (South Korea)

2.3. Multivariate statistical analysis

2.3.1. Principal components analysis

Principal components analysis (PCA) is an analytical technique to reduce the levels of several variables associated with each other while preserving as much information as possible. Principal component analysis has been widely used to evaluate the chemical composition of groundwater. The order of principal component analysis is as follows.

First, a linear combination of original variables calculates a set of uncorrected variables (factors or principal components). The number of principal components can be created by the maximum number of original variables, of which the first few principal components are ordered to have most of the variations of the original variables. Next, the values of the higher-dimensional raw data are converted into a small number of low-dimensional principal components. Dimensioning a dataset can visualize the data into a low-dimensional space. The reduction of data can greatly reduce some numerical algorithm computation times and can be used to provide linear associations with no correlation between covariates. PCA is based on decomposing the data matrix X (n by p) into two matrices U (n by k) and V (p by k). In higher dimensions (p), it transforms into lower dimensions (k). A suitable number of k can be found from the PCA results.

2.3.2. Factor analysis

Factor analysis, a multivariate statistical method, yields the general relationship between measured chemical variables by showing multivariate patterns that may help to classify the original data. It enables the geographical distribution of the resulting factors to be determined. The geological interpretation of factors yields

insight into the main processes, which may govern the distribution of hydro-chemical variables. (Chen-Wuing Liu, 2003) Factor Analysis is to divide data matrix X into products of two data matrices, U and V.

$$\boldsymbol{X = UV^T}$$

Eq.1

U is "score matrix" and **V** is "loading matrix". The loading is the weight for each of the original variables when calculating principal components. The weight means the coefficient of principal components.

In the principal component analysis, factor 1 explains for most of the variance. The rotation of a factor redefines a factor so that the loading of several factors is very large (near -1 or 1) or very small (near 0). There are orthogonal and oblique modes of rotation. Orthogonal rotation is a rotation that maintains the independence of factors. There are varimax and quartimax in the orthogonal rotation, the orthogonal rotation is common in analysis.

2.3.3. k-means Clustering Analysis

Cluster analysis is an analytical method of grouping several variable values into several clusters with similar characteristics. And it is a multivariate analysis technique that analyzes the relationships among the clusters by analyzing the characteristics of the formed clusters.

It is a hierarchical cluster analysis method that is widely used in previous studies as a method to classify groundwater or surface water quality. However, since hierarchical cluster analysis is difficult to handle and visualize large amounts of data, k-means cluster analysis method is used in this study.

The algorithm requires as input a matrix of M points in N dimensions and a matrix

of K initial cluster centers in N dimensions. Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$, the algorithm proceeds by alternating between two steps.

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the nearest mean.

$$\boldsymbol{S_i^{(t)} = \{x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \forall j,\ 1 \leq j \leq k\}} \qquad \text{Eq.2}$$

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$\boldsymbol{m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j} \qquad \text{Eq.3}$$

The algorithm has converged when the assignments no longer change. (Hartigan, 1979) The k-means clustering is a simple and fast algorithm and can handle more data than hierarchical clusters.

2.3.4. Association Rule Analysis

Data mining is a technology that identifies associations in big data and converts it into information to quantify. Data mining discovers hidden regularity in the database and the extraction of useful information. Association rule is one of data mining techniques that calculate the frequency and probability of simultaneous occurrence of data. This is an analysis method that reveals the association between one variable and another, or groups of variables. Association rule analysis finds hidden rules in the database and extract useful information. The basis of association rules is the calculation of frequency. A rule is in the form R: X $\Rightarrow$ Y, where X and Y are item groups that do not have the same elements. X is the condition of the rule and Y is the result. There are many interesting measures to identify only meaningful rules among

large number of total rules. Support is a measure of how many data out of the total data are simultaneously included and how many trends are identified. A rule with high confidence and high lift and low support is highly relevant rule but unlikely to occur. Confidence represents the probability that another variable will be included under the conditions of one variable and can determine the degree of association. As a conditional probability, the closer to 1 the greater the probability and the greater association. Lift is the rate of occurrence associated with the condition compared to the random result. If Lift is greater than 1 then it is more likely to have positive correlation and occur simultaneously.

3. RESULTS AND DISCUSSION

3.1. Multivariate Analysis (Hoengseong)

3.1.1. Principal Component Analysis (Hoengseong)

Principal component analysis is a multivariate analysis that uses an orthogonal transformation to reduce variables (Na, 2017). PCA uses the loading matrix of n × k to traverse the n × p matrix into the score matrix, which is the lower dimension of k × p. Case of Hoengseong, eight ions were used from 227 wells, size of original data set is the matrix of 227×8. To find a suitable number of k for the analysis, the variance associated with the principal component was. The first three components are shown to explain for most of the variation in the data, based on the variance of 1.0, in Hoengseong calculated (Figure 3). Therefore, three main components were selected. Varimax orthogonal transformation was chosen, this is the orthogonal transformation of principal components to maximize the factor loading (Jöreskog et al., 1993).

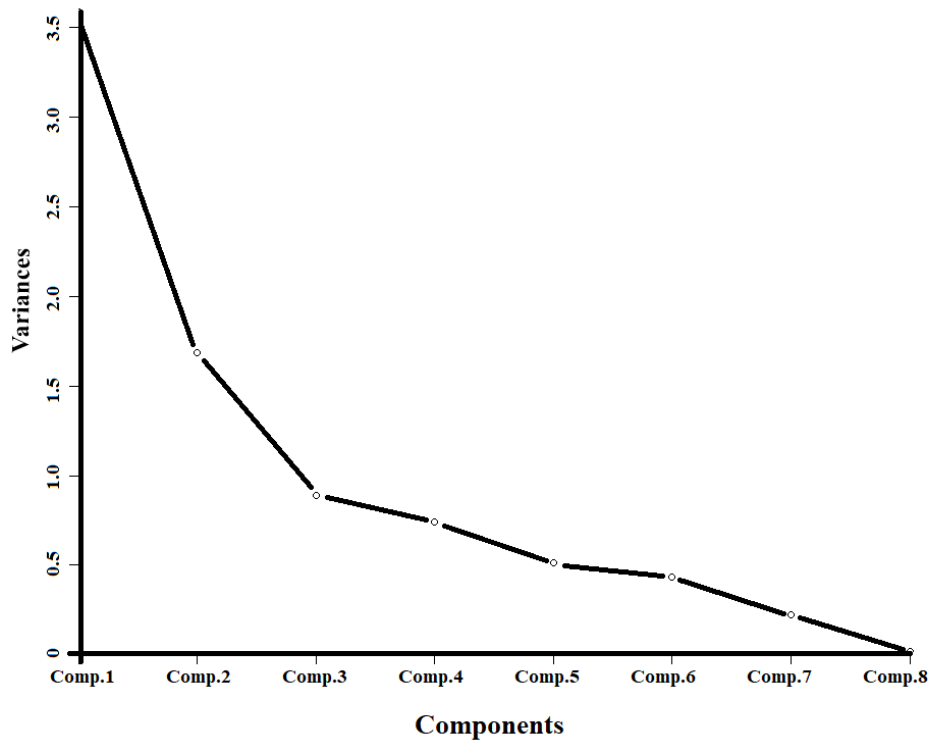

Figure 3. The graph of variances related to each principal component in Hoengseong area.

3.1.2. Factor Analysis

Factor Analysis is useful method for explaining groundwater quality data (Lawrence et.al., 1982). Three main components were selected as the main component analysis method. The original data matrix was 8 × 227, which could be separated by score matrix of 227 × 3 and loading matrix of 3 × 8. Loading matrix of 3 × 8 provides loading values of each factor about original variables of eight ions (Table 3) The loading values in this loading matrix allow the interpretation of the meaning of each factor. Positive values mean positive correlations, and negative values are negative correlations. If there is no value, there is no correlation between the variables. And the score matrix of 227 × 3 was a new matrix of data that could be obtained as a result of factorial analysis.

In factor 1, the loading value of SO_4 was 0.471, Cl was 0.710, and NO_3 was 0.697. Sulfate, chlorine, and nitrates are the main contents which represent the pollution. Since there were high correlations with SO_4, Cl, and NO_3, factor 1 indicates artificial contaminations. Ca was also highly correlated with 0.585, but Ca was much more correlated with 0.808 in factor 2. In factor 2, three ions of calcium, magnesium and bicarbonate were highly correlated with 0.808 on Ca, 0.732 on Mg and 0.590 on HCO_3. Therefore, factor 2 directs the abundance of carbonate ions of Ca, Mg and HCO_3. For example, if Factor2 is high in the region, it may be assumed that it is affected by carbonate minerals, which can be checked against the actual geological map. Factor 3 showed a high correlation between the two ions with Na of 0.98 and HCO_3 of 0.799.

Table 3. The loading matrix from factor analysis (Hoengseong)

	Factor 1	Factor 2	Factor 3
Na	0.175		0.982
K	0.452	0.111	0.144
Ca	0.585	0.808	
Mg	0.292	0.732	0.156
SO_4	0.471	0.393	0.104
Cl	0.710	0.286	0.159
NO_3	0.697	0.164	-0.214
HCO_3		0.590	0.799

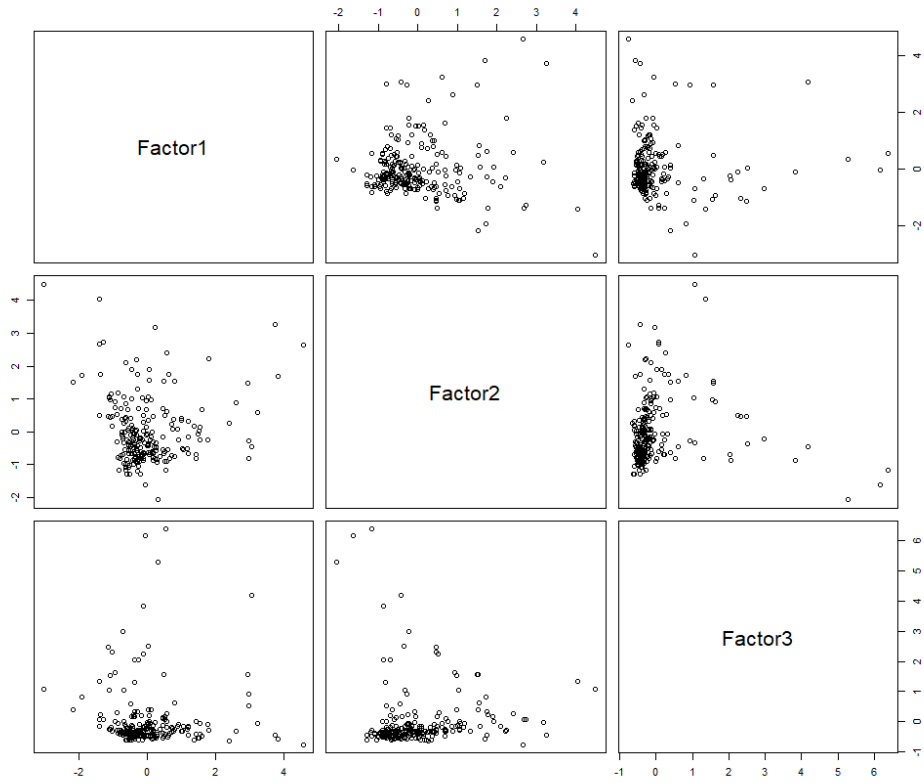

Figure 4. Scatter-plot of 3 factors (Hoengseong)

3.1.3. Cluster Analysis

Cluster Analysis is a multivariate analysis technique that groups several clusters of similar characteristics among different objects and identifies and analyzes characteristics of each cluster. K-means clustering is one of cluster analysis. This method specifies the initial value as much as the number of clusters of k, forms the cluster based on the nearest initial value in each data, and updates the center value by recalculating the mean of each cluster. This repeats the new allocation process for the new central value and forms the final cluster of k when the center value is no longer moving.

The score matrix, which consists of three factor values from the result of factor analysis, is used for k-means clustering. Using k-means clustering, 227 wells were grouped into four clusters, as shown in figure 5.

Figure 5 (a) it the scatterplot of factor 1 (y-axis) and factor 2 (x-axis). Factor 1 was a factor that directs artificial pollutants with high correlation between SO_4, Cl and NO_3. Since Factor 1 was higher than other clusters, Cluster 2 is a relatively contaminated cluster. Cluster 1 showed high values in factor 2 value. Factor 2 had a high correlation of Ca, Mg and HCO_3, indicating carbonate minerals. Cluster 1, a cluster with high Factor 2, was a cluster which was rich in carbonate minerals and relatively low levels of pollution. Figure 5 (b) is a scatterplot of factor 1 (y-axis) and factor 3 (x-axis). Also in this instance, cluster 2 was a cluster of contaminated one because the value of factor 1 is higher than other clusters. For cluster 4, factor 3 was high. Factor 3 directed Na and HCO_3, so it was shown that Cluster 4 was a cluster with high Na and HCO_3 concentrations. Figure 5 (c) is a scatterplot of factor 2 (y-axis) and factor 3 (x-axis). In this figure, cluster 1 had a higher factor 2 and cluster 4 had a higher factor 3, as well. In summary, Cluster 1 was a cluster of low pollution

and high carbonate minerals, cluster 2 was a cluster of high-contamination, cluster 3 was low-contamination and the lowest concentration of ions overall, and cluster 4 was rich in Na and HCO_3 with low contamination.

The results of cluster analysis also differed comparing each cluster with the geological conditions of wells. Among the wells in Hoengseong area, the proportion of wells located in gneiss rock zone was 24%. Twenty-two percent of cluster 1 was located on gneiss and 78% of cluster 1 was located on granite zone. Cluster 2 was located in 9% of gneiss, 9% of volcanic rocks, and 82% of granite areas. Cluster 3 was located at 34 % of gneiss, 2 % of volcanic rock and 64 % of granite. Cluster 4 was located on 42% of gneiss and 58 % of granite. For Cluster 3 with the lowest contamination and the lowest ionic concentration, and cluster 4 had relatively high probability on gneiss rock zone.

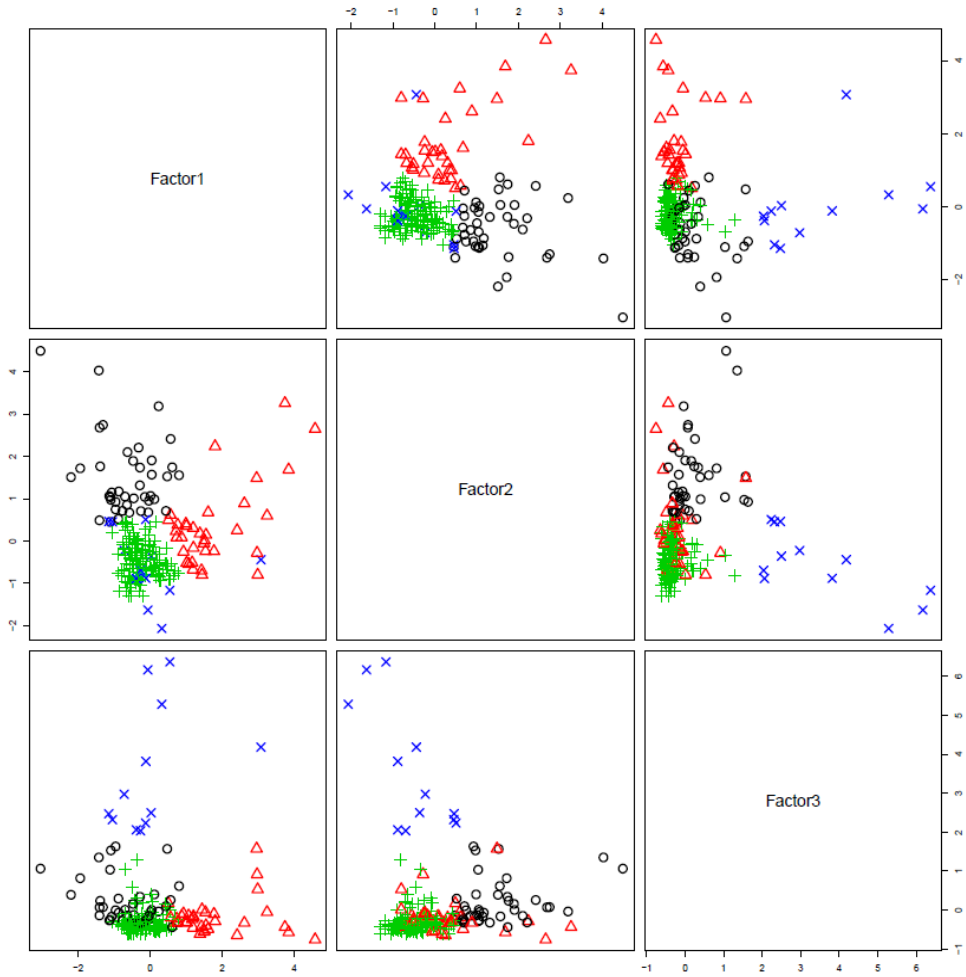

Figure 5. Scatter-plot of 4 clusters (Hoengseong)

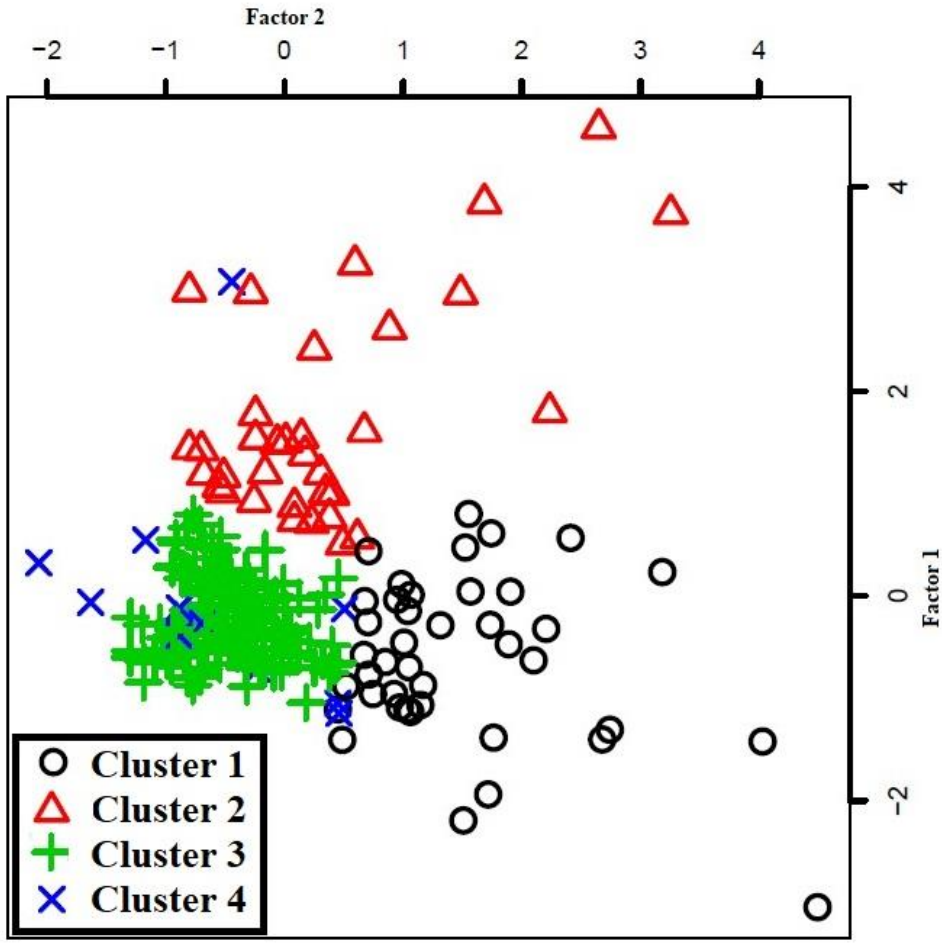

Figure 5. (a) Scatter-plot of factor 1 and factor 2. Four color represent each 4 clusters from k-means clustering analysis. (Hoengseong)

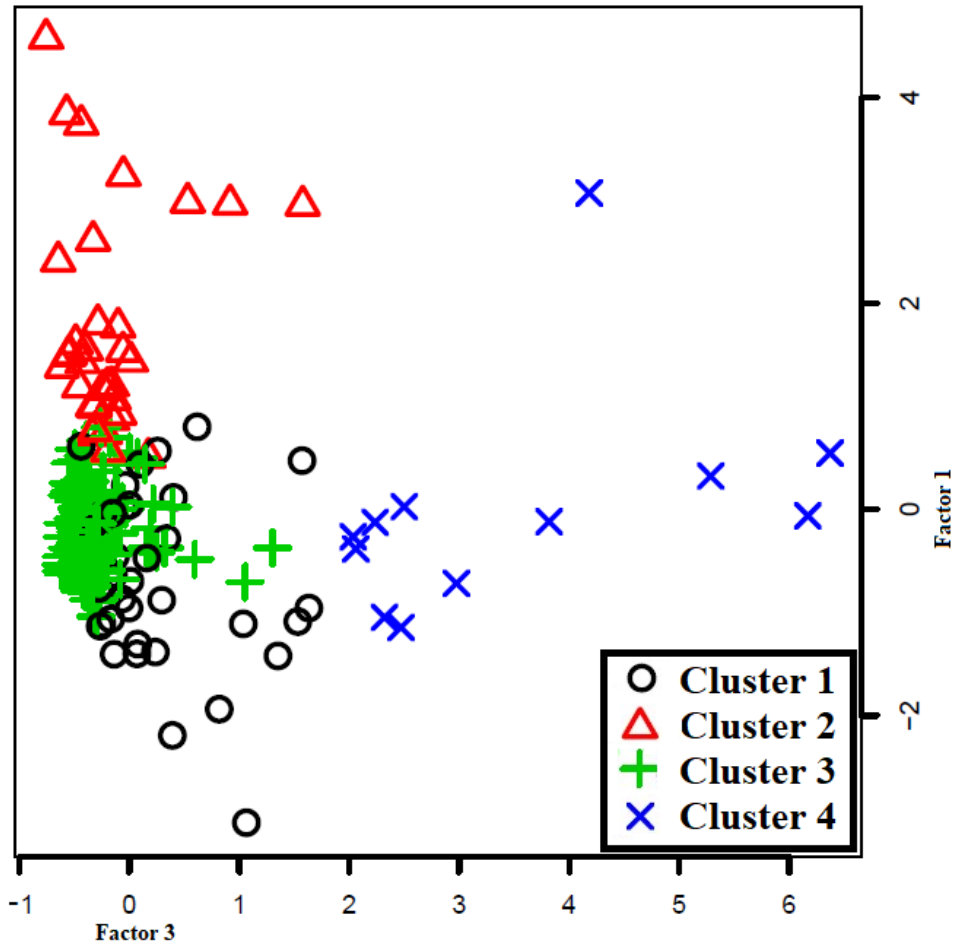

Figure 5. (b) Scatter-plot of factor 1 and factor 3. Four color represent each 4 clusters from k-means clustering analysis. (Hoengseong)

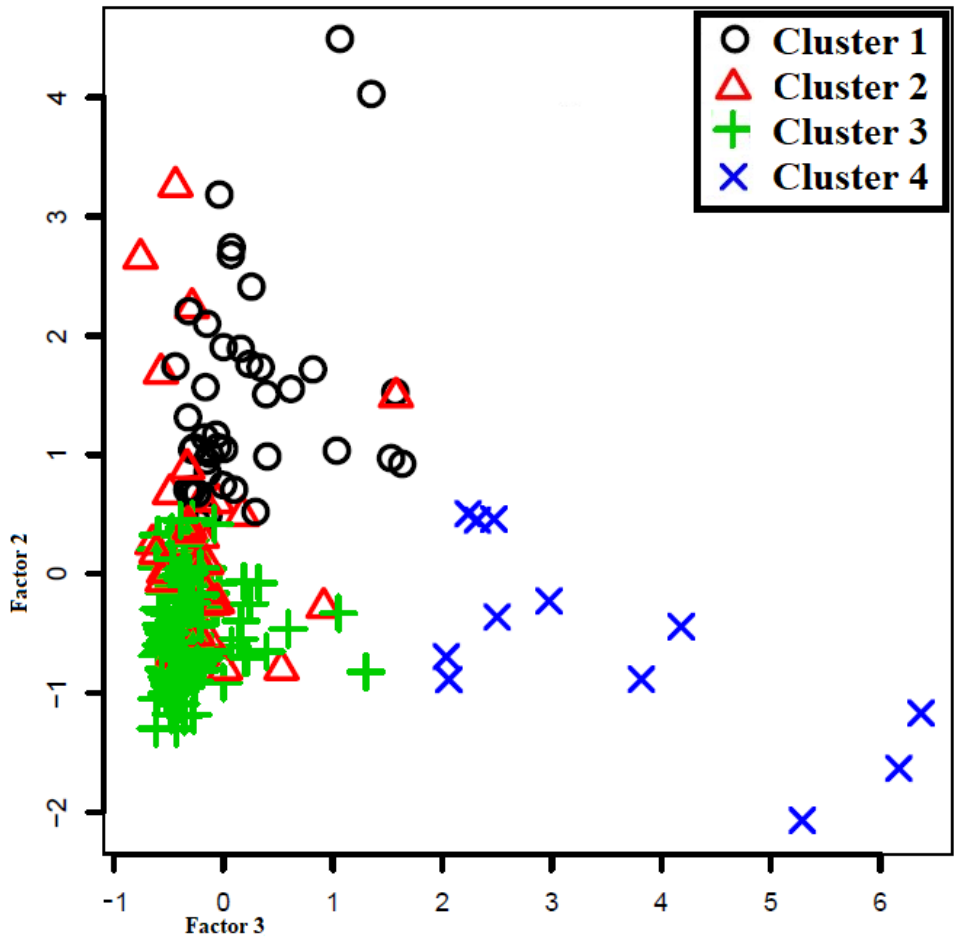

Figure 5. (c) Scatter-plot of factor 2 and factor 3. Four color represent each 4 clusters from k-means clustering analysis. (Hoengseong)

3.2. Association Rules (Hoengseong)

3.2.1. Discretization of data

In order to perform association rules, continuous data should be discrete. Geological features were classified into 3 types of rock of granite, gneiss, and volcanic rock. The depth is divided into three categories: Shallow, Medium and Deep, based on 50m and 100m. Water quality was categorized into two categories: suitable for drink and not suitable. This is based on eight major ion concentration (Na, K, Ca, Mg, SO_4, NO_3, Cl, HCO_3) and pH by 6 types of bottled water quality data. Land use of the area (Radius 100 m) is divided into three categories: forest, medium and farm area. Slope gradient is divided into three categories: Gradual, Medium, and Steep. Slope gradient was divided by 7 and 12 degrees.

a) Geology

There are nine sub-categories specifically, granite, banded gneiss, granite gneiss, schist, amphibole, leucocratic gneiss, volcanic rock, granite porphyry, and andesite. This is based on the data of geology within 100 m radius from the well. Most wells were found to comprise 100 % of a single geology. Of the total 227 wells, there are 152 wells that are located on granite, 5 wells that are located on banded gneiss, 51 wells that are located on granite gneiss, 3 wells that are located on schist, 2 wells are on leucocratic gneiss, one well is located on sedimentary rock or volcanic rock, one well is on granite porphyry, and 5 wells are located on andesite. Total of 220 wells were located in a single geological area (Table 4). The remaining seven wells were classified as geology if the well had a dominant presence of more than 75%. For example, in case of HSR-0089, it was classified as granite since it was located on 98 % of granite and 2 % of banded gneiss. In the case of HSR-0714, this well is located on 44% of granite and 56% of granite porphyry (Table 5). It could not be

classified by this method, was classified as granite, because granite and granite porphyry have similar chemical composition.

In conclusion, geological conditions of this area were largely categorized into three main types: 157 wells on granite, 64 wells on gneiss, 6 wells on volcanic rock. These three main types were used in this study.

b) Depth

Data on the depth of wells are continuous data. The depth of the well was divided into three categories on the basis of 50 m and 100 m. Wells with a depth of less than 50 m were classified as "shallow". Wells with a depth of above 50 m and not more than 100 m were classified as "medium". And wells with more than 100 m of depth were classified as “deep”. There were 60 shallow depth wells, 55 medium depth wells, and 112 deep depth wells.

c) Water quality (ionic concentration)

Na, K, Ca, Mg and pH were divided into two categories of “Suitable for drinking” and “Not suitable (for drinking)” based on water quality data of 6 bottled water (Table 1). Based on minimum and maximum values, five variables were divided into three categories: "Low", "Suitable", and "High". Na ranged 4.7 ~ 21.14 (mg/L). K ranged 0.38 ~ 8.89 (mg/L). Ca ranged 10.2 ~ 34.74 (mg/L). Mg ranged 1.16 ~ 6.1 (mg/L). NO3, Cl, SO_4 and HCO_3 were divided into two categories of "Suitable” and "Not suitable” based on maximum value only (Table 2). In conclusion, water quality that was suitable for all 9 variables was classified as "Suitable for drinking" and the rest as "Not suitable". There are 26 wells of “Suitable for drinking” and 201 wells of “Not suitable”.

d) Land use

Land use data was based within 100 m radius from the well. This was divided into three categories: forest, medium and farm. Forest zone is less than 33 % of farm zone. Farm zone was where more than 67 percent of the area is covered by farms. Medium zone was between 33 and 67 percent. There were 129 wells on forest area, 43 wells on medium zone and 55 wells on farm zone.

e) Slope gradient

Slope gradient was divided into 3 categories: Gradual, Medium and Steep. This classification was based on 1 quartile of data (7 degree) and 3 quartile of data (12 degree). There were 57 wells of gradual slope, 113 wells of medium slope and 57 wells of steep slope gradient.

Table 4. Number of wells located on single type of geological rock.

	Number of wells located on single type of geological rock
Granite	152
Banded gneiss	5
Granite gneiss	51
Schist	3
Amphibole	0
Leucocratic gneiss	2
Volcanic rock	1
Granite porphyry	1
Andesite	5
Total	220

Table 5. Seven wells located on multiple types of geological rock. Part of total data in Hoengseong.

Well Number	Granite	Banded gneiss	Granite gneiss	Schist	Amphibole	Leucocratic gneiss	Volcanic rock	Granite porphyry	Andesite	Original Category	Large Category
HSR-0089	0.98	0.02	0	0	0	0	0	0	0	Granite	Granite
HSR-0164	0	0.99	0.01	0	0	0	0	0	0	Banded gneiss	Gneiss
HSR-0208	0	0.13	0	0.87	0	0	0	0	0	Schist	Gneiss
HSR-0399	0	0.03	0	0	0	0.97	0	0	0	Leucocratic gneiss	Gneiss
HSR-0421	0.94	0	0	0	0.06	0	0	0	0	Granite	Granite
HSR-0714	0.44	0	0	0	0	0	0	0.56	0	Unclassified	Granite
HSR-0742	0.96	0	0	0	0	0	0	0.04	0	Granite	Granite

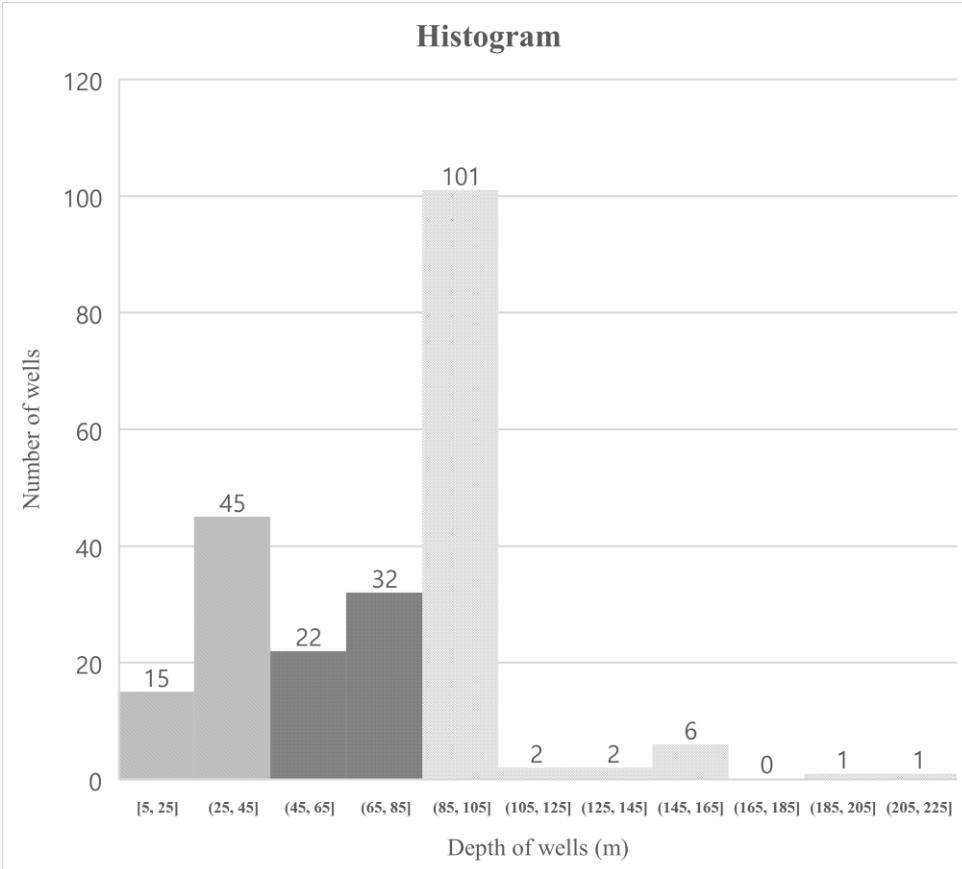

Figure 6. Histogram of well depth distribution (Hoengseong)

Table 6. Part of categorized data of geology, depth and water quality.

Well number	Class	Geology	Depth
HSR-0088	Not suitable	Gneiss	Med
HSR-0089	Not suitable	Granite	Deep
HSR-0091	Suitable for drinking	Granite	Shallow
HSR-0093	Not suitable	Granite	Med
HSR-0104	Not suitable	Granite	Med
HSR-0114	Not suitable	Granite	Shallow
HSR-0119	Not suitable	Granite	Deep
HSR-0124	Not suitable	Granite	Deep
HSR-0126	Not suitable	Granite	Deep
HSR-0127	Suitable for drinking	Granite	Deep
HSR-0135	Not suitable	Granite	Shallow
HSR-0143	Not suitable	Gneiss	Shallow
HSR-0153	Not suitable	Gneiss	Shallow
HSR-0157	Not suitable	Gneiss	Med
HSR-0161	Not suitable	Gneiss	Deep

Table 7. Categorized data in Hoengseong area (Total 227 wells).

Variables	Category	Number of wells
Water quality	Suitable	26
	Not suitable	201
Geology	Granite	157
	Gneiss	64
	Volcanic rock	6
Depth	Shallow	60
	Medium	55
	Deep	112
Land use	Forest	129
	Medium	43
	Farm	55
Slope gradient	Gradual	57
	Medium	113
	Steep	57

3.2.2. Association Rules Analysis (Hoengseong)

The result of association rules using 5 variables of water quality, depth of the well, geology, land use and slope gradient, there were 1186 rules in Hoengseong area (Minimum support value = 0.01, Minimum confidence value = 0.1). Top 10 results in descending order for confidence are in table 8. The highest rule for confidence was land use of forest as a result and volcanic rock as a condition. There are only six wells that are on volcanic rock zone. So additional study using large amounts of volcanic rock data will be needed, in order to determine whether this result can be seen as a significant result. In case of depth, artificial pollutants are mainly located on the surface, it was expected that the deeper the well, the better the quality of the water, but the result was the opposite. This is because not only artificial contamination was considered, but also high concentrations of cations were classified as "Not suitable for drinking". A high concentration of Ca and Mg results in a high hardness, which is not suitable for drinking.

Since the purpose of this study is to find the conditions for good drinking water, only 42 rules for "suitable for drinking" of water quality as results were compiled out of the whole 1186 association rules (Table 8). The highest rule for lift and confidence was in conditions with gneiss and forest and steep slope area. Considering only geology as condition, there was the highest probability that suitable water for drinking was in gneiss, the probability at granite zone was lower than gneiss zone. There were 6 wells in volcanic rock zone, and all of those 6 wells were not suitable for drinking.

Table 8. Top 10 rules in descending order by confidence. Part of total 1186 rules. (Hoengseong)

	Lhs (Condition)		Rhs (Result)	support	confidence	lift
[1]	{Geology=Volcanic Rock}	⇒	{LandUse=Forest}	0.026	1.000	1.760
[2]	{Geology= Volcanic Rock }	⇒	{Class=Not Suitable}	0.026	1.000	1.129
[3]	{Geology= Volcanic Rock, Depth=Deep}	⇒	{LandUse= Forest }	0.018	1.000	1.760
[4]	{Geology= Volcanic Rock, Depth=Deep}	⇒	{Class=Not Suitable}	0.018	1.000	1.129
[5]	{Geology= Volcanic Rock, LandUse=Forest}	⇒	{Class=Not Suitable}	0.026	1.000	1.129
[6]	{Class=Not Suitable, Geology= Volcanic Rock }	⇒	{LandUse= Forest }	0.026	1.000	1.760
[7]	{LandUse=Medium, SlopeGradient=Steep}	⇒	{Class=Not Suitable}	0.044	1.000	1.129
[8]	{Geology= Volcanic Rock, Depth=Deep, LandUse=Forest}	⇒	{Class=Not Suitable}	0.018	1.000	1.129
[9]	{Class=Not Suitable, Geology= Volcanic Rock, Depth=Deep}	⇒	{LandUse= Forest }	0.018	1.000	1.760
[10]	{Class=Suitable for drinking, Geology=Gneiss, LandUse=Medium}	⇒	{SlopeGradient=Medium}	0.013	1.000	2.009

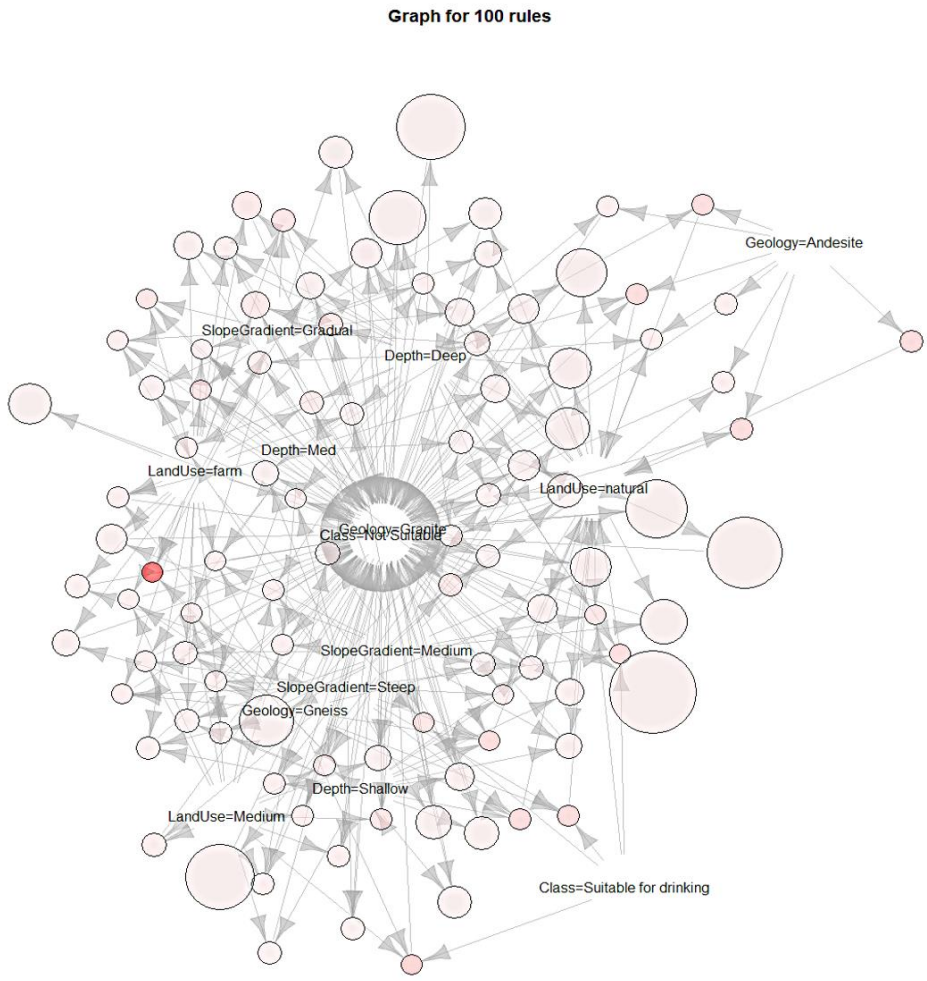

Figure 7. Network graph of total 67 rules (Hoengseong)

Table 9. Top 10 rules about "suitable for drinking" (Total 42 rules, Hoengseong)

lhs (condition)		rhs (result)	support	confidence	lift
{Geology=Gneiss, Land Use=Forest, Slope Gradient=Steep}	→	{Class=Suitable for drinking}	0.018	0.174	1.518
{Geology=Gneiss, Land Use=Medium, Slope Gradient=Medium}	→	{Class=Suitable for drinking}	0.018	0.167	1.455
{Geology=Granite, Depth=Shallow, Slope Gradient=Gradual}	→	{Class=Suitable for drinking}	0.044	0.156	1.364
{Depth=Shallow, Slope Gradient=Gradual}	→	{Class=Suitable for drinking}	0.022	0.135	1.180
{Geology=Gneiss, Depth=Deep, Land Use=Forest}	→	{Class=Suitable for drinking}	0.035	0.133	1.164
{Geology=Gneiss, Slope Gradient=Steep}	→	{Class=Suitable for drinking}	0.031	0.127	1.111
{Geology=Gneiss, Land Use=Forest}	→	{Class=Suitable for drinking}	0.018	0.111	0.970
{Geology=Granite, Land Use=farm, Slope Gradient=Medium}	→	{Class=Suitable for drinking}	0.070	0.102	0.890
{Depth=Deep, Land Use=Forest, Slope Gradient=Steep}	→	{Geology=Granite}	0.070	0.615	0.890
{Geology=Gneiss, Depth=Shallow}	→	{Depth=Deep}	0.048	0.423	0.857
{Geology=Gneiss, Land Use=Forest, Slope Gradient=Steep}	→	{Geology=Gneiss}	0.044	0.385	1.364
{Geology=Gneiss, Land Use=Medium, Slope Gradient=Medium}	→	{Depth=Shallow}	0.035	0.308	1.164
{Geology=Granite, Depth=Shallow, Slope Gradient=Gradual}	→	{Depth=Med}	0.031	0.269	1.111

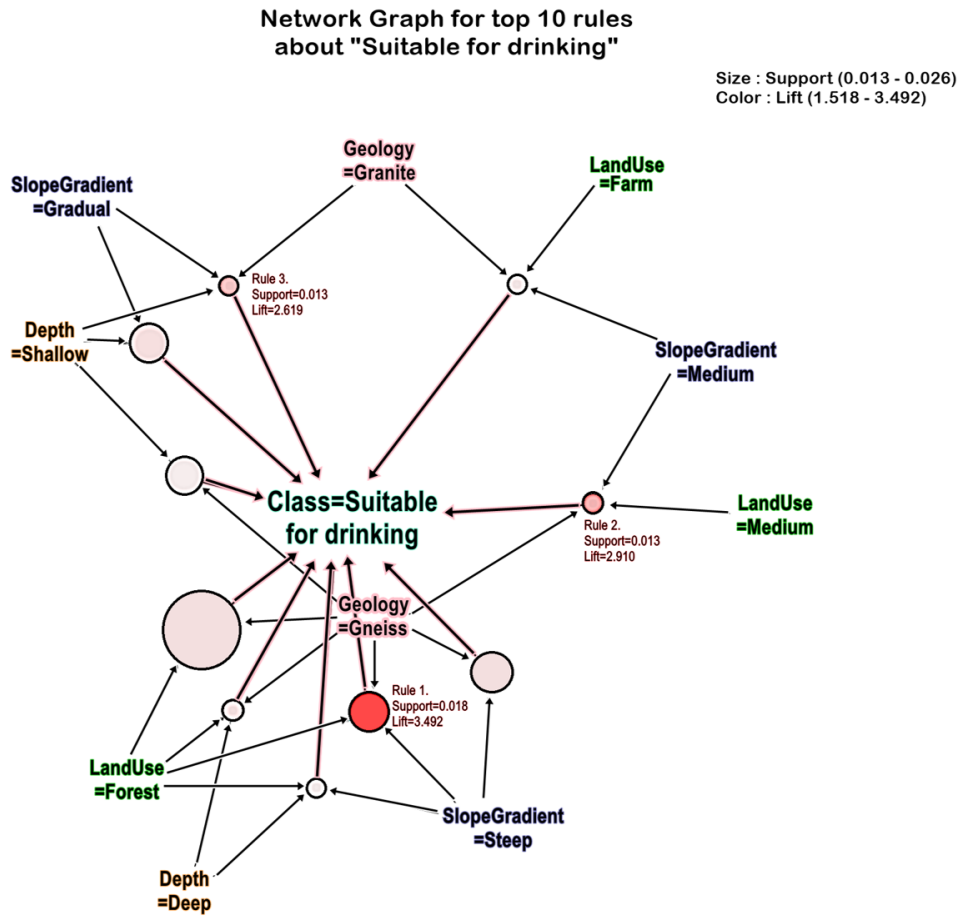

Figure 8. Network graph of top 10 rules for "Suitable for drinking" as result. Size of a circle means support values and color of a circle represents lift values.

3.3. Association Rules (National Data)

Additional association rules were applied to national-wide data to verify the result in Hoengseong. Total 1269 data collected across the country from the National Institute of Environmental Research were used.

3.3.1. Discretization of data

Associated rules require a categorical change of continuous data. Geological variable used eight classifications. The depth is divided into three categories, Shallow, Medium and Deep, based on 50m and 100m. Water quality was categorized into two categories of "suitable for drinking" and "not suitable" based on bottled water using 10 variables of eight major ions, and pH.

a) Geology

97.3% of the wells are located in gneiss or granite, and data were divided into 3 categories of granite, granite and volcanic rock. Nationwide data were evenly distributed in the number of wells belonging to each geographical category comparing with the data in Hoengseong. The purpose of this part is to examine whether the same rules appear when different numbers of categories are used with different data. There are 336 well on granite, 336 wells on gneiss, 264 wells on unconsolidated sediments, 167 wells on clastic sedimentary rock, 96 wells on non-vesicular volcanic rock, 46 wells on limestone, 12 wells on semi-consolidated sedimentary rock, and 12 vesicular volcanic rock.

b) The depth

As in Hoengseong, the depth of the well was divided by 50 m and 100 m with 3 categories of “Shallow”, “Medium”, and “Deep”. Shallow 845 wells, Medium 416 wells, and Deep 8 wells, there were few deep wells compared to Hoengseong.

c) Water quality

Based on the bottled water data, maximum and minimum values of Na, K, Ca, and Mg were selected. The corresponding values were classified as "suitable for drinking" and rest of data were classified as "not suitable" (for drinking). There were 187 wells for "suitable for drinking" and 1,082 wells for "not suitable".

3.3.2. Association Rules Analysis

From 1269 data nationwide, there were a total of 90 association rules (minimum support value 0.01, minimum confidence value 0.1). The highest association rule was "not suitable" in water quality under the condition of clastic sedimentary rock and medium depth of the well (Table 9). Only considering geological conditions, the rock most associated with not suitable for drinking water was clastic sedimentary rock, followed by unconsolidated sediments and limestone. Only considering the depth of wells, there was the highest probability that medium-depth wells (50 m ~ 100 m) would yield "not suitable" water, and then shallow-depth wells (~ 50 m) next.

To find conditions about drinking water, so only the rules related to "Suitable for drinking" were compiled separately. There were five rules under condition, nine rules in result, and 14 rules in total (Table 10). The highest association rule was the condition in which the depth was shallow (less than 50 meter), located in the gneiss zone. The following was the case of wells with medium depths (50 m ~ 100 m) located in intrusive igneous rocks. Considering only geological characteristics, the most powerful rule was about gneiss rock, followed by intrusive igneous rock and non - porous volcanic rock. For depth, the probability was higher under shallow wells and then at medium depths.

Table 10. Top 10 rules in descending order by confidence. Part of total 90 rules. (National-wide data)

Condition		Result	Support	Confidence	Lift
{Geology=Clastic Sedimentary Rock, Depth=Medium}	⇒	{Class=Not suitable }	0.042	0.946	1.110
{Geology=Limestone, Depth=Medium}	⇒	{Class=Not suitable }	0.012	0.938	1.100
{Geology=Unconsolidated Sediments, Depth=Medium}	⇒	{Class=Not suitable }	0.065	0.932	1.093
{Geology=Clastic Sedimentary Rock }	⇒	{Class=Not suitable }	0.122	0.929	1.089
{Geology= Clastic Sedimentary Rock, Depth=Shallow}	⇒	{Class=Not suitable }	0.080	0.919	1.078
{Geology=Unconsolidated Sediments}	⇒	{Class=Not suitable }	0.191	0.917	1.075
{Geology=Unconsolidated Sediments, Depth=Shallow}	⇒	{Class=Not suitable }	0.126	0.909	1.066
{Depth=Medium}	⇒	{Class=Not suitable }	0.282	0.861	1.009
{Depth=Shallow}	⇒	{Class=Not suitable }	0.565	0.849	0.995
{Geology=Limestone}	⇒	{Class=Not suitable }	0.031	0.848	0.994

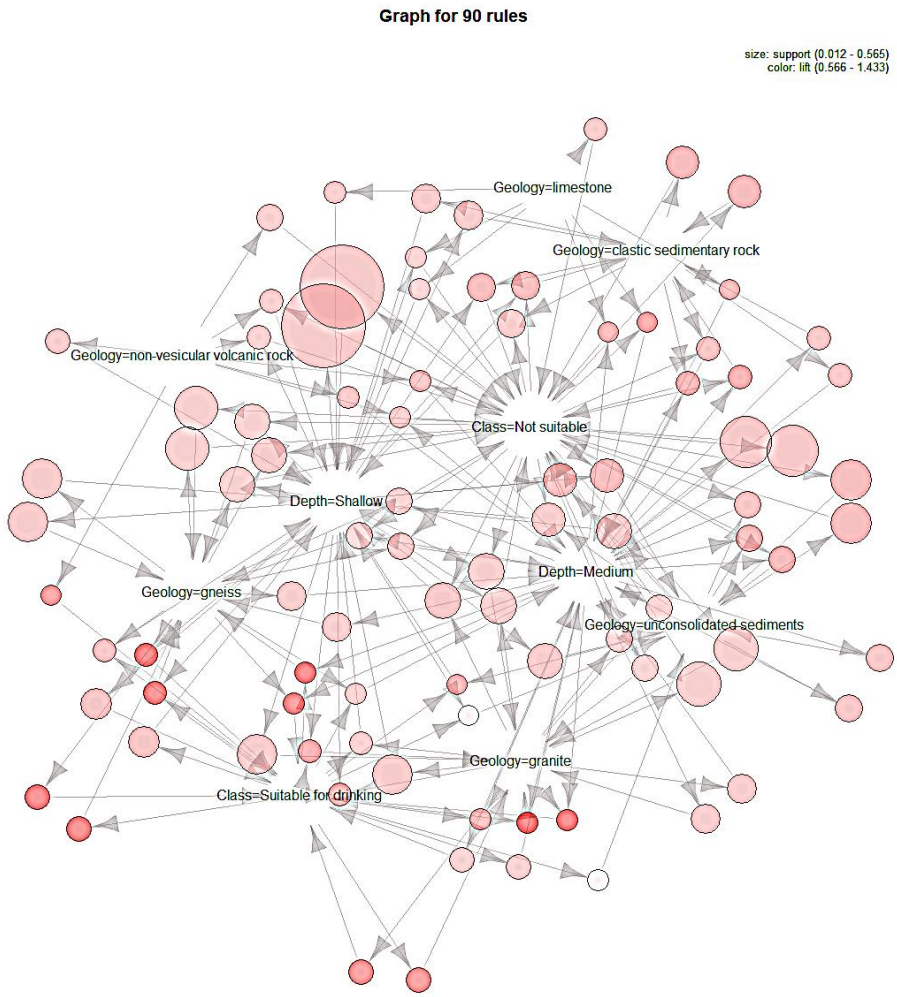

Figure 9. Network graph of total 90 rules (national data).

Table 11. Fourteen rules related with "Suitable for drinking".

lhs (condition)		rhs (result)	Support	Confidence	Lift
{Geology=Gneiss, Depth=Shallow}	⇒	{Class=Suitable for drinking}	0.036	0.205	1.394
{Geology=Granite, Depth=Medium}	⇒	{Class=Suitable for drinking}	0.017	0.204	1.382
{Geology=Gneiss}	⇒	{Class=Suitable for drinking}	0.053	0.199	1.353
{Geology=Gneiss, Depth=Medium}	⇒	{Class=Suitable for drinking}	0.017	0.194	1.320
{Geology=Granite}	⇒	{Class=Suitable for drinking}	0.049	0.185	1.252
{Geology=Vesicular volcanic rock}	⇒	{Class=Suitable for drinking}	0.013	0.177	1.202
{Geology=Granite, Depth=Shallow}	⇒	{Class=Suitable for drinking}	0.031	0.174	1.182
{Depth=Shallow}	⇒	{Class=Suitable for drinking}	0.101	0.152	1.028
{Depth=Medium}	⇒	{Class=Suitable for drinking}	0.046	0.139	0.946
{Class=Suitable for drinking}	⇒	{Depth=Shallow}	0.101	0.685	1.028
{Class=Suitable for drinking}	⇒	{Geology=Gneiss}	0.053	0.358	1.353
{Class=Suitable for drinking}	⇒	{Geology=Granite}	0.049	0.332	1.252
{Class=Suitable for drinking}	⇒	{Depth=Medium}	0.046	0.310	0.946
{Class=Suitable for drinking}	⇒	{Geology=Unconsolidated Sediments }	0.017	0.118	0.566

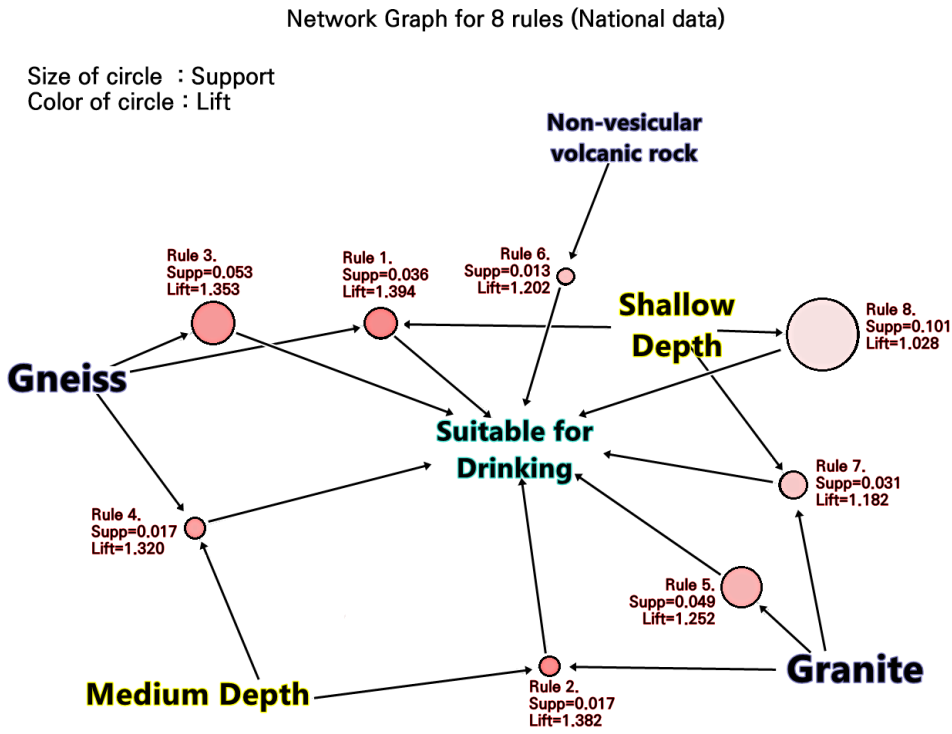

Figure 10. Network graph of only for "suitable for drinking". There are 9 rules related with "suitable for drinking". Size of a circle means support values and color of a circle represents lift values.

4. CONCLUSIONS

The purpose of this study was to find the environmental variables producing water that is suitable for drinking. The characteristics of the target groundwater clusters were identified through multivariate analysis, and association rules were used to in this study specify environmental conditions.

1) The groundwater quality data of 227 wells in the Hoengseong area were analyzed by multivariate analysis of eight kinds of Na, K, Ca, Mg, Cl, NO_3, SO_4 and HCO_3. Using principal component analysis and factor analysis, variables were reduced to 3 factors. Factor 1 with high correlation of SO_4, NO_3, and Cl, was the factor about artificial contamination. Factor 2 was a highly correlated factor of Ca, Mg and HCO_3, and factor 3 was a factor with high correlation of Na and HCO_3. With these three factors, as a result of k-means cluster analysis, it was classified into four clusters. Cluster 1 was a low-pollution, carbonate-rich cluster, and cluster 2 was a highly polluted. Cluster 3 was low in pollution level and low in total ion concentration. Finally, Cluster 4 was low in pollution level and rich in Na and HCO_3.

2) For association rules, categorized 5 variables were used (Water quality, geology, depth, land use, and slope gradient). There were total 1186 rules, and 42 rules related to water quality of suitable for drinking as result. The highest rule was about gneiss and forest area with steep slope gradient as conditions. Only considering the geological condition, it was more likely that water was suitable for drinking in the order of gneiss, granite, and volcanic rock.

3) The method of categorizing 1269 data across the country for association rules was conducted. There were 3 variables of water quality, geology and depth of the well. Geology was used as original 8 classification of basic survey (Clastic

sedimentary rock, Granite, Limestone, Gneiss, Semi-consolidated sedimentary rock, Unconsolidated sedimentary rock, Unconsolidated sediments, and Vesicular volcanic rock). As a result of association rules, the highest related condition to yield water suitable for drinking was the condition of shallow depth well located in the gneiss zone. Considering only geological characteristics, probability was higher in order of gneiss, followed by intrusive igneous rock and non - porous volcanic rock. For depth, the probability was higher under shallow wells and then at medium depths.

While previous studies covered small amounts of data, this study covered hundreds and thousands of units of groundwater data. Association rules were used to find environmental variables that produce water suitable for drinking. Thousands of ground water quality data will be gathered when the nation's groundwater fundamental research would be completed. Later, this study would be used the big data mining to derive meaningful findings.

5. REFERENCES

Aboytes-Ojeda, M., Castillo-Villar, K. K., Yu, T. H. E., Boyer, C. N., English, B. C., Larson, J. A., ... & Labbé, N. (2016). A Principal Component Analysis in Switchgrass Chemical Composition. Energies, 9(11), 913.

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Bartlett, M. S. (1951). The effect of standardization on a χ 2 approximation in factor analysis. Biometrika, 38(3/4), 337-344.

Cerny, B. A., & Kaiser, H. F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivariate behavioral research, 12(1), 43-47.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and data Engineering, 8(6), 866-883.

Fetter, C. W. (2018). Applied hydrogeology. Waveland Press..

Hahsler, M., Grün, B., & Hornik, K. (2005). A computational environment for mining association rules and frequent item sets.

Hamm, S. Y., Kim, K. S., Lee, J. H., Cheong, J. Y., Sung, I. H., & Jang, S. (2006). Characteristics of groundwater quality in Sasang Industrial Area, Busan Metropolitan city. Economic and Environmental Geology, 39(6), 753-770.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hashimoto, S., Fujita, M., Furukawa, K., & Minami, J. I. (1987). Indices of drinking water concerned with taste and health. Journal of Fermentation Technology, 65(2), 185-192.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics),

28(1), 100-108.

Jolliffe, I. (2011). Principal component analysis. In International encyclopedia of statistical science (pp. 1094-1096). Springer, Berlin, Heidelberg.

Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Scientific Software International.

Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31-36.

Kim, J., Ryoo, R., Lee, J., Song, J., Lee, Y. J., & Jun, H. B., (2016). Study of major mineral distribution characteristics in groundwater in South Korea. J. Korean Soc. Environ. Eng., 38(10), 566~573.

Ko, K. S., Kim, Y., Koh, D. C., Lee, K. S., Lee, S. G., Kang, C. H., Seong, H. J., & Park, W. B. (2005). Hydrogeochemical characterization of groundwater in Jeju Island using principal component analysis and geostatistics. Economic and Environmental Geology, 38(4), 435-450.

Liu, C. W., Lin, K. H., & Kuo, Y. M. (2003). Application of factor analysis in the assessment of groundwater quality in a blackfoot disease area in Taiwan. Science of the Total Environment, 313(1-3), 77-89.

Ministry of Environment, Korea. (2014). National Groundwater Background Water Quality Assessment Network in 2014.

Moon, S. K., Woo, N. C., & Lee, K. S. (2004). Statistical analysis of hydrographs and water-table fluctuation to estimate groundwater recharge. Journal of Hydrology, 292(1-4), 198-209.

Na, J., (2017). R applied multivariate analysis. Freedom academy press.

Na, J., (2017). R data mining. Freedom academy press.

Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. Water research, 39(12), 2621-2635.

Park, J. S., Yu, P. S., & Chen, M. S. (1997, January). Mining association rules with adjustable accuracy. In Proceedings of the sixth international conference on Information and knowledge management (pp. 151-160). ACM.

Tan, P. N. (2007). Introduction to data mining. Pearson Education India.

Todd, D. K. (2007). Groundwater hydrology third edition, Jhon Wiley and Sous, Third Reprint. Inc. India. 535p.

Winter, T. C., Mallory, S. E., Allen, T. R., & Rosenberry, D. O. (2000). The use of principal component analysis for interpreting ground water hydrographs.

Groundwater, 38(2), 234-246.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in quantitative methods for psychology, 9(2), 79-94.

국문 초록

지하수는 땅 속을 천천히 이동하며 물과 암석 사이의 상호작용을 통해 많은 미네랄을 함유한다. 따라서 지하수는 그 지역의 지질학적인 특징을 반영하며, 지하수 수질의 특징은 지역에 따라 다른 분포를 띈다. 이 연구의 주된 목적은 먹는 샘물에 적합한 지하수를 산출하는 지질학적인 특징을 찾는 것이다. 먼저 횡성 지역의 지하수의 특징을 파악하기 위해 다변량 분석을 시행했다. 다변량분석에는 8개 주요 이온을 이용했다(Na, K, Ca, Mg, SO_4, NO_3, Cl, HCO_3). 먼저 주성분분석을 이용한 결과 8개 이온의 상관성이 3개의 주요 주성분으로 해석 가능했기 때문에, 변수를 8개에서 3개로 줄였다. 이 세가지 주성분을 이용하여 요인 분석을 시행하여 각 주성분과 이온과의 상관성 분석을 통해 각 주성분의 의미를 알 수 있었다. 첫번째 주성분은 인공적인 오염원을 지시했고, 두번째 주성분은 탄산염 광물을 지시했고, 세번째 주성분은 나트륨과 중탄산염을 지시했다. 요인 분석을 통해 얻은 주성분의 score를 값으로 하는 새로운 행렬을 입력 변수로 하여 k-평균 군집 분석을 시행한 결과, 횡성 지역의 데이터는 4개의 군집으로 분류할 때 가장 잘 해석되는 것을 알 수 있었다. 각 군집의 주성분 값을 비교하여 군집의 특성을 파악할 수 있었다. 군집 1, 3 그리고 4는 상대적으로 덜 오염된 군집이었으며, 군집 2는 오염된 군집이었다. 군집 1은 탄산염 광물이 가장 풍부했고, 군집 4는 나트륨이 풍부했으며, 군집 3은 전반적인 이온 농도가 가장 낮은 군집으로 군집들 중 가장 먹는 샘물에 적합한

군집이었다. 이러한 군집 분석 방법은 물 고유의 수질 특성을 분석하고 이해하는데 적합한 방법이지만, 환경적인 변수를 입력 변수로 분석하기는 적합하지 않았다. 다변량분석은 수질 데이터를 얻을 수 있는 기존 관정 중에서 먹는 샘물에 적합한 관정을 찾는 방법에는 적합하다. 환경적인 변수와 수질의 관계를 이해하기 위해서 새로운 분석 방법을 필요로 했고, 본 연구에서는 연관 규칙을 사용하였다. 연관 규칙은 데이터 마이닝 방법 중 하나로, 연관 규칙을 이용하여 관정이 없는 타겟 지역에서 먹는 샘물에 적합한 물이 산출될 확률이 높은 곳을 찾을 수 있다. 횡성 지역의 수질, 관정의 심도, 지질, 토지이용도, 경사도의 다섯가지 변수를 연관 규칙의 입력 변수로 사용하였다. 그 결과, 총 1186개의 연관 규칙이 있었으며, 규칙의 조건부와 결과부 중 결과 부분이 먹는 샘물에 적합한 수질을 지시하는 연관 규칙은 42개가 있었다. 연관 규칙을 판단하는 세가지 측도를 이용하여, 가장 먹는 샘물에 적합한 물을 산출하는 조건은 편마암, 숲, 가파른 경사도 세가지 조건일 때였다. 횡성에서의 연관 규칙 결과를 검증하기 위해 전국의 1269개 데이터를 이용하여 연관 규칙을 시행하였다. 수질, 지질, 심도의 세가지 데이터를 입력 변수로 사용했다. 전체 90개의 연관 규칙이 있었으며, 그 중 먹는 샘물에 적합한 결과에 대한 연관 규칙은 8개가 있었다. 먹는 샘물에 적합한 물이 산출될 확률이 높은 환경적인 변수는 편마암, 얕은 심도의 조건이었다.

주요어 : 지하수, 지구화학 분석, 통계 분석, 다변량 분석, 연관 규칙, 수질