Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2022.28696

268 (2022) 242–253
August

Geochemical classification for bottled natural waters in China: using unsupervised and supervised machine leaching algorithm

Yao Shan, Yilan Wang*, Bo Yang, Hongtao Li, Jian Li

School of Emergency Management, North China Institute of Science and Technology, Yanjiao 101601, China, email: wangyilanuem@163.com (Y. Wang)

Received 17 May 2022; Accepted 12 July 2022

ABSTRACT

Major elements in still bottled natural water were analyzed in China. An unsupervised machine learning algorithm, principal component analysis (PCA), was used to classify the samples. The PCA result suggested four groups to discriminant the chemical composition of the samples. By using the labelled data, supervised machine learning methods, random forest, and support vector classification, were used to train models. The models were then applied on the data obtained from literature. To analyze the water–rock interactions of the samples from different groups, reverse modeling in the software PHREEQC was implemented.

Keywords: Bottled water; Chemical composition; Multivariate analysis; Principal component analysis; Random forest; Reverse modeling

1. Introduction

The bottled mineral water is of great importance to health hydration and plays a key role in people's lives for their taste, quality, convenience, and healthy considerations. The production and consumption of bottled drinking water have increased steadily in the last decades. According to the report, China has become the largest market in the world. However, the consumption per capita in China was 32 L/person (8.45 gallons per person) in 2018, which was lower than the average level of the world, 44 L/person (11.62 gallons per person). A high increasing ratio is expected in the next years.

In the European Union, the bottled water quality is regulated by Directive 80/777/EEC and Directive 96/70/EC, "natural mineral water" is water that is microbiologically wholesome, "originating in an underground water table or deposit and emerging from a spring tapped at one or more natural or bore exits". The US FDA classify natural water into spring water, well water, artesian well water, and mineral water, etc. The mineral water means that comes from an underground source and contains at least 250 ppm total dissolved solids. Minerals and trace elements must come from the source of the underground water, without adding later. In China, natural spring water is defined in national standard (GB8537-2018) for drinking natural spring water as the water flowing naturally to the surface or derived from an underground formation through boreholes, containing some minerals and trace elements, and no contaminants in a district where techniques are applied to preventing pollution. The physical parameters, including chemical composition, flow, temperature, are relatively stable. One or more parameters of Li, Sr, Zn, Se, metasilicate acid, free CO_2, total dissolved solids (TDS), should satisfy the minimum regulated levels. Concentrations of harmful trace elements, organics, and microorganisms should not exceed the limiting values.

The compositions of bottled waters result from water–rock interaction in the underground aquifers. In this process, the origin content in the water, geochemical and physical parameters of aquifer rock, flow duration, and other factors may influence the geochemical content of

* Corresponding author.

outflow. Water type, and its content are the expressions of the water–rock interaction process, and they also represent the water quality and its taste. Therefore, researchers have investigated the water type and the classifying methods all around the world.

Classification of bottled water is not a trivial task, for large variety and different usages of natural water. Classically, hydrogeochemical parameters were used in young sedimentary environment in the Netherlands [1], chlorinity and alkalinity were used as the main hydrological factors to classify the waters with the main cations and anions as the secondary consideration in Netherlands and Turkey [2], sulphate and carbonate contents as indicators for possible scaling [3].

With the development of AI algorithm and more powerful computers, it became easier to use multivariate techniques to classify water analyses. Examples of this approach use principal component analysis (PCA), Hierarchical and *K*-Means clustering [4–9]. In Portugal [10], the bottled waters are classified into three groups, including high mineralization waters, low mineralization waters, and medium mineralization waters from evaporitic origin, they are controlled by two types of process, namely deep faults circulation in metamorphic and/or magmatic environments, and rock dissolution processes. In Italy, water parameters were reduced using PCA and grouped by clustering analysis, then a classification model was built using Discriminant Analysis [11]. In Korea, major elements and stable isotopes of oxygen, carbon, and hydrogen, were successfully used to classify various types of bottled water using statistical method [12]. Grošelj et al. [13] collected bottled water samples around the Europe and trained artificial neural networks using geological origin as labels. In Nigeria, PCA was used to classify natural springs and other drinking water sources [14].

In China, bisphenol analogues and microplastic pollution were investigated [15–17]. Zhang et al. [18] collected mineral samples and found the water type changed from Ca–HCO_3 face to Ca–Mg–HCO_3 from 2011–2015. However, systematical research of water type and classifying method for the bottled water are rare. In this study, 30 randomly selected bottled waters were obtained from the public market, hydro-geochemical composition of which were measured.

Table 1
Origin of the samples and label reported data

No.	K^+	Na^+	Ca^{2+}	Mg^{2+}	pH	TDS	Origin
1	1.4–12.0	4.0–12.0	3.0–5.8	2.1–5.8		80–170	Changbai Mountains, Jilin Province
2	1.5–6.5	5.5–19.5	8.0–25.6	6.6–22.9	7.25–7.90	93.6–230.0	Changbai Mountains, Jilin Province
3	2.4–4.1	1.6–9.9	4–6.9	4–6.8	7.02–7.98	92–162	Changbai Mountains, Jilin Province
4	1.0–6.0	5.0–15.0	2.0–10.0	2.0–9.0		80.0–200.0	Changbai Mountains, Jilin Province
5	1.0–2.5	2.0–6.8	4.0–10.0	1.5–5.0		35–100	Changbai Mountains, Jilin Province
6	0.35–7.0	0.8–20.0	4.0–20.0	0.5–10.0		20.0–100.0	Changbai Mountains, Jilin Province
7	1.0–8.6	3.2–12	2.8–15.5	0.6–7.5		75–160	Foshan, Guangdong Province
8	1.1–4.9	1.8–14.8	0.9–15.6	0.8–10.2		60.5–181.3	Changbai Mountains, Jilin Province
9	0.5–10.0	1.0–15.0	2.0–15.0	0.1–10.0		50.0–180.0	Huizhou, Guangdong Province
10	0.2–1.0	6.5–9.0	20.0–30.0	4.0–9.0	7.0–8.0	50.0–400.0	Changbai Mountains, Jilin Province
11	0.1–0.5	1.0–4.7	45–65	5–11	7.3–8.2	250–360	Hechi, Guangxi Province
12	1.0–5.0	4.0–8.0	0.5–7.9	0.5–5.0		68–160	Yifeng, Jiangxi Province
13	0.5–10.0	3.0–35.0	5.0–40.0	0.5–10.0		40–300	Huangshan, Anhui Province
14	0.5–5.0	1.0–10.0	5.0–35.0	1.0–10.0		60.0–60.0	Yichun, Jiangxi Province
15	≥35	≥80	≥400	≥50	7.3 ± 0.5		Chunan, Zhejiang Province
16	≥35	≥80	≥400	≥50	7.3 ± 0.5		Jiande, Zhejiang Province
17	0.6–2.0	4.8–17.4	10.8–32.4	1.5–5.7		120–300	Huzhou, Zhejiang Province
18	0.5–10.0	1.0–25.0	2.0–35.0	0.1–15.0		50.0–250.0	Yuyao, Zhejiang Province
19		5.7	20.9	7	7	106	Nantou, Taiwan Province
20	0.5–5.0	2.0–8.0	3.0–15	0.5–6.0	7.45 ± 0.6	50–300	Luan, Anhui Province
21	0.6–2.0	4.8–17.4	10.8–32.4	1.5–5.7		120–300	Huzhou, Zhejiang Province
22		125.0–280.0				450.0–1,000.0	Kedong, Heilongjiang Province
23	0.1–1.5	100–150	0.5–3.0		8.5 ± 0.5	400–450	Baiquan, Heilongjiang Province
24	0.3–2.0	10.0–20.0	0.2–2.0	0.1–0.4	7.2–7.7	80.0–150.0	Mohe, Heilongjiang Province
25	1.0–2.4	53.0–72.0	15.0–22.0	4.0–6.8		220–580	Leshan, Sichuan Province
26	0.8–1.8	40.0–100.0	10.0–30.0	2.0–4.0	7.8 ± 0.6	200–500	Kedong , Heilongjiang Province
27	1.0–3.0	10–65	30–75	30–60	7.0–8.5	400–850	Qinghai Province
28		≥80	≥210	≥60			Puyang, Henan Province
29	0.1–5	1–80	1.5–15	0.5–5		>40	Dali, Yunnan Province
30	0.1–5.0	1.0–10.0	1.0–10.0	0.1–5.0			Xiamen, Fujian Province

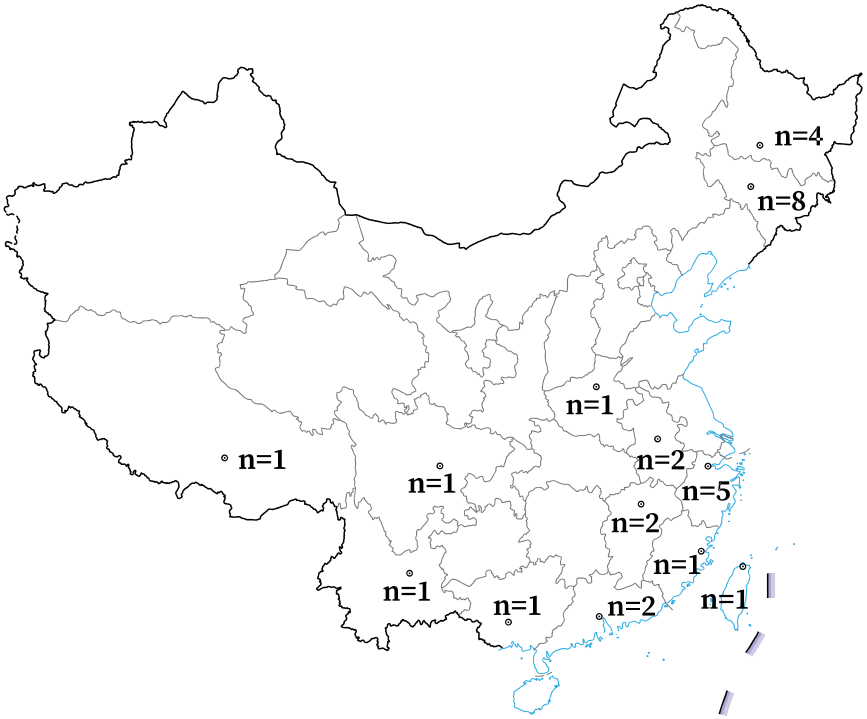

Fig. 1. Original locations of the bottled water in province scale.

The objective of this study is composed by three steps: First, the geochemical feature of the bottled natural water in China are investigated, especially in some key regions; Second, waters are classified into several groups using the method of multivariate analysis, then the group labels were used to train supervised discriminant models; Third, water–rock interaction for all groups are then evaluated and discussed using the geochemical modeling technology.

2. Materials and methods

2.1. Sampling collection and analysis

30 still bottled water commercially available from supermarkets in China were collected. According to the national standards, the bottled water should be filled and sealed at the water source, essential information, including the source of the water, total dissolved solid and concentrations of important cations ($[K^+]$, $[Na^+]$, $[Ca^{2+}]$, $[Mg^{2+}]$) should be marked. Table 1 important data copy shown on the sample bottles as a reference for the water test and analysis. Fig. 1 shows the original locations where the bottled water collected in province scale.

Major ions and physical parameters of water samples were measured complied with Chinese Standard Protocols in Jiangsu Provincial Coal Geology Research Institute. K and Na were analysed by flame atomic absorption spectrophotometry (GB 11904-89). Mg and Ca were measured using atomic absorption spectrophotometric (GB 11905-89). Fe was measured by phenanthroline spectrophotometry (HJ/T 345-2007). Sulphate and chloride were determined using flame atomic absorption spectrophotometry (GB 13196-91) and silver nitrate titration (GB 11896-89), respectively. Total dissolved solids and hardness were analysed using the standard GB/T 8538 method. pH was analysed in the laboratory using Glass electrode method (GB/T 6920-86).

2.2. Statistical methods of data analysis

The geochemical composition of a bottled water sample shows the result changes pattern occurring in the water source, which depend on water geochemistry of the aquifers, lifetime of the water, pH changes due to the system's degasification, etc. [19]. The important cations, anions, and parameters were measured to classify the water types and analyze the water–rock interaction of the water source. However, univariate statistical analysis of a large scale of data could be cumbersome and cause misunderstanding and error in the interpretation. In virtue of the thought of machine learning, a distinguish method with higher dimensions may get a better result. Based on this foundation, principal component analysis can be applied while no labelled parameters can be originally obtained. Instead of univariate statistical, the water type was calculated by machine learning in a higher dimension space. While the samples are labelled, supervised method can be used to build discriminant models. In the area of hydro-chemical studies, an unsupervised classifying method, principal component analysis (PCA) has been widely used to reduce dimensions and analyze the relations among the variates and samples [20–28]. Shan and Shi [9] reviewed the application of multivariate analysis on the source apportionment of trace elements in water and source matrix, the PCA is one of the widest used methods. Based on the correlation matrix, the PCA calculate loadings of all parameters and samples on principal components to represent information

of them. The data dimension may be reduced to three or more, depending on the need to present variance of the original data [19]. While the axis of coordinates rotated, the axis was marked as RCs. Then the matrix with principal components in lower dimensions could be clustered.

In our study, the loadings of every drawn show co-existing pattern of parameters ions, and scores of every drawn show the co-existing pattern of samples. The clustering result of loadings show similar pattern among ions and parameters. Therefore, the co-existing behavior of parameters and ions could be summarized. The clustering result of scores show similar and different pattern among samples. Therefore, the co-existing behaviour of samples, which means types of water samples, could be summarized. The clustering method was based on the Gaussian Mixture Model (GMM). The GMM algorithm fit all the groups into Gaussian distribution, then the probability that all attributes and samples belong to the groups are calculated. Comparing with *K*-means algorithm, the GM Model does not divide different group by stiff border but allow some mixture of different groups. So, the classifying probability to each group can be calculated.

By using the method of PCA and GMM clustering, the water samples could be labelled into different groups. Then the data could be used to train supervised models. In the feature selection step, a random forest algorithm implemented to identify the importance of the attributes that could classify the samples. Then the selected features were used in the supervised model training, in which the algorithms of random forest and support vector classification were implemented. The models were then applied on the data of natural water of China and other countries from the literature.

We have applied software R as tool. The packages psych and mclust were used to calculate PCA and GM model clustering result, the packages Boruta was used to select important features in the model, then the package e1071 was used to train supervised support vector classification and, package random Forest for the random forest model, respectively.

2.3. Geochemical modelling

Geochemical modelling carried out using the software PHREEQC. The following calculation were carried out: (1) speciation and saturation-index calculations; (2) batch-reaction and one-dimensional (1D) transport calculations with reversible and irreversible reactions; and (3) inverse modelling, which finds sets of mineral and gas mole transfers that account for differences in composition between waters within specified compositional uncertainty limits. The aim of this study was to find the water–rock interaction process based on the water content. Therefore, the reverse modelling module was used.

3. Result and discussion

3.1. Geochemical analysis

Table 2 and Fig. 2 show concentrations of major ions and parameters. Fig. 2a displays profiles of $[Na^+] + [K^+]$, $[SiO_3^-]$, and $[HCO_3^-] + [CO_3^{2-}]$ of all the 30 water samples. $[Na^+] + [K^+]$ shows roughly two patterns. Four samples had concentrations that higher than 80 mg/L, including three from Heilongjiang, and one from Sichuan Province. Except for the four samples, others had the mean $[Na^+] + [K^+]$ value of 4.74 mg/L. The concentrations of HCO_3^- and CO_3^{2-} also shows two-type pattern. Six samples had concentrations that higher than 200 mg/L. The others had means values of 59.7 mg/L. According to the correlation calculation, the values of $[Na^+] + [K^+]$, TDS, and $[HCO_3^-] + [CO_3^{2-}]$ had high correlation indexes, that is, 0.86 between $[Na^+] + [K^+]$ and TDS, 0.92 between $[HCO_3^-] + [CO_3^{2-}]$ and TDS, and 0.92 between $[Na^+] + [K^+]$ and $[HCO_3^-] + [CO_3^{2-}]$, respectively. Mean value and standard deviation of $[SiO_3^-]$ was 10.43 mg/L and 5.18, respectively, with the highest value of 19.10 mg/L.

Fig. 2b displays profiles of $[Ca^{2+}] + [Mg^{2+}]$, $[Cl^-]$, and $[SO_4^{2-}]$ of all the 30 water samples. This graph shows that seven out of thirty samples had concentrations higher than 20 mg/L Ca^{2+}, two samples had concentrations higher than 10 mg/L Mg^{2+}, two samples had concentrations higher than 20 mg/L Cl^-, and six samples had concentrations higher than 20 mg/L SO_4^{2-}.

Higher $[Ca^{2+}]$ was observed in the water sourced from Jilin, Guangxi, Jiangxi, Zhejiang, Sichuan, and Tibet Provinces, higher $[Mg^{2+}]$ in Jilin and Tibet Provinces, higher $[Cl^-]$ in Heilongjiang and Tibet Provinces, and higher $[SO_4^{2-}]$ in Zhejiang, Taiwan, Sichuan, Heilongjiang, and Tibet Provinces, respectively. The correlations indexes between $[Ca^{2+}]$ and $[Mg^{2+}]$, and that among $[Mg^{2+}]$, $[Cl^-]$, and $[SO_4^{2-}]$ were relatively high, suggesting a similar pathway and reaction process. On the other hand, the samples with high $[K^+] + [Na^+]$ usually had low $[Ca^{2+}]$ and $[Mg^{2+}]$ values, suggesting different reaction mechanisms.

Fig. 2c displays profiles of $[Fe^{3+}] + [NH_4^+]$, and $[NO_3^-]$ of all the 30 water samples. This graph shows that 23 out of 30 samples had concentrations lower than 0.1 mg/L Fe^{3+}, 22 samples had concentrations lower than 1 mg/L NH_4^+, two samples had concentrations higher than 20 mg/L Cl^-, and all samples had concentrations lower than 0.5 mg/L NO_3^-, respectively.

3.2. Classification of bottled water

In this study, PCA was firstly applied to a matrix of concentration of nine major ions observed in 30 bottled waters. Based on the correlation matrix, the parameters and samples were projected on three orthogonal (rotated) axes in a three-dimensional space. Eigen values of them were 2.82, 2.34, 1.58, respectively. All the other six Eigenvalues were all smaller than one.

The nine parameters were then relocated in a 3D dimensional space based on its loadings on the three axes, then the parameters were then clustered using a Gaussian mixture model (GMM), which is shown in Fig. 3a. The 30 samples were relocated in a 3D dimensional space based on its scores on the three axes, then they were clustered using the method of GMM, shown in Fig. 3b. Principal component loadings of the nine variables are shown in Table 3, and the principal component loadings of the 30 bottled waters are shown in Table 4.

As shown in Fig. 3a, the nine parameters were classified into three groups, the groups I included $K^+ + Na^+$, Cl^-, SO_4^{2-}, and HCO_3^-, which are marked in solid circles. The groups II

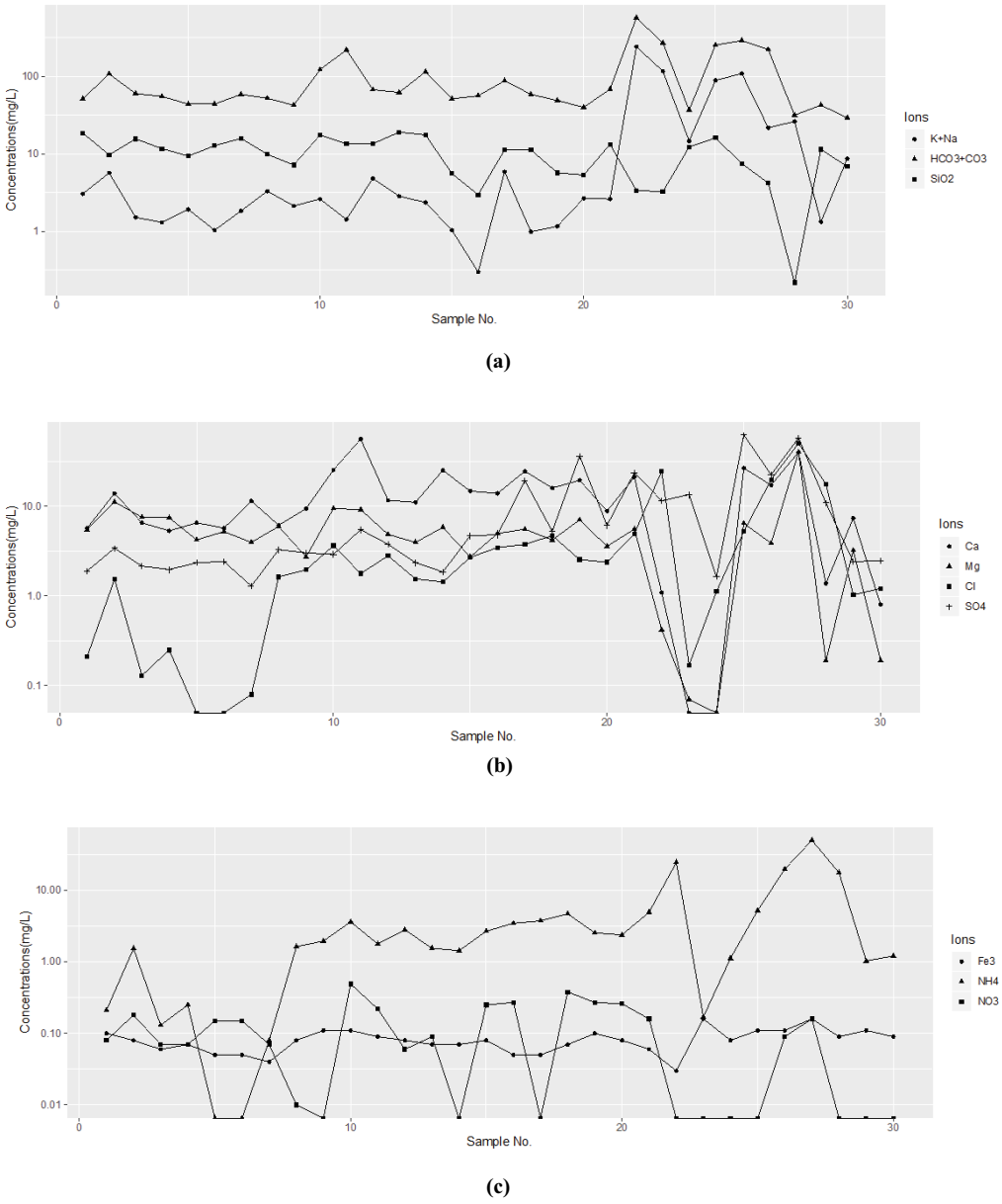

Fig. 2. Concentrations of major ions in water samples.

Table 2
Concentrations of major ions and origins of the water samples

Locations	$K^+ + Na^+$	Ca^{2+}	Mg^{2+}	Fe^{3+}	Cl^-	SO_4^{2-}	HCO_3^-	Source location
1	3.06	5.71	5.46	0.10	0.21	1.90	51.26	Jilin
2	5.70	13.89	11.14	0.08	1.54	3.40	107.40	Jilin
3	1.52	6.53	7.53	0.06	0.13	2.15	60.04	Jilin
4	1.31	5.31	7.42	0.07	0.25	1.97	55.16	Jilin
5	1.93	6.53	4.22	0.05	0.00	2.34	43.93	Jilin
6	1.04	5.71	5.20	0.05	0.00	2.42	43.93	Jilin
7	1.84	11.44	3.96	0.04	0.08	1.29	58.58	Guangdong
8	3.31	6.13	5.94	0.08	1.63	3.27	52.23	Jilin
9	2.14	9.40	2.72	0.11	1.96	3.02	42.47	Guangdong

(Continued)

Table 2 Continued

Locations	$K^+ + Na^+$	Ca^{2+}	Mg^{2+}	Fe^{3+}	Cl^-	SO_4^{2-}	HCO_3^-	Source location
10	2.62	25.33	9.41	0.11	3.63	2.90	122.04	Jilin
11	1.43	55.95	9.16	0.09	1.78	5.47	219.67	Guangxi
12	4.83	11.60	4.85	0.08	2.80	3.77	67.37	Jiangxi
13	2.85	11.02	3.96	0.07	1.55	2.35	62.00	Anhui
14	2.37	25.15	5.80	0.07	1.43	1.85	114.72	Jiangxi
15	1.04	14.71	2.72	0.08	2.69	4.68	51.26	Zhejiang
16	0.30	13.89	4.96	0.05	3.46	4.81	56.14	Zhejiang
17	5.89	24.51	5.54	0.05	3.75	19.25	87.87	Zhejiang
18	0.99	15.93	4.20	0.07	4.70	5.24	58.58	Zhejiang
19	1.17	19.44	7.04	0.10	2.54	36.08	48.82	Taiwan
20	2.67	8.82	3.57	0.08	2.38	6.11	40.03	Anhui
21	2.62	21.08	5.54	0.06	4.94	23.63	68.34	Zhejiang
22	240.70	1.09	0.42	0.03	24.54	11.48	550.16	Heilongjiang
23	116.96	0.00	0.07	0.16	0.17	13.44	246.52	Heilongjiang
24	14.67	0.00	0.00	0.08	1.12	1.64	36.61	Heilongjiang
25	88.78	26.55	6.43	0.11	5.22	62.14	245.06	Sichuan
26	108.74	17.15	3.86	0.11	19.76	22.39	277.27	Heilongjiang
27	21.71	40.02	40.13	0.16	50.18	57.78	222.11	Tibet
28	26.24	1.38	0.19	0.09	17.59	10.88	31.73	Henan
29	1.33	7.35	3.22	0.11	1.03	2.40	42.47	Yunnan
30	8.72	0.80	0.19	0.09	1.21	2.49	29.29	Fujian
	CO_3^{2-}	NO_3^-	NH_4^+	pH	TDS	CO_2	SiO_2	COD
1	0.00	0.08	0.25	7.30	112.00	9.40	18.62	0.39
2	0.00	0.18	0.26	7.48	136.00	5.98	9.74	0.47
3	0.00	0.07	0.34	6.95	70.00	12.81	15.63	0.99
4	0.00	0.07	0.29	7.32	54.00	12.81	11.73	0.83
5	0.00	0.15	0.19	7.79	56.00	8.20	9.38	0.72
6	0.00	0.15	0.19	7.91	56.00	7.69	12.77	0.87
7	0.00	0.07	0.20	6.68	116.00	15.38	15.84	0.45
8	0.00	0.01	0.50	7.63	48.00	8.54	9.88	0.06
9	0.00	0.00	0.46	6.06	24.00	23.06	7.21	0.37
10	0.00	0.49	0.22	6.85	116.00	11.96	17.55	0.56
11	0.00	0.22	2.80	7.74	200.00	12.81	13.48	0.60
12	0.00	0.06	1.27	6.60	64.00	13.67	13.52	0.52
13	0.00	0.09	1.91	6.88	64.00	11.96	19.10	0.45
14	0.00	0.00	2.16	6.04	134.00	41.00	17.57	0.22
15	0.00	0.25	0.17	7.29	112.00	4.27	5.60	0.39
16	0.00	0.27	0.09	7.69	28.00	4.27	2.98	0.64
17	0.00	0.00	0.16	6.18	102.00	32.80	11.29	0.75
18	0.00	0.38	0.46	6.23	74.00	23.06	11.31	0.22
19	0.00	0.27	0.44	7.47	82.00	4.27	5.71	0.14
20	0.00	0.26	0.01	7.03	18.00	3.42	5.36	0.99
21	0.00	0.16	2.32	6.13	118.00	25.63	13.27	0.68
22	19.21	0.00	0.56	8.46	562.00	0.00	3.38	0.68
23	24.01	0.00	0.46	8.94	564.00	0.00	3.25	0.49
24	0.00	0.00	0.49	7.11	42.00	8.24	12.32	0.29
25	8.64	0.00	0.50	8.29	354.00	0.00	16.09	0.29
26	12.00	0.09	1.08	8.20	332.00	0.00	7.50	0.82
27	0.00	0.16	0.17	7.94	368.00	15.72	4.22	0.60
28	0.00	0.00	0.50	6.84	36.00	6.83	0.22	0.14
29	0.00	0.00	1.32	7.33	66.00	6.83	11.44	0.68
30	0.00	0.00	2.28	7.16	30.00	5.13	6.87	0.45

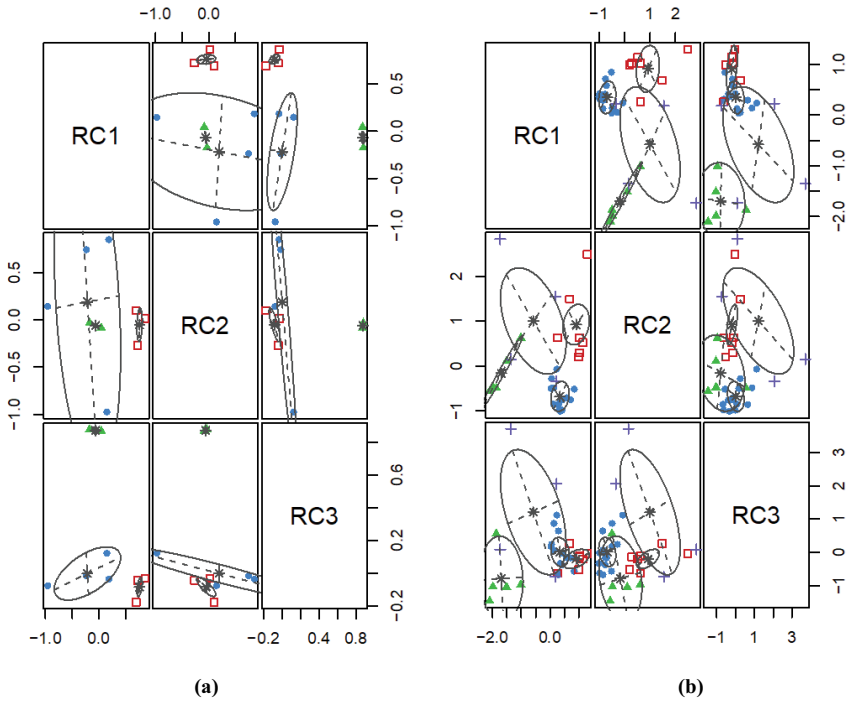

Fig. 3. Classification of major ions and samples using PCA.

Table 3
Principal component loadings of the nine variables

	RC1	RC1	RC1
$K^+ + Na^+$	–0.959	0.144	
Ca^{2+}	0.865		
Mg^{2+}	0.722	–0.270	
Fe^{3+}			0.870
NH_4^+	–0.175		0.878
Cl^-	–0.236	0.744	
SO_4^{2-}	0.184	0.852	
HCO_3^-	0.147	–0.974	0.123
NO_3^-	0.695	0.100	–0.176

included Ca^{2+}, Mg^{2+}, and NO_3^-, which are marked in hollow squares. The groups III included Fe^{3+} and NH_4^+, which are marked in solid triangles.

The sample classification result is shown in Fig. 3b. Four groups had 14, 7, 5, and 4 samples, which are marked in solid circles, hollow squares, solid triangles, and crosses, respectively. The samples in the group I are listed as first 14 in Table 1, the group II include the sample No. 15–21, the group III include the sample No. 22–26, the group IV include the sample No. 27–30. In consideration of water sources, the group I include eight from Jilin Province, two from Guangdong Province, two from Jiangxi Province, and one each from Guangxi and Anhui Province. The group II include five from Zhejiang Province, one each from Anhui and Taiwan Province, respectively. The group III include four from Heilongjiang Province, and one from Sichuan Province. The group IV include one each from Tibet, Henan, Yunnan, and Fujian Province, respectively.

Water samples show a significant geographical dependency, especially for the Jilin, Zhejiang, and Heilongjiang Province, suggesting its identity feature of the water reaction mode in these water source regions. Fig. 4 shows concentration profile of major ions in every group. Group I and II had lower K^+, and Na^+ concentrations, and higher Ca^{2+}, and Mg^{2+} concentrations than the group III and IV. Group I had higher HCO_3^- concentrations than other groups. Group 4 had relative higher Fe^{3+} and lower NO_3^- concentrations than other groups.

The four groups of samples were drawn in a piper plot (Fig. 5), which are shown using solid circles, hollow circles, solid stars, and solid triangles, respectively. It can be concluded that the group III can be divided from others based on the ion composition. According to the classification method of Shug Kalev, the water of group III belongs to Na–HCO_3 type. Group I and II are similar, with a roughly distinguishing pattern, which belong to Ca–HCO_3 type. The group IV dispersive in the piper plot.

3.3. Supervised machine learning model

The supervised machine learning algorithm was used based on two premises: a supervised ML method, rather than the unsupervised ML method can be used to build a model, which is feasible to reuse in some other scenario; the supervised ML model usually has higher precise than the unsupervised ML model. An important factor that constrains the using of supervised ML method is the difficulty to get labels of observations (samples). In this study, the observations had been grouped using a series of

Table 4
Principal component loadings of the 30 bottled waters

	RC1	RC2	RC3
1	0.291	–0.864	0.645
2	0.286	–0.854	–0.675
3	1.015	0.291	–0.140
4	0.990	0.203	–0.523
5	–2.097	–0.566	–1.467
6	0.262	0.630	–0.632
7	0.358	–1.012	–0.334
8	0.417	–0.989	–0.019
9	0.577	–0.710	–0.161
10	0.178	1.549	–0.732
11	0.708	–0.752	–0.164
12	0.312	–0.983	–0.628
13	–1.988	–0.477	–1.036
14	0.034	–0.500	0.177
15	1.138	0.520	–0.201
16	0.234	–0.068	1.122
17	–1.868	–0.495	0.559
18	1.290	2.492	–0.031
19	–1.735	2.828	0.088
20	–1.014	0.622	–0.971
21	–1.513	0.105	–1.048
22	0.221	–0.342	2.059
23	0.839	–0.516	–0.566
24	0.412	–0.761	–0.378
25	–1.357	0.152	3.716
26	1.022	0.631	–0.111
27	0.051	–0.285	0.243
28	0.134	–0.488	0.873
29	0.681	1.489	0.266
30	0.122	–0.850	0.071

unsupervised ML method, and the labels were marked on every sample.

Before the supervised ML modeling, two pre-steps were needed: selecting the observations and features. The observations should be those could be distinguished correctly and clearly. And some features that contribute the most to the distinguishing should be select. The number of the features should be suitable to the modelling. A too small feature group would miss some part of the variable characteristics, leading to wrong classification. On the other hand, too many features would lead to over-fitting problem.

The random forest classification (RFC) is an ensemble ML algorithm of multiple classification trees. Compared with simple decision trees, RFC runs efficiently on high-dimensional space data, and it is more accurate and robust to noise. At the same time, this method can handle many input variables while assessing the importance of variables. The RFC draws multiple samples based on the bootstrap resampling method from the original samples, and then constructs the decision trees model for the samples. Then the prediction output is obtained by calculating the average of all prediction trees. In order to determine the importance of cations and anions as distinguish indexes. A supervised algorithm, random forest, and a corresponding R package, Boruta, were used.

As the first three groups could be clearly grouped, they are used in the machine learning calculation, which sample number 1–26 in Table 1. The feature selection result is shown in Fig. 6. As the result showing, five parameters were found to be important, one was found to be medium important. This result suggested that the six major ions can be used as indexes to build a discriminant model.

Several algorithms could be used in the supervised ML model, such as logistic regression, artificial neural network, Bayesian network, random forest, support vector classification, etc. Considering that this task is a multi-group classification, and relatively small data size, the algorithm random forest and support vector classification were used.

For the RFC, the R package randomForest was used, and the Gini index was used to calculate for establishing and pruning of every tree. Two important parameters were defined, the number of trees in the forest (ntree), and the number of the random variables of the split nodes (mtry). When the ntree is defined too small, the RFC prediction error is large and unstable; on the other hand, too large ntree number need more computation time and memory. The default setting of the ntree is 500. By repeating operation, it is found that the model tends to be stable, while the ntree achieve 200. The parameter mtry is the number of the random variables of the split nodes. A vulgar method to determine the mtry is to calculate square root of the number of variables. While the number was six, mtry should be set to be 2 or 3. After comparison, the mtry value was set to be 2.

The SVC algorithm has superior prediction performance in various fields of data modeling for its high accuracy of prediction and low probability of over-fitting. When the data cannot be classified by linear algorithm, the data is calculated using kernel formulas to project the data to a higher-dimension space, then the data may be classified by hyperplanes. The most popular kernel includes polynomial and Radial Basis Function, which were used in this study. The R package e1071 was used in the SVC training. When using the polynomial kernel, parameters were setting to cost = 1, gamma = 1, degree = 2, coef = 1, and parameters were setting to cost = 1, gamma = 1, while using the Radial Basis Function kernel.

Distinguishing models were built using the six major ions under support vector classification algorithm and random forest algorithm. In the discriminant model calculation, the molar equivalents per cent was used. The RFC model and SVC models with two types of kernel, polynomial, and radial, could all have got 100% distinguishing result. To compare the performance of the three models, the root mean square error (RMSE) and the coefficient of determination R^2 were utilized, which is shown in Table 5. The kernel redial means Radial Basis Function was used in the SVC training. The result shows that the random forest model and polynomial kernelled SVC got better performance.

We have applied the models on data obtained from literature. Table 6 lists the molar equivalents per cent of the

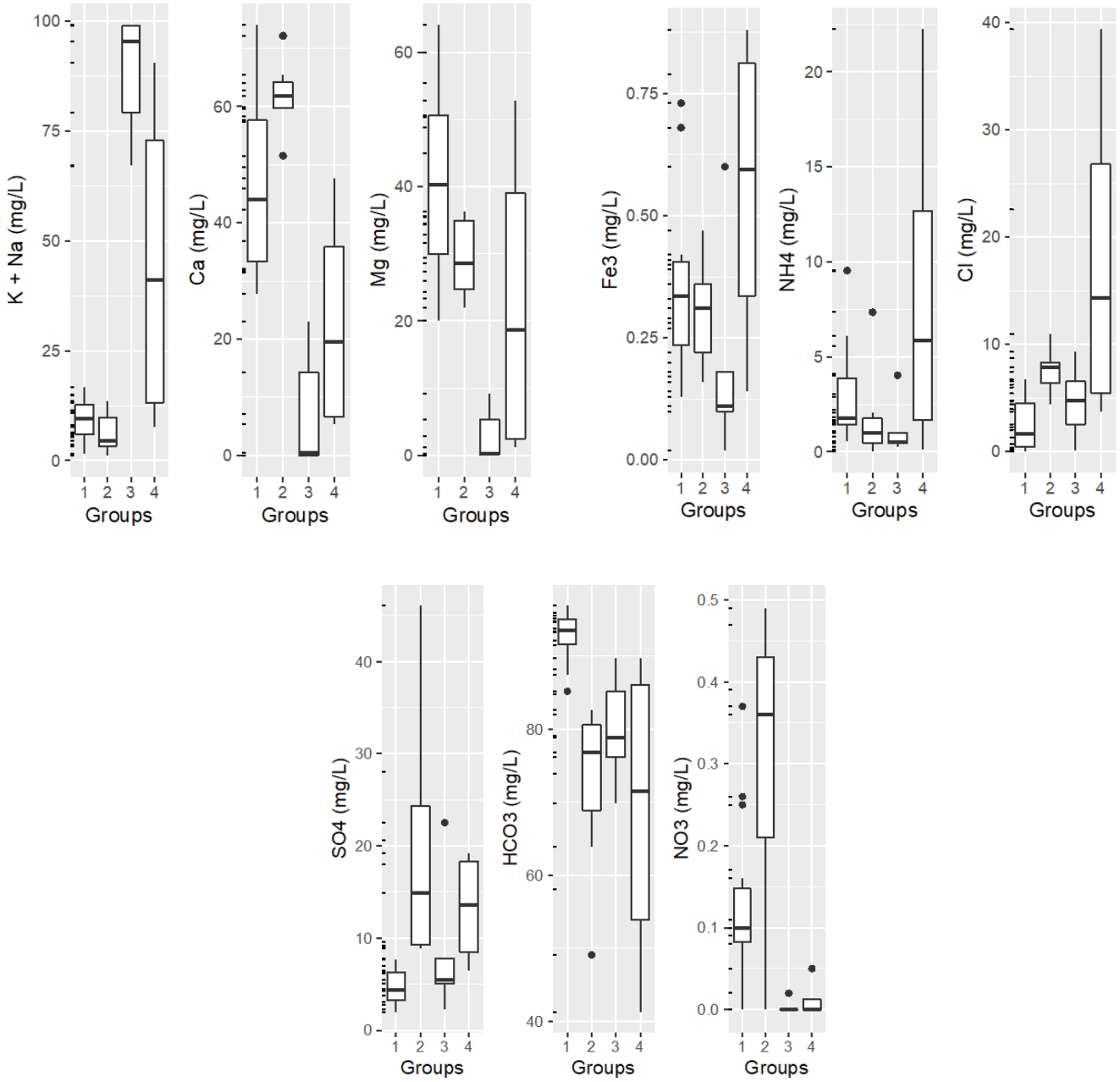

Fig. 4. Boxplot of major ions in every water type.

major ions and the discriminant result using the models of random forest and radial kernelled SVC.

By using the models, the spring waters were classified into three types. Although the model may not be precise to distinguish all types of water, the result to classify spring water was acceptable. It suggested that the spring water undergoes some limited types of water–rock interaction, or some types plays important role in the bottled natural water, or spring water.

In sight of the model application, the models can be compared further. Comparing with the SVC algorithm, the random forest shows a better performance on the classification of the three groups, and a better feasible in model generalization. The SVC gave a good criterion of RMSE and R^2, but it not suitable for outside of the training data, suggesting an over-fitting effect. Comparing with the polynomial kernel, the Radial Basis Function shows a better performance, the group I and group II can be divided successfully. However, the Radial Basis Function kernelled SVC was weak to distinguish group III from others.

3.4. Water–rock interaction processes

To analyze the water–rock interaction process, reverse modeling using PHREEQC was implemented.

Table 5
Root mean square error (RMSE) and the coefficient of determination R^2 of the discriminant models

	RFC (ntree = 200)	RFC (ntree = 500)	SVC (kernel = polynomial)	SVC (kernel = radial)
RMSE	0.099	0.097	0.111	0.254
R^2	0.984	0.984	0.980	0.895

Table 6
Molar equivalents % of the major ions and the discriminant result for the literature data

Location	K + Na (%)	Ca (%)	Mg (%)	HCO_3 (%)	Cl (%)	SO_4 (%)	RFC	SVC (kernel = radial)
Hebei	13.2	64.9	22.0	49.7	17.7	32.5	2	2
Hebei	10.9	59.3	29.8	75.8	7.8	16.5	2	2
Inner Mongolia	93.4	5.1	1.5	23.1	30.7	46.2	3	2
Inner Mongolia	93.9	5.9	0.2	25.3	29.4	45.3	3	2
Anhui	14.8	59.7	25.6	88.1	6.9	4.9	1	1
Anhui	13.8	60.9	25.3	89.1	5.7	5.2	1	1
Yunnan	23.9	43.2	32.8	92.5	3.3	4.1	1	1
Yunnan	17.2	55.5	27.3	92.4	5.1	2.4	1	1
Xinjiang	70.8	27.6	1.6	9.3	43.4	47.3	3	2
Xinjiang	71.4	24.2	4.4	15.4	42.5	42.1	3	2
Shandong	33.6	44.6	21.8	46.8	26.7	26.6	2	2
Jiangxi	12.3	50.3	37.5	96.1	1.7	2.1	1	1
Qingdao Shandong	16.1	68.6	15.3	73.1	16.0	10.9	2	2
Qingdao Shandong	19.7	62.7	17.7	44.3	24.2	31.5	2	2

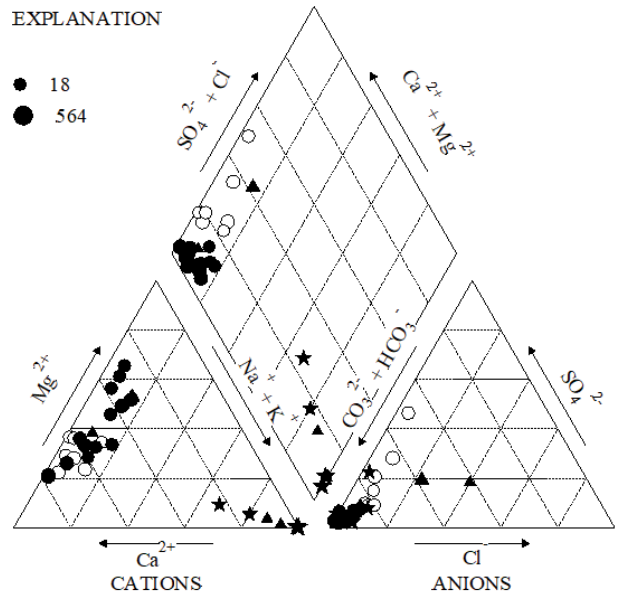

Fig. 5. Piper plot of the water samples.

Inverse modeling was used to calculate water–rock reaction modes by input water geochemistry pre and post the reaction unit and rock/mineral content. We selected one sample from group I/II/III each (sample 1 to present group I, sample 15 to present group II, and sample 13 to present group III) to present the post-unit water geochemistry. The rock should present high soluble mineral and Na/Mg/Ca containing minerals in igneous rocks. Therefore, we selected the minerals dolomite, calcite, gypsum, halite, quartz, albite, anorthite, akermanite, and pyroxene. PHREEQC 3.5.0 and database llnl.dat was used to simulate this process. Table 7 shows probable reactions according to the water and rock settings.

For group I, the minerals dolomite, gypsum, halite, quartz or anorthite, albite dissolved, and minerals akermanite and pyroxene precipitated. Group II had similar geochemistry characteristics and showed similar reaction mode. However, albite showed precipitation tendency, and some Ca/Mg containing mineral dissolved. Group III show significant higher [Na^+] comparing with groups I and II. The simulation result showed that Na^+ in water released from the dissolution of Na-containing mineral, such as albite, Ca^{2+} and Mg^{2+} tend to precipitate to form mineral such as anorthite, akermanite, Pyroxiene, or clay minerals.

In our study, the water samples both from our test and literature in group III were originated from Heilongjiang, Inner Mongolia, and Xinjiang Provinces. According to the geological analysis, they are all located in the Tianshan Xing'an orogenic system. The group I and II are located Sino-Korean paraplatform or Yangzi paraplatform. Relative

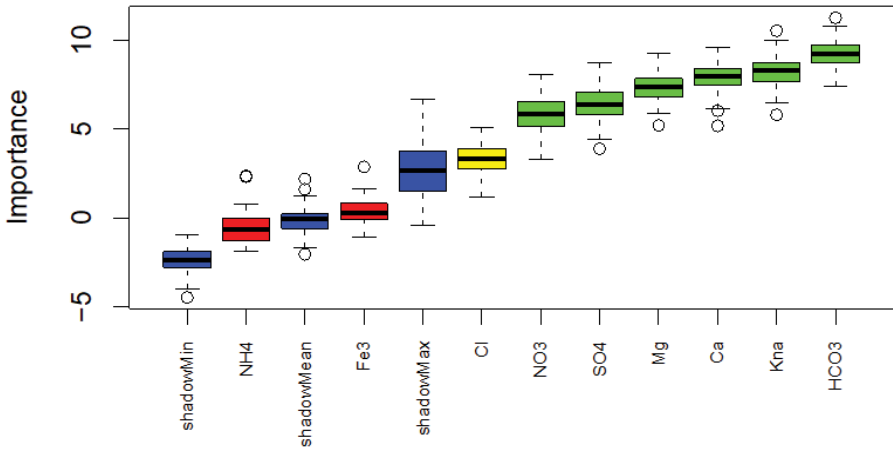

Fig. 6. Important indexes of the major ions in the supervised ML modelling.

Table 7
Reverse simulation results of hydrogeochemistry (mmol/kgw)

Group	Sample/pattern	Dolomite	Calcite	Gypsum	Halite	Quartz	Albite	Anorthite	Akermanite	Pyroxene
I	Sample 1/1	3.1×10^{-4}	–	3.9×10^{-5}	4.8×10^{-6}	4.9×10^{-4}	1.3×10^{-4}	–	-7.4×10^{-5}	-6.3×10^{-5}
I	Sample 1/2	3.1×10^{-4}	–	3.9×10^{-5}	4.8×10^{-6}	–	1.3×10^{-4}	4.9×10^{-4}	-7.4×10^{-5}	-5.6×10^{-4}
II	Sample 15/1	3.5×10^{-4}	2.6×10^{-4}	6.1×10^{-5}	1.7×10^{-4}	7.8×10^{-4}	-1.3×10^{-4}	–	-2.3×10^{-4}	6.5×10^{-5}
II	Sample 15/2	3.5×10^{-4}	2.6×10^{-4}	6.1×10^{-5}	1.7×10^{-4}	–	-1.2×10^{-4}	7.8×10^{-4}	-2.2×10^{-4}	-7.2×10^{-4}
III	Sample 13/1	7.1×10^{-4}	3.0×10^{-3}	1.4×10^{-4}	3.8×10^{-6}	–	4.9×10^{-3}	-1.1×10^{-2}	-7.0×10^{-4}	8.4×10^{-3}
III	Sample 13/2	7.1×10^{-4}	3.0×10^{-3}	1.4×10^{-4}	3.8×10^{-6}	-1.1×10^{-2}	4.9×10^{-3}	–	-7.0×10^{-4}	-2.5×10^{-3}

different Na, Ca, Mg content in the minerals led to different water geochemistry and water–rock interaction between the group I and II.

4. Conclusion

To analyze the geochemical composition, their classification behavior, and water–rock interaction in the source underground aquifers of the bottled water in China, 30 samples were obtained commercially available in market, then the analytical techniques, geochemical analysis, principal component analysis, random forest, support vector machine, reverse modeling were applied.

In summary, the bottled water can be classified into four groups. The samples in Groups I and II show similar geochemical composition with each other, which were Ca–HCO_3 type water. The water source located in the Sino-Korean paraplatform or Yangzi paraplatform, which undergo dissolution or precipitation of Na/K-containing or Ca/Mg containing mineral in the underground aquifers depending on the relative content of Na/K, and Ca/Mg in the igneous minerals. The samples in Group III showed significant higher [Na^+], [K^+] than Group I, and total dissolved solid level, these water samples were originally collected from Heilongjiang, Inner Mongolia, and Xinjiang Provinces. In the sense of geological plate, they are all included in the Tianshan Xing'an orogenic system. The water–rock interaction mode was mainly dissolution of Na/K-containing minerals, such as albite. The samples in the group IV were relatively dispersive in view of geochemical composition.

Acknowledgements

The test of samples was carried out in the Jiangsu Provincial Coal Geology Research Institute. We would like to thank all of them for their support!

Funding

Our study is funded by the Fundamental Research Funds for the Central Universities (No. 3142014005) and the Colleges and universities in Hebei Province science and technology research project (No. ZD2016204).

References

[1] P.J. Stuyfzand, Hydrochemistry and Hydrology of the Coastal Dune Area of the Western Netherlands, KIWA N.V. Research and Consultancy Division, Nieuwegein, The Netherlands, 1993.
[2] A. Firat Ersoy, H. Ersoy, Stuyfzand Hidrojeokimyasal Modelleme Sistemi: Gümüşhaciköy (Amasya) Akiferi Örneği, Jeoloji Mühendisliği Dergisi, 2008, pp. 37–51.
[3] S. El-Manharawy, A. Hafez, A new chemical classification system of natural waters for desalination and other industrial uses, Desalination, 156 (2003) 163–180.
[4] C. Lourenço, L. Ribeiro, J. Cruz, Classification of natural mineral and spring bottled waters of Portugal using principal component analysis, J. Geochem. Explor., 107 (2010) 362–372.

[5] V.H. McNeil, M.E. Cox, M. Preda, Assessment of chemical water types and their spatial variation using multi-stage cluster analysis, Queensland, Australia, J. Hydrol., 310 (2005) 181–200.
[6] Y.H. Liu, K. Zhang, Z.J. Li, Z.Y. Liu, J.F. Wang, P.N. Huang, A hybrid runoff generation modelling framework based on spatial combination of three runoff generation schemes for semi-humid and semi-arid watersheds, J. Hydrol. (Amsterdam), 590 (2020) 125440, doi: 10.1016/j.jhydrol.2020.125440.
[7] K. Yekdeli Kermanshahi, R. Tabaraki, H. Karimi, M. Nikorazm, S. Abbasi, Classification of Iranian bottled waters as indicated by manufacturer's labellings, Food Chem., 120 (2010) 1218–1223.
[8] H.G.M. Eggenkamp, J.M. Marques, A comparison of mineral water classification techniques: occurrence and distribution of different water types in Portugal (including Madeira and the Azores), J. Geochem. Explor., 132 (2013) 125–139.
[9] Y. Shan, J.J. Shi, Data Mining for Source Apportionment of Trace Elements in Water and Solid Matrix, M.A. Murillo-Tovar, H. Saldarriaga-Noreña, A. Saeid, Eds., Trace Metals in the Environment – New Approaches and Recent Advances, IntechOpen, 2019. Available at: http://dx.doi.org/10.5772/intechopen.88818, ISBN 978-1-83880-332-2.
[10] C. Lourenço, L. Ribeiro, J. Cruz, Classification of natural mineral and spring bottled waters of Portugal using principal component analysis, J. Geochem. Explor., 107 (2010) 362–372.
[11] G. Ragno, M. De Luca, G. Ioele, An application of cluster analysis and multivariate classification methods to spring water monitoring data, Microchem. J., 87 (2007) 119–127.
[12] Y.-S. Bong, J.-S. Ryu, K.-S. Lee, Characterizing the origins of bottled water on the South Korean market using chemical and isotopic compositions, Anal. Chim. Acta, 631 (2009) 189–195.
[13] N. Grošelj, G. van der Veer, M. Tušar, M. Vračko, M. Novič, Verification of the geological origin of bottled mineral water using artificial neural networks, Food Chem., 118 (2010) 941–947.
[14] I.C. Nnorom, U. Ewuzie, S.O. Eze, Multivariate statistical approach and water quality assessment of natural springs and other drinking water sources in Southeastern Nigeria, Heliyon, 5 (2019) e01123, doi: 10.1016/j.heliyon.2019.e01123.
[15] H. Wang, Z.-h. Liu, Z. Tang, J. Zhang, H. Yin, Z. Dang, P.-x. Wu, Y. Liu, Bisphenol analogues in Chinese bottled water: quantification and potential risk analysis, Sci. Total Environ., 713 (2020) 136583, doi: 10.1016/j.scitotenv.2020.136583.
[16] X.-j. Zhou, J. Wang, H.-y. Li, H.-m. Zhang, H. Jiang, D.L. Zhang, Microplastic pollution of bottled water in China, J. Water Process Eng., 40 (2021) 101884, doi: 10.1016/j.jwpe.2020.101884.
[17] X.J. Wang, Y. Zhang, M.H. Luo, K. Xiao, Q.Q. Wang, Y. Tian, W.H. Qiu, Y. Xiong, C.M. Zheng, H.L. Li, Radium and nitrogen isotopes tracing fluxes and sources of submarine groundwater discharge driven nitrate in an urbanized coastal area, Sci. Total Environ., 763 (2021) 144616, doi: 10.1016/j.scitotenv.2020.144616.
[18] Q. Zhang, X.J. Liang, C.L. Xiao, The hydrogeochemical characteristic of mineral water associated with water-rock interaction in Jingyu County, China, Procedia Earth Planet. Sci., 17 (2017) 726–729.
[19] M.J. Canto Machado, Águas minerais: Sua exploração industrial. jornadas hispano-lusas sobre as águas subterrâneas no noroeste da Península Ibérica, Corunha. Inst. Geol. e Mineiro de Espanha, Madrid, 2000, pp. 353–367.
[20] C.K. Hwang, J.-M. Cha, K.-W. Kim, H.-K. Lee, Application of multivariate statistical analysis and a geographic information system to trace element contamination in the Chungnam coal mine area, Korea, Appl. Geochem., 16 (2001) 1455–1464.
[21] K.P. Singh, A. Malik, D. Mohan, S. Sinha, Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study, Water Res., 38 (2004) 3980–3992.
[22] P. Liu, N. Hoth, C. Drebenstedt, Y.J. Sun, Z.M. Xu, Hydro-geochemical paths of multi-layer groundwater system in coal mining regions — using multivariate statistics and geochemical modeling approaches, Sci. Total Environ., 601–602 (2017) 1–14.
[23] C. Güler, M. Ali Kurt, M. Alpaslan, C. Akbulut, Assessment of the impact of anthropogenic activities on the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) using fuzzy clustering, multivariate statistics and GIS techniques, J. Hydrol., 414–415 (2012) 435–451.
[24] C. Güler, G.D. Thyne, J.E. McCray, K.A. Turner, Evaluation of graphical and multivariate statistical methods for classification of water chemistry data, Hydrogeol. J., 10 (2002) 455–474.
[25] A. Sako, O. Bamba, A. Gordio, Hydrogeochemical processes controlling groundwater quality around Bomboré gold mineralized zone, Central Burkina Faso, J. Geochem. Explor., 170 (2016) 58–71.
[26] J.E. Cortes, L.F. Muñoz, C.A. Gonzalez, J.E. Niño, A. Polo, A. Suspes, S.C. Siachoque, A. Hernández, H. Trujillo, Hydrogeochemistry of the formation waters in the San Francisco field, UMV basin, Colombia — a multivariate statistical approach, J. Hydrol., 539 (2016) 113–124.
[27] V. Carucci, M. Petitta, R. Aravena, Interaction between shallow and deep aquifers in the Tivoli Plain (Central Italy) enhanced by groundwater extraction: a multi-isotope approach and geochemical modeling, Appl. Geochem., 27 (2012) 266–280.
[28] H. Chihi, G. de Marsily, H. Belayouni, H. Yahyaoui, Relationship between tectonic structures and hydrogeochemical compartmentalization in aquifers: example of the "Jeffara de Medenine" system, south–east Tunisia, J. Hydrol., 4 (2015) 410–430.