Environmental Earth Sciences (2019) 78:612
https://doi.org/10.1007/s12665-019-8624-2

ORIGINAL ARTICLE

The multiple logistic regression recognition model for mine water inrush source based on cluster analysis

Hao Zhang[1] · Haofeng Xing[1] · Duoxi Yao[2] · Liangliang Liu[1] · Daorui Xue[3] · Fei Guo[4]

Received: 24 February 2019 / Accepted: 27 September 2019 / Published online: 13 October 2019

Abstract

Mine water inrush is one of the major geological hazards that threaten safe production in coal mines. The accurate identification of mine water inrush sources plays a vital role in mine water disaster control, and it is the key to preventing mine water inrush incidents. Ninety-three water samples were extracted from the three types of aquifers in the Qinan coal mine. The cluster analysis method was then used to analyze 82 of the original water samples, and the other 11 water samples that did not meet the requirements were removed. Then, the remaining 82 water samples were regarded as training samples, and the principal component analysis was completed. Taking the scores of the principal components as the independent variable and the types of water inrush sources as the dependent variable, the multiple logistic regression recognition model was established. Meanwhile, this recognition model was used to recognize the types of mine water inrush sources and verify the recognition accuracy for the 82 training samples. The comprehensive recognition accuracy reached 86.6%, which is much higher than the traditional recognition methods of water inrush sources. Based on cluster analysis, the multiple logistic regression recognition model fully considers the ion content measurement errors and the complex relationships between the internal ions, and this recognition model is more reasonable and improves the accuracy of water inrush source recognition. This paper provides a new method for recognizing the problem of water inrush sources, which also provides an effective basis for mine water inrush prevention and control.

Keywords Mine water inrush · Recognition of water source · Ion contents · Principal component analysis · Model validation

Introduction

Mine water hazards are one of the main geological hazards that can threaten the safety of coal mining. Coal mine water inrush often causes the partial submergence of a coal mine, causing huge economic losses and human casualties (Gui and Lin 2016; Hu et al. 2011; Wu et al. 2016). The Qinan coal mine is located in the Suxian mining area. Since the coal mine has been operational, there have been fewer occurrences of water bursting, and the water inflow has not been large. However, the Taoyuan coal mine, which belongs to the same hydrogeological unit as the Qinan coal mine, experienced a mine water inrush accident with a maximum water inflow of 29,000 m^3/h in 2013, which caused a serious flooding incident. To prevent the occurrence of similar large-scale mine water inrush accidents in the Qinan coal mine, it is necessary to carry out mine water prevention and control work. Among these tasks, the accurate judgment of mine water inrush sources is a prerequisite of coal mine water inrush prevention and control work, as well as an important part of preventing mine water inrush accidents (Ganyaglo et al. 2011; Zhang et al. 2017).

For a long time, experts and scholars had proposed many methods for judging water inrush sources in the problem of "Recognition of mine water inrush". Water inrush source recognition methods include geological analysis, hydrodynamic analysis, hydrochemical characteristics analysis,

✉ Haofeng Xing
hfxing@tongji.edu.cn

[1] Department of Geotechnical Engineering, Tongji University, Shanghai 200092, China

[2] College of Earth and Environment, Anhui University of Science and Technology, Huainan 232001, China

[3] Shanghai Municipal Engineering Design Institute (Group) Co., Ltd, Shanghai 200092, China

[4] Institute of Crustal Dynamics, China Earthquake Administration, Beijing 100085, China

water temperature, water level dynamic observation and geophysical prospecting (Biswas and Sharma 2017; Farnham et al. 2000; Panagopoulos et al. 2016; Keskin et al. 2015). Among them, the hydrochemical characteristics analysis method is a simple and effective way to identify a mine water inrush source. Li et al. (2017) used a hydrochemical approach to ascertain the mine water sources and to locate the potential seawater inrush seepage channels in the Xinli Mine. Wei et al. (2015) identifies a water source by analyzing hydrochemical characteristic ions at water inrush points. But, at present, the methods of multivariate statistical analysis have been relatively mature. And the following methods are applied mostly in multivariate statistical analysis: principal component analysis reduces an original set of variables into a smaller number of uncorrelated components without losing much information (Jolliffe 2002; Kim et al. 2005; Meglen 1992; Qian et al. 2016), cluster analysis can measure the similarities among samples (Bu et al. 2010; Reghunath et al. 2002), discriminant analysis (includes Distance discriminant, Fisher discriminant and Bayes discriminant) can establish an intuitive discriminant relation (Chen et al. 2009; Huang and Chen 2011; Lu et al. 2012; Huang and Wang 2018). Xu et al. (2012) selected six sets of ions ($K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^-) and their total dissolved solids (TDS) as discriminant factors for designing a GRA–SDA coupled model. Liu et al. (2013) proposed a Fisher recognition analysis for identifying a coal mining inrush water source under mining-induced disturbances. A comprehensive identification model combining hydrochemistry analysis, water source detection, and water channel exploration was proposed by Liu et al. (2018). Based on the constant ion content test results, including the pH values and total dissolved solid (TDS), Yin et al. (2006) used systemic clustering and stepwise distinguishing to analyze the sources of the inrush water in the Wanbei Mining area. The PCA–BP neural network model, based on laser-induced fluorescence technology, was also used to identify a water inrush source by Wang et al. (2017a, b).

However, the current recognition methods of mine water inrush sources did not fully consider the measurement errors of the ion content caused by external factors and have ignored the complex relationships between the ions. These recognition methods have certain deviations from the actual identification process for mine water inrush sources. Therefore, the objectives of this paper are to propose a new method for the accurate identification of mine water inrush source, and it is the multiple logistic regression recognition model based on cluster analysis, which fully considers the measurement errors of the ion content and the complex internal relationships of ions. This method uses cluster analysis to measure the similarities among original water samples, and its purpose is to screen the original water samples. Principal component analysis is used to extract the information of hydrochemical indexes, and multiple correlated indicator variables are converted into new independent sample indicators. The multiple logistic regression recognition model can predict and classify based on existing water samples. So it can effectively extract the variation information of the original water samples, eliminate the influence caused by the superposition of the information among variables and realize the recognition of mine water inrush sources. In addition, the recognition model was applied to the water samples to be discriminated to verify its accuracy. The results show that the multiple logistic regression recognition model based on cluster analysis has high accuracy. And it is easy to operate in the actual water source discrimination process, with straightforward discrimination results.

Hydrogeological conditions in the study area

The Qinan coal mine is located in the middle of Huaibei plain, and it is distributed in the Huaihe River valley, positions shown in Fig. 1. The Huihe River, a tributary of the Huaihe River, flows through the mining area, and It has high vegetation coverage. The study area belongs to the north temperature monsoon region ocean—continental climate and has distinctive four seasons. As a typical central plain climate, the annual average temperature and annual average precipitation are about 14.6 °C and 756 mm, respectively. Rainfall is concentrated in July and August. The evaporation capacity is higher than the precipitation, and the annual average relative humidity is 71%.

The Qinan coal mine is located in the southwest region of the Sunan syncline. The inclination of the strata in the northern coal mine is steep, generally ranging from 20° to 30°. However, the inclination of the strata in the middle and eastern regions of the Qinan coal mine is gentler, generally ranging from 7° to 15°. The coal-bearing strata within the study area are covered by a loose layer from the Cenozoic period.

The groundwater regime in the mining areas of Qinan consists of four subsystems: the loose aquifer of the Cenozoic, the coal-bearing sandstone fissure aquifer of the Permian, the limestone-karst fissure aquifer in the Taiyuan formation of the Carboniferous and the limestone karst fracture aquifer of the Ordovician. The hydrogeological characteristics of the aquifer and aquifuge are shown in Fig. 2.

Among them, the limestone karst fracture aquifer of the Ordovician is furthest away from the coal seam. Thus, under normal conditions, there is no direct water filling effect on the coal mine.

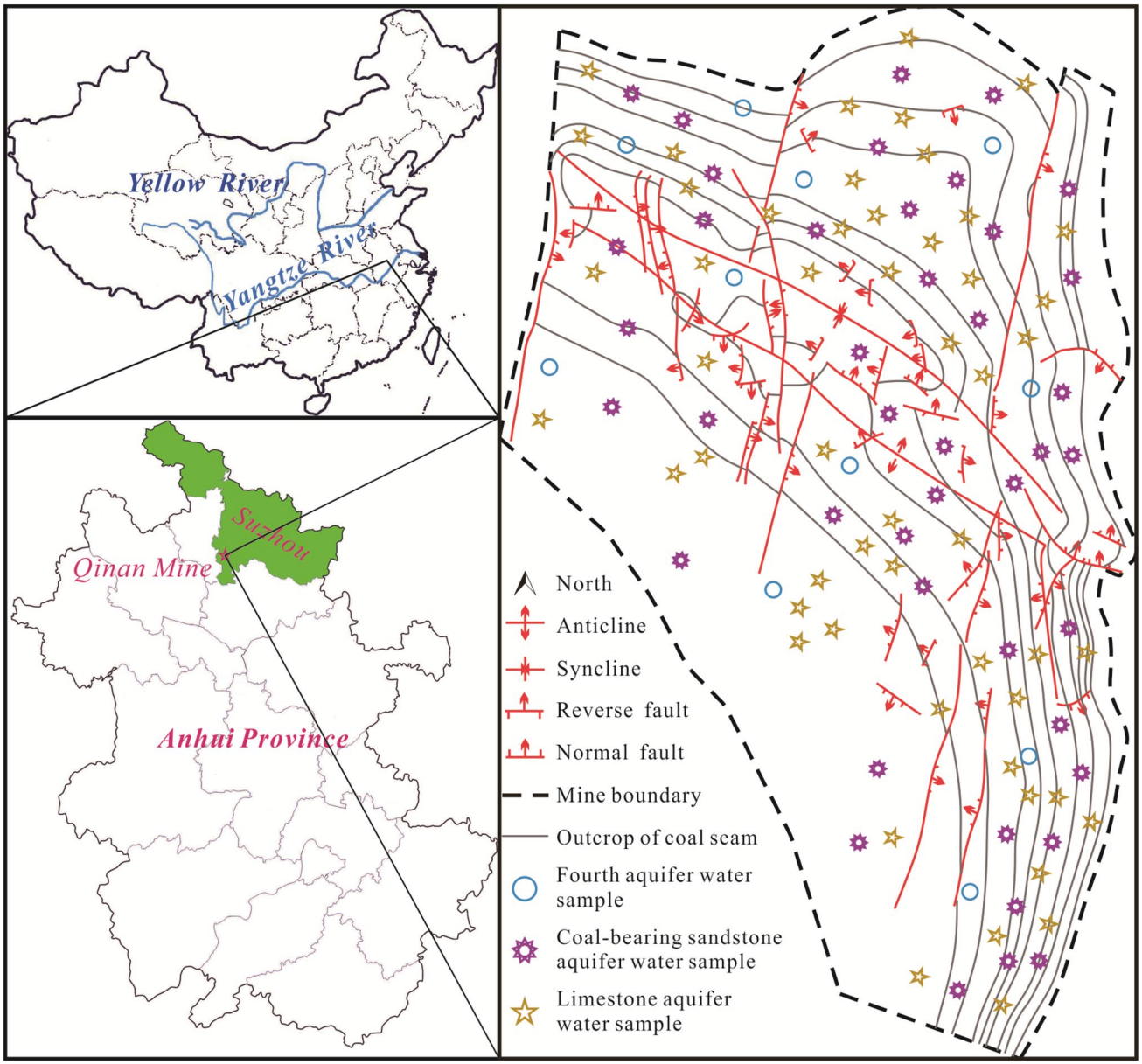


Fig. 1 Map showing the location of the study area, with the distribution of faults, folds, and sampling sites within the Qinan coal mine

Materials and methods

Sampling and test

We collected a total of 93 original water samples from the Qinan coal mine during the period of 2000–2017 (positions shown in Fig. 1), and the 93 original water samples were evenly distributed between 2000 and 2017. These water samples were used to establish the recognition model. Among them, there were 9 water samples of the fourth aquifer in the loose layer of the Cenozoic (referred to as "the fourth aquifer"), 39 water samples of the coal-bearing sandstone fissure aquifer of the Permian (referred to as "the coal-bearing sandstone aquifer") and 45 water samples of the limestone-karst fissure aquifer in the Taiyuan formation of the Carboniferous (referred to as "the limestone aquifer"). In addition, 16 water samples from the Qinan mining area were taken from the site for verification model. Among them, there were 2 water samples of the fourth aquifer, 4 water samples of the coal-bearing sandstone aquifer and 10 water samples of the limestone aquifer.

When the water samples were collected, plastic bottles and covers were rinsed three to five times using sampling water. Later, water samples were stored in a clean 550 ml plastic bottle. Before the test, the water samples were processed at low temperature to inhibit the redox

Chronostratigraphic System			Histogram	Thickness(m) Min~Max / Average	Position	Hydrogeological Feature
Erathem	System	Formation				
Cenozoic Erathem	Neogene System			0~17.3 / 12	The fourth aquifer	According to the pumping test data, the unit water inflow q_{91} is equal to 0.00044~2.389 L/s·m, and the water-rich is weak to strong. The permeability coefficient K is equal to 0.011~29.553 m/d, and the water conductivity is weak to strong. The salinity is 1.955~1.841 g/L, and the water quality type is SO_4^{2-}·Cl^--Ca^{2+}·Na^+
Paleozoic Erathem	Permian System	Lower Shihezi Formation		210.5~267.5 / 223.5	The aquifer between 6 coal and 9 coal	According to the pumping test data,the unit water inflow q_{91} is equal to 0.0004~0.002L/s·m, and the water-rich is weak. The permeability coefficient K is equal to 0.00992m/d, the salinity is 0.759~1.721g/L, and the water quality type is SO_4^{2-}·HCO_3^--Na^+
					The aquifuge between 9 coal and 10 coal upper	The lithology is generally mudstone and siltstone. The lithology of layer is dense, and the impermeability is better
		Shanxi Formation		118~136 / 127	The aquifer between 10 coal upper and lower	According to the pumping test data, the unit water inflow q_{91} is equal to 0.000091~0.001L/s·m,the permeability coefficient K is equal to 0.000208~0.00321m/d, and the salinity is 0.542~2.272g/L, and the water quality type is SO_4^{2-}·Cl^--Ca^{2+}·Na^+ or SO_4^{2-}·HCO_3^--Na^+
					The aquifuge between 10 coal lower and limestone upper	The lithology is generally mudstone and siltstone,with 1 to 2 sandstone, and it is relatively dense. In general, the impermeability is better
	Carboniferous System	Taiyuan Formation		52.7~62.2 / 60	The limestone karst fissure aquifer in Taiyuan formation	According to the pumping test data,the unit water inflow q_{91} is equal to 0.06919~0.4233L/s·m, and the water-rich is weak to medium. The permeability coefficient K is equal to 0.1853~1.8755m/d, and the water quality type are SO_4^{2-}·Cl^-·HCO_3^--K^++Na^+ and SO_4^{2-}·Cl^--K^++Na^+·Ca^{2+}

Mudstone, Sildstone, Packsand, Limestone, Magmatite
Clay, Clayey Sand, Fine Sand, Coal Seam, Pebbly Sandstone

Fig. 2 The synthesis column map of aquifer and aquifuge

reaction and biochemical action (Chen et al. 2013; Faghih Nasiri et al. 2018). The conventional water chemistry tests include the contents of $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^- and CO_3^{2-}. Among them, HCO_3^- and CO_3^{2-} were tested by dilute sulfuric acid-methyl orange titrimetry, Cl^- and SO_4^{2-} were tested by ion chromatography, Ca^{2+} and Mg^{2+} were tested by EDTA titration method and $K^+ + Na^+$ was tested by flame atomic absorption spectrophotometry. It was known from hydrogeological data that the 109 water samples were taken from drain holes, hydrogeological observation wells, extracting coal faces and underground roadways. The water levels of the observation wells did not show any abnormal changes during the collection of water samples. As such, this study was only concerned with the 109 water samples from a static perspective. The water sample data are shown in Table 1. Among them, X_1, X_2, X_3, X_4, X_5, X_6 and X_7 represent the contents of $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^- and CO_3^{2-}, respectively.

Cluster analysis

The principle of cluster analysis is that n different samples are regarded as n different classes, and the two classes with the closest properties (or the shortest distance) can be merged into the same class. Then, the next two classes with the closest properties (or the shortest distance), from the $n-1$ classes, are combined. This process continues until all the samples have been merged into a single class. In the cluster analysis, we usually divide it into Q-type cluster analysis and R-type cluster analysis based on the differences of classification objects. And Q-type cluster analysis is the classification of samples, while R-type cluster analysis is the classification of variables. The basic algorithm steps of cluster analysis are shown below:

1. At the beginning, each sample is a separate class, and the distance matrix between two pairs of n classes is calculated, denoted as:

$$D_0 = \begin{bmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ \vdots & \vdots & \vdots & & \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 0 \end{bmatrix}.$$

2. Find the minimum distance value d_{ij} in the distance matrix, and denoted as $d_{i_1 j_1}$, and combine the i_1 and j_1 classes into the $n-1$ class.
3. Calculate the distance between class $n-1$ and other classes;
4. Merge rows i_1, j_1 in the initial distance matrix D_0 into new row, and columns i_1, j_1 into new column, the number of classes is reduced by one. We can get the new distance matrix D_1.
5. Repeat steps (2) (3) and (4) until n samples are clustered into one class.
6. The clustering process was made into a cluster analysis diagram. And the original samples were screened according to the cluster analysis diagram to eliminate the samples that did not meet the requirements.

Principal component analysis

Principal component analysis is a method for original data compression and characteristic information extraction. It can replace many correlated variables with several comprehensive variables. These comprehensive variables not only express a great amount of information of the original variables but can also remain mutually independent (Jolliffe 2002; Kim et al. 2005; Meglen 1992; Qian et al. 2016; Huang et al. 2019). The basic principle is:

If $X_1, X_2, \ldots, X_n$ are defined as the original variables and Y_1, $Y_2, \ldots, Y_m$ $(m \le n)$ are new variables, the relationship between the original and new variables is

$$\left.\begin{aligned} Y_1 &= A_{11}X_1 + A_{12}X_2 + \cdots + A_{1n}X_n \\ Y_2 &= A_{21}X_1 + A_{22}X_2 + \cdots + A_{2n}X_n \\ &\cdots \\ Y_m &= A_{m1}X_1 + A_{m2}X_2 + \cdots + A_{mn}X_n \end{aligned}\right\}, \quad (1)$$

where Y_i is independent from Y_j $(i \neq j;\ i, j = 1, 2, \ldots, m)$. Y_1 is the item with the highest variance in all linear combinations of $X_1, X_2, \ldots, X_n$, and Y_2 is the item with the maximum variance in all linear combinations of $X_1, X_2, \ldots, X_n$ independent from Y_1. The rest can be performed in the same way. The new variables $Y_1, Y_2, \ldots, Y_m$ are the first, second, …, and the m principal component of the original variables $X_1, X_2, \ldots, X_n$.

Multiple logistic regression analysis

Among n multiple logistic regression analysis models, assuming P_i $(i = 1, 2, \ldots, n)$ is the probability of the sample belonging to the type i $(i = 1, 2, \ldots, n)$. Taking the reference type that the sample belongs to type n, the multiple logistic regression analysis models are as follows (Wang 2010; Wang and Guo 2001; Zhang 2002):

$$\left.\begin{aligned} G_1 &= \ln \frac{P_1}{P_n} = A_1X_1 + A_2X_2 + \cdots + A_tX_t + C \\ G_2 &= \ln \frac{P_2}{P_n} = B_1X_1 + B_2X_2 + \cdots + B_tX_t + D \\ &\cdots \\ G_n &= \ln \frac{P_n}{P_n} = 0 \end{aligned}\right\}. \quad (2)$$

Table 1 The water inrush source samples from the Qinan coal mine

N	X_1	X_2	X_3	X_4	X_5	X_6	X_7	N	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1 (The fourth aquifer)								3 (The limestone aquifer)							
1	63.6	63.4	34.2	162.4	22.7	32.1	457.9	49	258.1	141.5	82.3	511.1	257.3	600.9	418.8
2	46.4	48.1	26.3	15.4	10.7	394.1	0	50	249.5	177.9	95.9	526.4	261.6	627.3	449.1
3	189.8	122.7	84.9	400.2	192.6	486.4	359.8	51	236.0	210.8	91.9	542.0	252.9	672.9	451.6
4	233.8	143.8	102.9	203.3	648.7	402.7	0	52	250.2	223.4	89.3	565.5	261.6	699.7	463.8
5	76.71	91.1	60.5	61.7	220.6	399.3	0	53	197.4	189.9	95.9	523.9	265.9	648.3	448.1
6	207.9	33.8	65.3	309.6	113.4	357.7	292.9	54	257.5	192.8	90.8	545.7	256.6	694.4	418.8
7	49.8	21.6	35.8	17.9	36.2	258.7	19.2	55	259.0	202.6	87.5	550.8	254.3	704.7	417.4
8	581.3	94.3	67.6	74.9	1343.5	333.1	0	56	284.9	150.1	85.9	524.4	257.9	622.3	422.3
9	712.6	24.8	3.3	135.3	1232.7	123.9	31.7	57	314.6	78.2	41.0	227.4	309.5	495.5	0
A1	86.7	95	56.5	63	218	389.32	0	58	297.6	99.9	68.7	245.7	446.6	449.1	0
A2	197.9	29	61.3	319	113	357.74	292.9	59	261.3	100.8	85.1	242.3	448.2	441.8	0
2 (The coal-bearing sandstone aquifer)								60	261.3	100.8	85.1	242.3	448.2	441.8	0
10	602.6	6.7	2.5	611.7	100.7	342.4	993.1	61	353.6	101.9	67.9	239.5	455.2	444.2	0
11	463.4	3.5	1.5	528.3	98.7	9.1	922.8	62	192.5	103.2	67.6	234.4	269.2	422.3	0
12	368.2	4.6	3.0	376.5	110.7	6.2	809.4	63	349.5	110.7	71.4	239.6	591.9	461.3	0
13	136.9	65.9	29.9	232.7	85.1	89.3	453.9	64	295.1	80.4	52.9	219.3	414.1	390.5	0
14	321.6	3.9	2.6	331.2	116.1	2.5	628.8	65	357.4	19.7	9.1	248.9	199.6	374.8	0
15	319.5	4.8	3.2	336.9	127.2	13.2	664.9	66	259.4	183.7	94.9	262.3	673.0	422.4	0
16	438.1	10.8	4.5	454.7	254.1	236.8	418.4	67	229.9	203.9	82.7	253.1	644.6	393.0	0
17	438.5	10.8	4.6	455.2	253.2	237.9	422.3	68	284.3	157.1	87.1	256.3	646.2	410.7	0
18	437.7	11.2	4.9	455.1	254.9	237.9	414.9	69	347.2	20.0	12.2	226.0	116.9	500.4	4.9
19	467.8	12.0	5.7	486.4	251.6	286.9	425.9	70	277.9	189.2	75.4	255.1	650.3	436.3	0
20	418.9	9.8	4.4	439.3	245.9	205.8	412.5	71	295.4	192.5	73.4	251.1	638.4	497.3	0
21	302.8	7.4	2.6	313.2	78.14	34.6	593.1	72	483.3	45.6	13.8	225.5	245.3	793.8	0
22	261.8	20.2	15.5	297.9	117.3	12.4	510.1	73	214.9	173.6	91.8	247.6	611.6	358.1	0
23	261.0	29.4	26.0	316.9	134.6	66.3	520.9	74	157.7	247.3	109.1	255.4	677.9	422.0	0
24	399.1	8.9	3.1	411.6	232.8	197.6	378.3	75	162.2	251.2	105.5	253.6	680.8	428.9	0
25	429.4	6.9	4.9	442.1	254.8	233.4	360.2	76	151.8	248.3	107.7	253.6	667.2	419.7	0
26	440.3	66.2	28.88	536.9	246.1	494.7	468.6	77	174.1	238.1	110.5	256.2	685.7	433.4	0
27	396.6	36.8	24.3	459.8	250.5	402.5	351.5	78	161.8	239.5	112.0	251.9	679.1	428.9	0
28	202.5	16.2	11.6	321.7	160.1	7.41	290.5	79	166.6	244.4	107.3	252.7	684.5	424.3	0
29	293.5	6.8	6.9	307.8	134.1	123.07	446.7	80	164.7	237.5	110.3	257.1	676.7	415.5	0
30	668.2	0	2.8	84.1	85.2	578.5	473.0	81	173.7	240.5	106.1	252.7	693.6	413.4	0
31	375.9	104.2	50.0	232.6	561.0	454.0	0	82	279.5	19.9	14.6	220.9	27.2	463.8	0
32	250.5	13.6	11.9	89.3	3.7	585.7	10.7	83	28.0	24.6	18.6	220.8	49.4	425.4	23.0
33	282.2	8.9	4.4	88.3	6.6	583.6	26.3	84	268.9	21.8	15.1	225.4	8.2	418.4	23.7
34	306.8	4.7	2.4	87.4	12.8	594.2	39.8	85	28.0	24.6	18.6	220.8	49.4	425.4	23.0
35	264.5	11.0	6.7	80.7	17.7	581.1	13.3	86	391.6	19.8	13.0	221.6	158.9	512.4	34.1
36	578.1	44.4	21.5	157.7	309.9	1114.3	0	87	370.0	10.3	4.8	152.5	140.4	547.0	25.0
37	525.4	15.2	17.5	163.1	119.8	923.2	85.1	88	180.5	102.3	82.6	68.0	564.7	317.3	26.4
38	528.9	7.2	3.0	181.7	407.5	442.4	82.8	89	371.7	73.6	49.0	213.4	514.9	440.8	0
39	285.5	91.4	65.1	214.2	402.1	485.8	0	90	311.4	18.2	15.4	206.4	113.2	463.9	0
40	239.6	8.0	7.0	91.2	27.6	468.2	19.6	91	160.2	185.8	105.0	221.6	630.6	337.3	0
41	242.5	8.9	12.2	144.4	2.9	449.6	18.9	92	30.9	203.4	175.3	235.8	652.4	346.8	0
42	347.1	40.3	33.2	213.7	417.4	314.9	0	93	210.0	211.4	87.5	248.2	630.6	337.3	0
43	238.5	24.6	22.1	63.4	78.6	576.0	16.2	A7	276.9	214	81.5	278	667	428.1	5.17
44	241.8	8.7	17.3	59.9	44.9	541.9	26.4	A8	254.5	198	78.1	243	614	430.5	5.32
45	366.0	11.9	3.9	138.7	215.7	515.0	0	A9	270.7	215	82.9	265	606	428.1	4.19
46	248.9	19.8	13.4	68.8	57.2	549.3	25.0	A10	258.4	201	79.3	255	619	394.6	4.53

Table 1 (continued)

N	X_1	X_2	X_3	X_4	X_5	X_6	X_7	N	X_1	X_2	X_3	X_4	X_5	X_6	X_7
47	200.0	15.0	23.0	75.8	36.6	457.0	29.5	A11	269.9	209	80.5	262	608	425.7	5.95
48	248.3	8.7	13.5	97.0	0.4	507.8	38.6	A12	256.4	201	76.1	248	596	428.1	5.36
A3	264.0	29	26.4	317	135	66.34	520.8	A13	386.5	109	46.4	252	349	576.4	5.25
A4	379.1	9	3.1	412	233	197.6	378.3	A14	371.4	98	45.4	231	376	552.5	5.38
A5	459.4	7	4.9	442	255	233.4	360.2	A15	327.4	134	58.3	241	404	516.6	4.96
A6	427.3	66	28.9	537	246	494.7	468.6	A16	287.2	167	79.2	244	600	428.1	4.89

X_1, X_2, X_3, X_4, X_5, X_6 and X_7 represent the contents of $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^- and CO_3^{2-}, respectively

Because the sum of the probabilities that the samples belong to n types is 1, so we could get Formula (3):

$$P_1 + P_2 + \cdots P_n = 1. \tag{3}$$

Simultaneous Formulas (2) and (3), we were then able to derive the following Formula (4):

$$\left.\begin{aligned} P_1 &= \frac{e^{G_1}}{1 + e^{G_1} + e^{G_2} + \cdots + e^{G_{n-1}}} \\ P_2 &= \frac{e^{G_2}}{1 + e^{G_1} + e^{G_2} + \cdots + e^{G_{n-1}}} \\ &\cdots \\ P_n &= \frac{1}{1 + e^{G_1} + e^{G_2} + \cdots + e^{G_{n-1}}} \end{aligned}\right\}, \tag{4}$$

where, P_1, P_1, …, P_n are the probability functions of the respective recognition models of types 1, 2, …, n; X_i represent the value of independent variables; A_i and B_i represent the coefficient of constant ion contents, respectively.

Establishment and verification of recognition model

The sequence of steps taken using the water inrush source recognition methodology is described as follows (Fig. 3).

Q-type cluster analysis of the original water samples

To reduce the deviation of ion content caused by the external factors, for example, polluted water samples, large water evaporation because of poor sealing of the container, and the measurement deviation caused by human error in the testing process, the ion content of the original water samples were used as the analysis variables, we used MATLAB to complete the Q-type cluster analysis of the 93 original water samples. The results of the cluster analysis are shown in Fig. 4.

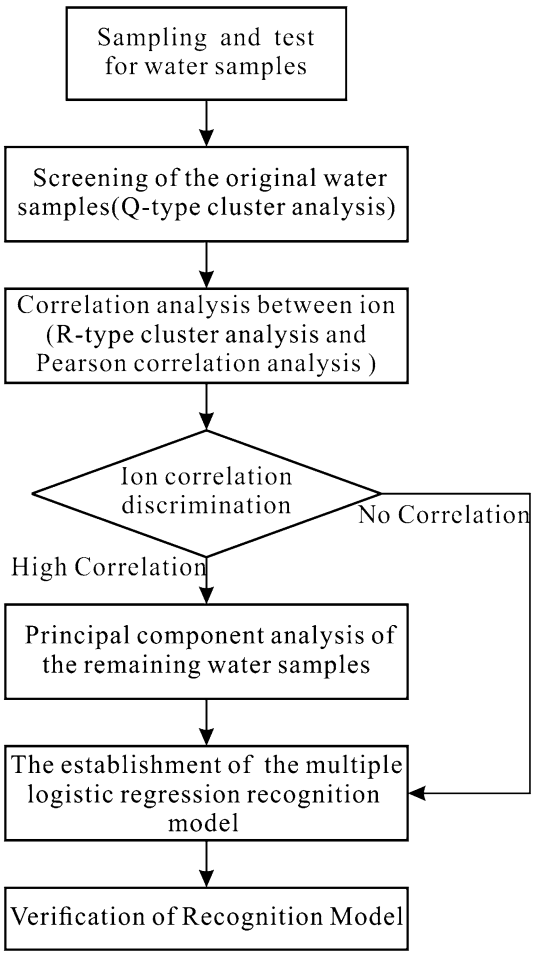


Fig. 3 Flowchart of mine water inrush source recognition methodology

From the results of Q-type cluster analysis of the original water samples displayed in Fig. 4 and according to the distance between the original water samples (Güler et al. 2002). We can re-classify the original water samples and get new classification results. Among them, there are differences between the new classification results and the original classification results for 93 original water samples, and those water samples are 1, 4, 13, 31, 38, 39, 42, 45, 72, 82 and

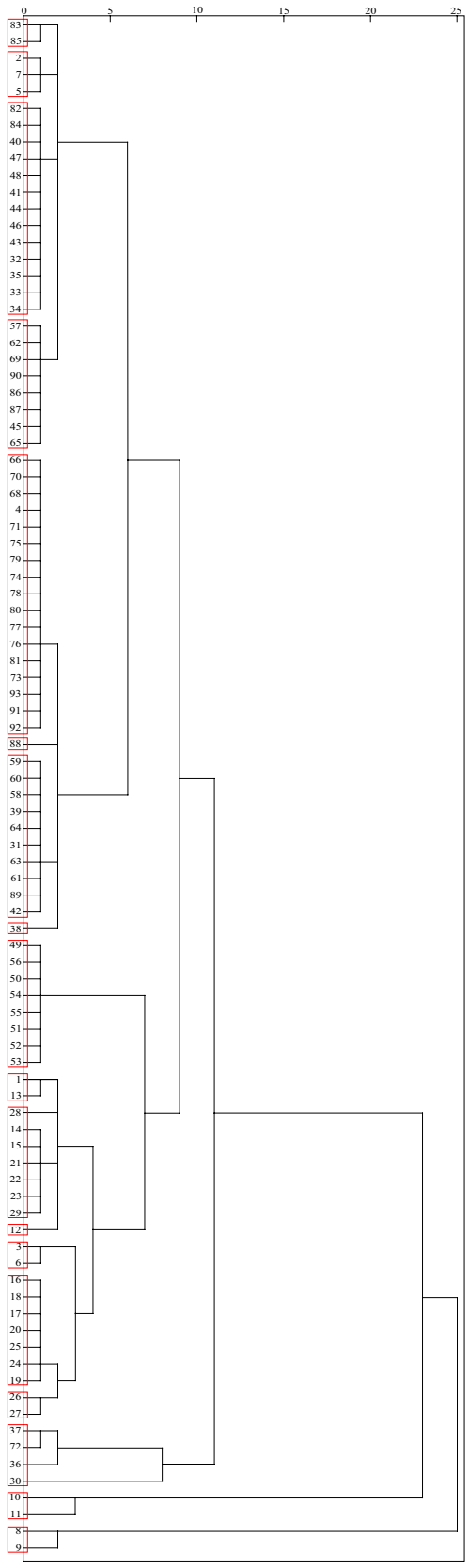


Fig. 4 Q-type cluster analysis diagram

84. In the process of discrimination, these water samples will have an impact on the results of discrimination, so we eliminate these water samples which are not consistent with the original classification results, and improve the accuracy of discrimination.

R-type cluster analysis of training samples

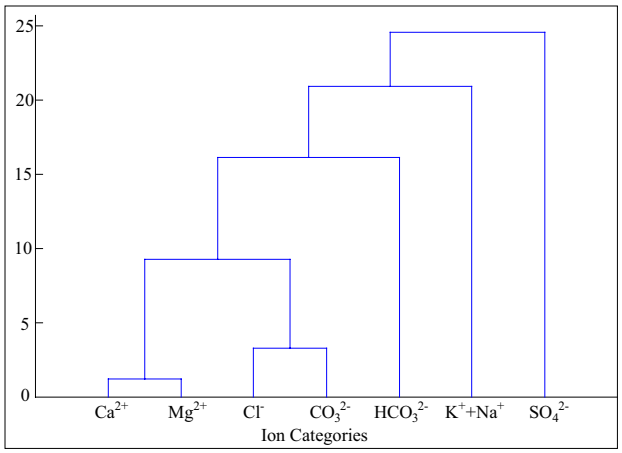
This paper used MATLAB to complete the R-type cluster analysis of the training samples. We regarded the content of $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^- and CO_3^{2-} as the clustering bases, and the R-type cluster analysis results for the three types of aquifer were obtained (Fig. 5).

It can be seen from the results of the R-type cluster analysis that the degree of similarity is high between Ca^{2+} and Mg^{2+} and between Cl^- and CO_3^{2-} in the fourth aquifer. In addition, there is a relationship between certain ions in the coal-bearing sandstone aquifer, Ca^{2+}, Mg^{2+} and SO_4^{2-} are closely related, as are $K^+ + Na^+$ and Cl^-. The cause of this phenomenon is the origin of the samples: Ca^{2+}, Mg^{2+} and SO_4^{2-} were derived from the dissolution of sulfate rocks, and $K^+ + Na^+$ and Cl^- came from soluble sodium–potassium salt rocks. In the limestone aquifer, ions such as Ca^{2+} and Mg^{2+} as well as $K^+ + Na^+$ and Cl^- are also closely related. The reason being that Ca^{2+} and Mg^{2+} came from the partial dissolution of insoluble carbonate rocks. However, the relationship between CO_3^{2-}, Ca^{2+} and Mg^{2+} was relatively small because of the reaction of $CO_3^{2-} + H_2O \rightleftharpoons HCO_3^- + OH^-$ proceeding in the positive direction when the concentration of CO_3^{2-} is increased. Thus, the concentration of CO_3^{2-} in the groundwater decreased, and the relationship between Ca^{2+} and Mg^{2+} was low. It is consistent with the alkalinity of the water samples from the limestone-karst fissure aquifer in the Taiyuan formation of the Carboniferous.

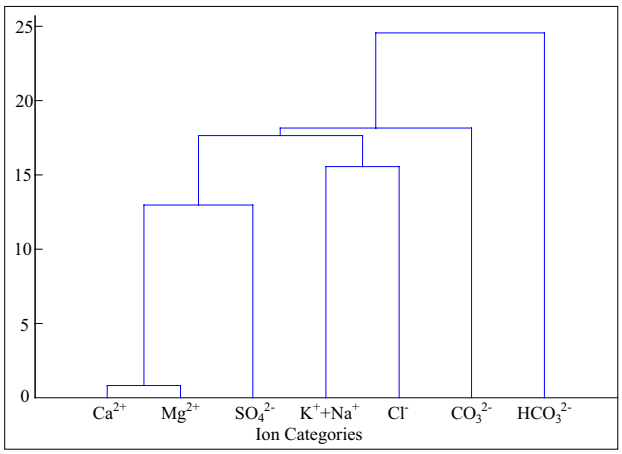
The ions in coal mine water have certain internal connections between them, and these inherent connections were often ignored in the process of establishing recognition models of mine water inrush sources, which led to excessive deviation in the practical applications of recognition models for water inrush sources. This deviation has brought a series of serious influences on the actual production of coal mines. To reduce this deviation, the training samples were preprocessed using the method of factor analysis. Finally, the recognition model was established.

Principal component analysis of the training samples

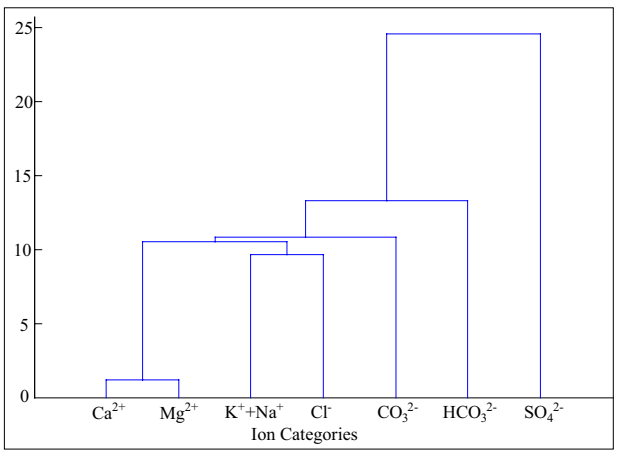
To verify the results of the R-type cluster analysis, a Pearson correlation analysis was conducted on the training samples (Chen et al. 2013; Huang and Wang 2018; Kim et al. 2005; Qian et al. 2016), and the Pearson correction coefficient



(a) The fourth aquifer water samples



(b) The coal bearing sandstone aquifer water samples



(c) The limestone aquifer water samples in the Taiyuan formation

Fig. 5 Ion content R-type cluster analysis diagram

of the three types of water samples were then obtained (Tables 2, 3, 4).

From Tables 2, 3 and 4, we could see that the correlations between some ions in each aquifer were remarkable (Qian et al. 2016). In the water sample of the fourth aquifer, the concentrations of Ca^{2+} and Mg^{2+} were positively correlated ($r=0.767$, $p<0.01$), Cl^{-} and CO_3^{2-} were significantly correlated ($r=0.971$, $p<0.01$); In the water sample of the coal-bearing sandstone aquifer, both Ca^{2+} and Mg^{2+} were moderately correlated with SO_4^{2-} (Ca^{2+} vs. SO_4^{2-}: $r=0.399$, $p<0.05$; Mg^{2+} vs. SO_4^{2-}: $r=0.359$, $p<0.05$; Table 3), and $K^{+}+Na^{+}$ showed positive correlations with SO_4^{2-} ($r=0.481$, $p<0.01$). In addition, Ca^{2+} and Mg^{2+} were also significant correlations ($r=0.877$, $p<0.01$) in the water sample of the limestone aquifer. Comparing the results of the R-type cluster analysis with the Pearson correlation coefficient, the correlation between the ions of each aquifer was basically consistent. It was, therefore, fully suggested that there is an internal connection between the ions in coal mine water.

To solve any problems with the connections among internal ions, the factor analysis of the training samples was then processed using SPSS. We used the principal component analysis to reduce the number of factors to 7, and the 7 original factors were then combined into 3 independent indicators to reflect the hydrochemical information.

Using the principal component analysis in factor analysis, the initial factors were extracted from the ion's correction coefficient matrix, and the initial eigenvalue and the variances explained by the principal component analysis were obtained (Table 5).

The number of principle components could be determined by the cumulative variance of the principle components. It is generally thought that the cumulative variance of extracting principal components is more than 80%, which means that the selected number of principal components can fully reflect the hydrochemical information of the training samples (Chen et al. 2013; Wang et al. 2017a, b; Yin et al. 2006; Zhang et al. 2017). Therefore, we extracted three principal components, which were consistent with the results of selecting the number of principal components according to the eigenvalues. To some extent, the number of principal components could be determined using eigenvalues greater than 1 as criteria. The eigenvalues of the principal components are shown in Fig. 6.

The maximum variance algorithm was used for the orthogonal rotation of the initial load matrix of factors so that loads of each ion on the same factor were distinctly different. The orthogonal rotation converges after 6 iterations, and the orthogonal rotation factor loading matrix (Table 6) and the orthogonal rotation factor loading diagram (Fig. 7) could then be obtained. After the orthogonal rotation of three types of water samples, each principal component

Table 2 Pearson correlation coefficients of the fourth aquifer water samples

	$K^+ + Na^+$	Ca^{2+}	Mg^{2+}	Cl^-	SO_4^{2-}	HCO_3^-	CO_3^{2-}
$K^+ + Na^+$	1						
Ca^{2+}	−0.048	1					
Mg^{2+}	−0.273	0.767*	1				
Cl^-	0.066	0.372	0.570	1			
SO_4^{2-}	0.959**	0.071	−0.214	−0.120	1		
HCO_3^-	−0.615	0.733	0.800*	0.390	−0.551	1	
CO_3^{2-}	−0.137	0.305	0.601	0.971**	−0.322	0.478	1

*Represents 0.05 levels of bilateral significant correlation; **represents 0.01 levels of bilateral significant correlation

Table 3 Pearson correlation coefficients of the coal bearing sandstone aquifer water samples

	$K^+ + Na^+$	Ca^{2+}	Mg^{2+}	Cl^-	SO_4^{2-}	HCO_3^-	CO_3^{2-}
$K^+ + Na^+$	1						
Ca^{2+}	0.068	1					
Mg^{2+}	−0.223	0.812**	1				
Cl^-	0.481**	0.119	−0.230	1			
SO_4^{2-}	0.558**	0.399*	0.359*	0.705**	1		
HCO_3^-	0.198	0.334	0.399*	−0.583**	−0.155	1	
CO_3^{2-}	0.441*	−0.174	−0.401*	0.792**	0.312	−0.704**	1

*Represents 0.05 levels of bilateral significant correlation; **represents 0.01 levels of bilateral significant correlation

Table 4 Pearson correlation coefficients of the limestone aquifer water samples

	$K^+ + Na^+$	Ca^{2+}	Mg^{2+}	Cl^-	SO_4^{2-}	HCO_3^-	CO_3^{2-}
$K^+ + Na^+$	1						
Ca^{2+}	−0.418**	1					
Mg^{2+}	−0.492**	0.877**	1				
Cl^-	0.041	0.355*	0.259	1			
SO_4^{2-}	−0.228	0.703**	0.711**	−0.318*	1		
HCO_3^-	0.280	0.069	−0.062	0.869**	−0.540**	1	
CO_3^{2-}	0.056	0.226	0.157	0.954**	−0.435**	0.889**	1

*Represents 0.05 levels of bilateral significant correlation; **represents 0.01 levels of bilateral significant correlation

Table 5 Interpreting total variance

Principal component	Initial eigenvalue			Extracted eigenvalue		
	Summation	Variance (%)	Cumulative variance (%)	Summation	Variance (%)	Cumulative variance (%)
1	2.985	42.637	42.637	2.985	42.637	42.637
2	1.774	25.347	67.984	1.774	25.347	67.984
3	1.026	14.661	82.644	1.026	14.661	82.644
4	0.989	14.129	96.773			
5	0.138	1.975	98.747			
6	0.058	0.834	99.581			
7	0.029	0.419	100.000			

represented the hydrochemical information of different ions. Among them, principal component 1 represented Ca^{2+}, Mg^{2+} and SO_4^{2-} and reflected the information of 42.637% of the training samples. Principal component 2 represented Cl^- and CO_3^{2-} and reflected the information of 25.347% of the training samples. Principal component 3 represented $K^+ + Na^+$ and reflected the information of 14.661% of the training samples.

Principal component analysis of the 82 training water samples was carried out using SPSS, and we obtained scores

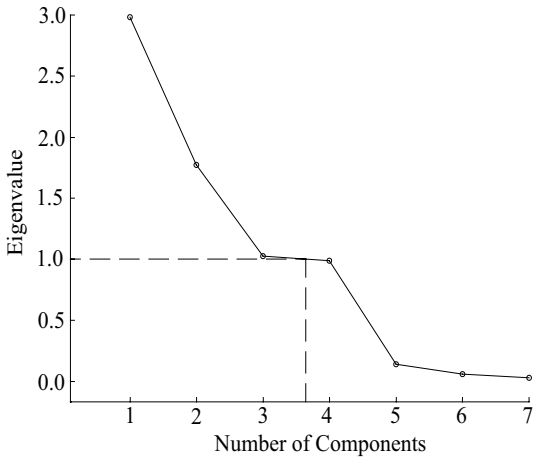


Fig. 6 Scree plot of the principal components

Table 6 Orthogonal rotation factor loading matrix

	Principal component			Communality	
	1	2	3	Initial	Extract
$K^+ + Na^+$	−0.099	0.169	0.898	1.000	0.845
Ca^{2+}	0.885	0.048	−0.420	1.000	0.962
Mg^{2+}	0.857	−0.013	−0.472	1.000	0.958
Cl^-	0.190	0.928	0.008	1.000	0.897
SO_4^{2-}	0.913	−0.191	0.298	1.000	0.958
HCO_3^-	0.131	−0.340	−0.312	1.000	0.230
CO_3^{2-}	−0.269	−0.905	0.211	1.000	0.936

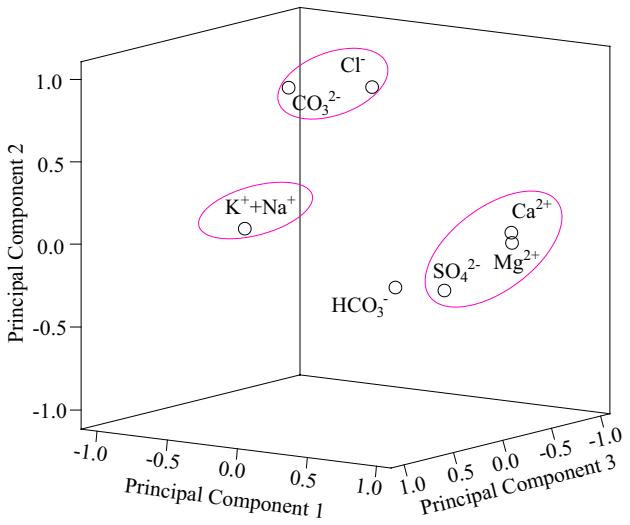


Fig. 7 Orthogonal rotation factor loading diagram

for three principal components from the 82 training water samples (Table 7). The scores of the principal components were expressed as Y_1, Y_2 and Y_3, respectively. The correlation coefficients between the three types of principal components and the original variables are shown in Table 8.

According to the principal component score coefficients, we could get the expression of principal component scores, relational expressions of the three extracted principal components Y_1, Y_2 and Y_3 with the original variables X_1, X_2, X_3, X_4, X_5, X_6 and X_7 were obtained as follows:

$$\left.\begin{aligned} Y_1 &= 0.153X_1 + 0.321X_2 + 0.296X_3 + 0.111X_4 + 0.485X_5 \\ &\quad - 0.015X_6 - 0.058X_7 \\ Y_2 &= -0.033X_1 + 0.107X_2 + 0.079X_3 + 0.531X_4 - 0.130X_5 \\ &\quad - 0.146X_6 + 0.479X_7 \\ Y_3 &= 0.707X_1 - 0.167X_2 - 0.208X_3 - 0.083X_4 + 0.475X_5 \\ &\quad - 0.185X_6 - 0.010X_7 \end{aligned}\right\}, \tag{5}$$

where Y_1, Y_2 and Y_3 represent the scores of principal component 1, principal component 2 and principal component 3 of the training samples; X_1, X_2, X_3, X_4, X_5, X_6 and X_7 represent the contents of $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^- and CO_3^{2-}, respectively.

Construction of the recognition model

We regarded the principal component scores Y_1, Y_2 and Y_3 of the 82 training water samples as independent variables for implementing the multiple logistic regression recognition analysis. The parameters of the multiple logistic regression recognition model are shown in Table 9.

The recognition function of the solution is as follows:

$$\left.\begin{aligned} G_1 &= -0.940Y_1 - 0.312Y_2 + 0.675Y_3 - 1.561 \\ G_2 &= -6.630Y_1 + 1.416Y_2 + 3.166Y_3 - 3.479 \end{aligned}\right\}. \tag{6}$$

Formula (6) could be simplified using Formula (5). We were then able to derive the following Formula (7).

$$\left.\begin{aligned} G_1 &= 0.343X_1 - 0.448X_2 - 0.443X_3 - 0.326X_4 \\ &\quad - 0.095X_5 - 0.065X_6 - 0.102X_7 - 1.561 \\ G_2 &= 0.027X_1 - 3.395X_2 - 3.254X_3 - 1.016X_4 \\ &\quad - 3.699X_5 - 0.314X_6 + 0.741X_7 - 3.479 \end{aligned}\right\}. \tag{7}$$

Finally, the expressions for the probability functions of the three types of water inrush sources are as follows:

Table 7 The scores of the principal components

Number	Y_1	Y_2	Y_3	Actual category	Predicted category	Number	Y_1	Y_2	Y_3	Actual category	Predicted category
1	−1.25	−1.12	−1.43	1	1	42	0.65	1.44	−1.06	3	3
2	0.12	0.90	−1.12	1	3	43	0.72	1.51	−1.24	3	3
3	−0.41	−0.95	−1.20	1	3	44	0.79	1.59	−1.20	3	3
4	−0.51	0.44	−0.72	1	1	45	0.64	1.44	−1.36	3	3
5	−1.22	−1.00	−1.24	1	3	46	0.69	1.41	−1.10	3	3
6	2.20	−1.49	3.35	1	3	47	0.70	1.43	−1.11	3	3
7	1.49	−1.26	4.45	1	1	48	0.51	1.33	−0.77	3	3
8	−0.58	2.60	1.48	2	2	49	−0.06	−0.59	0.13	3	3
9	−0.82	2.76	1.12	2	2	50	0.45	−0.48	0.13	3	3
10	−0.93	1.68	0.73	2	2	51	0.53	−0.44	−0.13	3	3
11	−0.97	1.20	0.54	2	2	52	0.53	−0.44	−0.13	3	3
12	−0.95	1.28	0.53	2	2	53	0.53	−0.51	0.44	3	3
13	−0.44	1.00	1.08	2	2	54	0.02	−0.39	−0.69	3	3
14	−0.44	1.00	1.08	2	2	55	0.82	−0.57	0.61	3	3
15	−0.43	0.99	1.08	2	2	56	0.20	−0.57	0.25	3	3
16	−0.38	1.08	1.16	2	2	57	−0.62	−0.52	0.53	3	2
17	−0.49	0.96	1.01	2	2	58	1.32	−0.34	0.04	3	3
18	−1.05	1.08	0.35	2	2	59	1.22	−0.33	−0.10	3	3
19	−0.88	0.92	0.15	2	2	60	1.14	−0.40	0.23	3	3
20	−0.73	0.99	0.05	2	2	61	−0.78	−0.64	0.22	3	2
21	−0.56	0.82	0.92	2	2	62	1.18	−0.40	0.18	3	3
22	−0.45	0.85	1.06	2	2	63	1.17	−0.46	0.20	3	3
23	−0.04	1.30	0.57	2	2	64	1.10	−0.33	−0.19	3	3
24	−0.25	0.85	0.56	2	2	65	1.55	−0.24	−0.66	3	3
25	−0.85	0.59	−0.05	2	2	66	1.54	−0.26	−0.63	3	3
26	−0.91	0.71	0.31	2	2	67	1.51	−0.24	−0.70	3	3
27	−0.84	−0.44	1.90	2	2	68	1.55	−0.26	−0.56	3	3
28	−1.22	−1.10	−0.47	2	2	69	1.54	−0.26	−0.64	3	3
29	−1.26	−1.10	−0.25	2	2	70	1.54	−0.26	−0.59	3	3
30	−1.25	−1.10	−0.10	2	2	71	1.53	−0.24	−0.61	3	3
31	−1.24	−1.14	−0.33	2	2	72	1.54	−0.27	−0.51	3	3
32	−0.13	−1.43	1.12	2	3	73	−1.20	−0.44	−1.52	3	3
33	−0.66	−1.06	0.77	2	2	74	−1.20	−0.44	−1.52	3	3
34	−1.24	−1.01	−0.34	2	2	75	−0.67	−0.64	0.51	3	2
35	−1.20	−0.79	−0.41	2	2	76	−0.87	−0.93	0.43	3	2
36	−1.01	−1.17	−0.45	2	2	77	0.51	−0.95	−0.13	3	3
37	−1.16	−1.15	−0.40	2	2	78	0.40	−0.70	0.80	3	3
38	−1.11	−1.13	−0.36	2	2	79	−0.83	−0.67	0.06	3	2
39	−1.14	−1.00	−0.60	2	2	80	1.19	−0.36	−0.49	3	3
40	−1.23	−0.96	−0.42	2	2	81	1.66	−0.14	−1.53	3	3
41	0.42	1.28	−0.85	3	3	82	1.24	−0.28	−0.21	3	3

Table 8 Principal component score coefficients

	Principal component		
	1	2	3
$K^+ + Na^+$	0.153	−0.033	0.707
Ca^{2+}	0.321	0.107	−0.167
Mg^{2+}	0.296	0.079	−0.208
Cl^-	0.111	0.531	−0.083
SO_4^{2-}	0.485	−0.130	0.475
HCO_3^-	−0.015	−0.146	−0.185
CO_3^{2-}	−0.058	0.479	−0.010

$$\left.\begin{aligned} P_1 &= \frac{e^{G_1}}{1+e^{G_1}+e^{G_2}} \\ P_2 &= \frac{e^{G_2}}{1+e^{G_1}+e^{G_2}} \\ P_3 &= \frac{1}{1+e^{G_1}+e^{G_2}} \end{aligned}\right\}, \tag{8}$$

where P_1, P_2 and P_3 are the probability functions of the respective recognition models of types 1, 2, and 3; X_1, X_2, X_3, X_4, X_5, X_6 and X_7 represent the contents of $K^+ + Na^+$, Ca^{2+}, Mg^{2+}, Cl^-, SO_4^{2-}, HCO_3^- and CO_3^{2-}, respectively; and the final item of the discriminant function is a constant.

Verification of water inrush source recognition model

The 82 groups of training samples in Table 7 were integrated into the established multiple logistic regression recognition model based on cluster analysis one by one for cross-validation (Table 10). The results showed that all water samples were discriminated with a discrimination rate of 87.8%. Among them, the recognition accuracy of water samples from the fourth aquifer is 42.8%, the recognition accuracy of water samples from the coal-bearing sandstone aquifer is 96.9% and the recognition accuracy of water samples from the limestone aquifer is 88.1%. The reason for the difference in the recognition accuracy of various aquifers lies in the difference in the number of training water samples. Because coal mining is less threatened by water inrush from the fourth aquifer, the limited number of water samples were collected from the fourth aquifer. However, the recognition model is established based on a certain amount of water samples. Therefore, the recognition accuracy of water samples from the fourth aquifer significantly different from the coal-bearing sandstone aquifer and the limestone aquifer. Meanwhile, this result can be compared to the traditional multiple logistic regression recognition model, which incurred multiple errors in its rediscrimination steps and had a correct discrimination rate of less than 78.5%. Therefore, the multiple logistic regression recognition model based on cluster analysis was more accurate, had a higher degree of stability, and could meet the actual requirements of water inrush source recognition.

In addition, to further verify the accuracy of the established multiple logistic regression recognition model based on cluster analysis, 16 water samples to be discriminated from the Qinan mining area were substituted

Table 10 Classification results of cross-validation

	Type	Water inrush aquifer types			Total
		FA	CBSA	LA	
Recognition model					
Count	FA	3	0	4	7
	CBSA	0	32	1	33
	LA	0	5	37	42
Correct rate %	FA	42.9	0	57.1	100
	CBSA	0	96.9	3.1	100
	LA	0	11.9	88.1	100

FA fourth aquifer, *CBSA* coal-bearing sandstone aquifer, *LA* limestone aquifer

Table 9 Multiple logistic regression recognition analysis model parameters

Type	Variable	*B*	Standard error	Wald value	Degree of freedom	Significance level	Exp(*B*)
1	Intercept	−1.561	0.601	6.759	1	0.009	
	Y_1	−0.940	0.574	2.678	1	0.102	0.391
	Y_2	−0.312	0.735	0.180	1	0.671	0.732
	Y_3	0.675	0.462	2.133	1	0.144	1.965
2	Intercept	−3.479	1.548	5.049	1	0.025	
	Y_1	−6.630	2.076	10.196	1	0.001	0.001
	Y_2	1.416	0.728	3.784	1	0.052	4.121
	Y_3	3.166	1.027	9.498	1	0.002	23.713

The reference type is the limestone aquifer water samples, as denoted by 3

Table 11 Classification results of the water inrush source discriminant model

Number	Constant ion content (mg/l)							Actual result	Predicted result
	$Na^+ + K^+$	Ca^{2+}	Mg^{2+}	Cl^-	SO_4^{2-}	HCO_3^-	CO_3^{2-}		
A1	86.7	95	56.5	63	218	389.32	0	1	1
A2	197.9	29	61.3	319	113	357.74	292.92	1	2
A3	264.0	29	26.4	317	135	66.34	520.83	2	2
A4	379.1	9	3.1	412	233	197.62	378.34	2	2
A5	459.4	7	4.9	442	255	233.46	360.22	2	2
A6	427.3	66	28.9	537	246	494.7	468.68	2	2
A7	276.9	214	81.5	278	667	428.17	5.17	3	3
A8	254.5	198	78.1	243	614	430.56	5.32	3	3
A9	270.7	215	82.9	265	606	428.17	4.19	3	3
A10	258.4	201	79.3	255	619	394.68	4.53	3	3
A11	269.9	209	80.5	262	608	425.77	5.95	3	3
A12	256.4	201	76.1	248	596	428.17	5.36	3	3
A13	386.5	109	46.4	252	349	576.47	5.25	3	3
A14	371.4	98	45.4	231	376	552.55	5.38	3	3
A15	327.4	134	58.3	241	404	516.67	4.96	3	3
A16	287.2	167	79.2	244	600	428.17	4.89	3	3

into the multiple logistic regression recognition model for discrimination (Table 11). Table 11 shows that 16 water samples are classified accurately by the established multiple logistic regression recognition model based on cluster analysis and only one sample is wrongly discriminated, showing an accuracy of 93.8%. Water sample A2 is actually the fourth aquifer water sample, but it is discriminated as the coal-bearing sandstone aquifer water sample in the model. Through comprehensive comparison, the multiple logistic regression recognition model based on cluster analysis was seen to be more accurate and to have greater extensive applicability than those of the traditional multiple logistic regression recognition model. Therefore, the multiple logistic regression recognition model based on cluster analysis has significant engineering relevance.

Results and discussion

Based on the hydrogeological conditions of the mining area, cluster analysis of water quality samples was carried out in this paper. The analysis results were then utilized to analyze and extract typical water samples. At last, the multiple logistic regression recognition model based on cluster analysis was established. According to the results of the model recognition and the engineering application, the conclusions were drawn as follows:

1. Through the cluster analysis of the original water samples, the nonconforming water samples were eliminated. The 82 water samples that accurately reflect the hydrochemical characteristics of the water inrush aquifer were screened from 93 original water samples, and they were used as training samples to establish the recognition model, which reduced the influence of the errors caused by the water quality analysis on the accuracy of the mode.
2. In the process of establishing recognition model, to eliminate the internal connections between the ions, this paper adopted the principal component analysis method to cut down the dimension of the initial seven types of variables and combine the original seven factors into a few independent indexes to comprehensively reflect the hydrochemical information.
3. The overall recognition accuracy of the multiple logistic regression recognition model based on cluster analysis reaches 87.8% and has high accuracy. It is easy to operate in the actual water source discrimination process, with straightforward discrimination results. This recognition model provides a new way to discriminate mine water inrush sources and has important guiding significance for mine water prevention and control work.
4. Because the recognition model is based on hydrological data from a certain amount. And the quantity of water sample has certain influence on the accuracy of the recognition model. Therefore, we should collect more water sample data to improve accuracy. In addition, given the complexity of hydrogeological conditions, temperature, and human activities on aquifers in the study area, future studies should fully consider the impact of these factories to promote the applications of the model.

Acknowledgements The project was supported by the National Natural Science Foundation of China (Grant nos. 41672273, 51474008), the Fundamental Research Funds for the Central Universities (22120180313) and the Anhui Natural Science Foundation of China (1508085QE89). The research was also substantially supported by the Key Laboratory of Geotechnical and Underground Engineering of Ministry of Education (Tongji University).

References

Biswas A, Sharma SP (2017) Geophysical surveys for identifying source and pathways of subsurface water inflow at the Bangur chromite mine, Odisha, India. Nat Hazards 88(2):947–964

Bu HM, Tan X, Li SY, Zhang QF (2010) Water quality assessment of the Jinshui River (China) using multivariate statistical techniques. Environ Earth Sci 60(8):1631–1639

Chen HJ, Li XB, Liu AH, Peng SQ (2009) Identifying of mine water inrush sources by Fisher discriminant analysis method. J Cent South Univ 40:1114–1120

Chen LW, Yin XX, Liu X, Gui HR (2013) Multivariate statistical analysis on hydrochemical evolution of groundwater in the concealed coal mines in North China. Coal Geol Explor 41(6):43–51

Faghih Nasiri E, Yousefi Kebria D, Qaderi F (2018) An experimental study on the simultaneous phenol and chromium removal from water using titanium dioxide photocatalyst. Civ Eng J 4(3):585

Farnham IM, Stetzenbach KJ, Singh AK, Johannesson KH (2000) Deciphering groundwater flow systems in Oasis Valley, Nevada, using trace element chemistry, multivariate statistics, and geographical information system. Math Geosci 32(8):943–968

Ganyaglo SY, Banoeng-Yakubo B, Osae S, Dampare SB (2011) Water quality assessment of groundwater in some rock types in parts of the eastern region of Ghana. Environ Earth Sci 62:1055–1069

Gui HR, Lin ML (2016) Types of water hazards in China coalmines and regional characteristics. Nat Hazards 84(2):1501–1512

Güler C, Thyne GD, McCray JE, Turner KA (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol J 10(4):455–474

Hu W, Dong S, Yan L (2011) Water hazard control technology for safe extraction of coal resources influenced by faulted zone. Procedia Earth Planet Sci 3:1–10

Huang PH, Chen JS (2011) Fisher identify and mixing model based on multivariate statistical analysis of mine water inrush sources. J China Coal Soc 36(S1):131–136

Huang PH, Wang XY (2018) Piper-PCA-Fisher recognition model of water inrush source: a case study of the Jiaozuo mining area. Geofluids 2018:1–10

Huang PH, Yang ZY, Wang XY, Ding FF (2019) Research on Piper-PCA-Bayes-LOOCV discrimination model of water inrush source in mines. Arab J Geosci 12:334

Jolliffe IT (2002) Principal component analysis. Wiley, Hoboken

Keskin TE, Düğenci M, Kaçaroğlu F (2015) Prediction of water pollution sources using artificial neural networks in the study areas of Sivas, Karabük and Bartın (Turkey). Environ Earth Sci 73(9):5333–5347

Kim JH, Kim RH, Lee J, Cheong TJ, Yum BW, Chang HW (2005) Multivariate statistical analysis to identify the major factors governing groundwater quality in the coastal area of Kimje. South Korea. Hydrol Process 19(6):1261–1276

Li GQ, Meng ZP, Wang XQ, Yang J (2017) Hydrochemical prediction of mine water inrush at the Xinli Mine, China. Mine Water Environ 36(1):78–86

Liu X, Chen LW, Lin ML, Li SD (2013) Fisher recognition analysis for coal mining inrush water source under mining-induced disturbance and inversion of groundwater recharge relation. Hydrol Eng Geol 40(4):36–43

Liu Q, Sun YJ, Xu ZM, Xu G (2018) Application of the comprehensive identification model in analyzing the source of water inrush. Arab J Geosci 11(9):189

Lu JT, Li XB, Gong FQ (2012) Recognizing of mine water inrush sources based on principal components analysis and fisher discrimination analysis method. China Saf Sci J 22(7):109–115

Meglen RR (1992) Examining large databases: a chemometric approach using principal components analysis. Mar Chem 39(1):217–237

Panagopoulos GP, Angelopoulou D, Tzirtzilakis EE, Giannoulopoulos P (2016) The contribution of cluster and discriminant analysis to the classification of complex aquifer systems. Environ Monit Assess 188:591

Qian J, Wang L, Ma L, Lu YH, Zhao WD, Zhang Y (2016) Multivariate statistical analysis of water chemistry in evaluating groundwater geochemical evolution and aquifer connectivity near a large coal mine, Anhui, China. Environ Earth Sci 75(9):747

Reghunath R, Murthy TRS, Raghavan BR (2002) The utility of multivariate statistical techniques in hydrogeochemical studies: an example from Karnataka, India. Water Res 36(10):2437–2442

Wang LB (2010) Multivariate statistical analysis: models, case study and application of SPSS. Economic Science Press, Beijing

Wang JC, Guo ZG (2001) Logistic regression model-methods and applications. Higher Education Press, Beijing

Wang XY, Zhao W, Liu XM, Wang TT, Zhang JG, Guo JW, Chen GS, Zhang B (2017a) Identification of water inrush source from coalfield based on entropy weight-fuzzy variable set theory. J China Coal Soc 42(9):2433–2439

Wang Y, Zhou MR, Yan PC, He CY, Liu D (2017b) Identification of coalmine water inrush source with PCA-BP model based on laser-induced fluorescence technology. Spectrosc Spectr Anal 37(3):978–983

Wei WX, Han J, Shi LQ, Lu XM, Zhang XJ (2015) Application of modern data analysis in mine water gushing prediction. Coal Industry Press, Beijing

Wu Q, Guo XM, Shen JJ, Xu S, Liu SQ, Zeng YF (2016) Risk assessment of water inrush from aquifers underlying the Gushuyuan coal mine, China. Mine Water Environ 36(1):1–8

Xu B, Zhang Y, Jiang L (2012) Coupled model based on grey relational analysis and stepwise discriminant analysis for water source. Rock Soil Mech 33(10):3122–3228

Yin XX, Xu GQ, Gui HR, Chen LW (2006) Analyzing for sources of inrush-water in Wanbei Mining Area by systemic clustering and stepwise distinguishing. Coal Geol Explor 34(2):61–64

Zhang WT (2002) SPSS 11.0 statistical analysis tutorial (advanced). Beijing Hope Electronic Press, Beijing

Zhang H, Yao DX, Lu HF, Zhu NN, Xue L (2017) Application of principal component analysis and bayes discrimination approach in water source identification. Coal Geol Explor 45:87–93