Hindawi
Geofluids
Volume 2021, Article ID 6663827, 17 pages
https://doi.org/10.1155/2021/6663827

Research Article

Application of Partial Least Squares-Discriminate Analysis Model Based on Water Chemical Compositions in Identifying Water Inrush Sources from Multiple Aquifers in Mines

Yaoshan Bi,[1] Jiwen Wu,[1] Xiaorong Zhai,[1] Shuhao Shen,[1] Libin Tang,[1] Kai Huang,[1] and Dawei Zhang[2]

[1]School of Earth and Environment, Anhui University of Science and Technology, Huainan 232001, China
[2]Renlou Coal Mine, Wanbei Coal and Electricity Group Co., Ltd., Huaibei 235123, China

Correspondence should be addressed to Jiwen Wu; jwwuaust@163.com

Received 7 November 2020; Revised 21 January 2021; Accepted 1 February 2021; Published 17 February 2021

Academic Editor: Qingquan Liu

Mine water inrush seriously threatens the safety of coal mine production. Quick and accurate identification of mine water inrush sources is of great significance to preventing mine water hazards. This paper combined partial least squares-discriminate analysis (PLS-DA) with inrush water chemical composition to identify the source of water inrush from multiple aquifers in mines. The Renlou Coal Mine in the Linhuan mining area was selected for this study, and seven conventional water chemical compositions from 54 water samples in three aquifers were collected and tested, of which 45 water samples were used to establish the PLS-DA discriminant model, and nine were used to test the prediction effect. To improve model accuracy and predictive ability, hierarchical clustering analysis method was used to eliminate seven unqualified water samples to reduce the errors caused by improper data. PCA and PLS-DA methods were used to analyze and process the remaining water sample data, and on the basis of PCA analysis, the remaining 38 water samples were used to establish the PLS-DA discriminant model. The model was validated using permutation and external prediction tests. The research shows the following results: (1) Both PCA and PLS-DA methods can distinguish water samples from three different water sources, but the classification effect of PLS-DA was better than PCA because it can strengthen the difference of water chemical composition between different water sources. (2) The correct discrimination rate of the PLS-DA discriminant model was as high as 100%, and permutation tests showed that the model was not overfit. External validation found that the model had good stability and discrimination. (3) HCO_3^- and total dissolved solids (TDS) were the most important differential marker compositions that affected the discrimination results based on Variable Importance for the Projection (VIP) scores. The discriminant model established in this study combined the advantages of principal component analysis and multiple regression analysis, providing a new method for accurately identifying the sources of water inrush in mines.

1. Introduction

Coal resource is an important basic resource for the long-term rapid and stable development of the national economy in China [1]. As mining depth has increased, water inrush disaster occurrence has also increased, which poses a serious threat to the safety of coal mine production [2, 3]. Quickly and accurately identifying the sources of mine water inrush is important for the prevention and control of coal mine water inrush disasters, and it is also a top concern in mine water disaster management research [4]. Many scholars have proposed various methods to identify the source of water inrush in mines, such as groundwater chemistry [5, 6], trace elements and isotopes [7, 8], water temperature [9, 10], and groundwater level dynamic observations [11]. After comparing groundwater chemistry with other methods, Wu et al. [12] concluded that the water chemistry discrimination method had more advantages in practical applications. Because groundwater chemistry can reflect the essential characteristics of groundwater, and can accurately, quickly, and economically identify water sources, it has been more commonly used to identify water inrush water sources in mines [13].

At present, mathematical statistical methods and machine learning methods are typically applied when using the chemical composition of groundwater to identify water inrush sources, such as fuzzy mathematical theory [14, 15], grey relational analysis [16, 17], back propagation neural network (BP neural network) [18], Fisher's discriminant method [19], Bayes' discriminant method [20], distance discriminant method [21], extension identification method [22], support vector machine (SVM) [23], and extreme learning machine (ELM) [24]. The application of these mathematical and machine learning methods enriches the content of mine water source discrimination theory, improves identification accuracy, and demonstrates good practicability and effectiveness. However, most of these present discriminant methods have not considered the complicated information superposition problem between water chemistry indicators, a problem that results in misdiscrimination of the established model in the practical application process, and their recognition accuracy still needs to be further improved [25]. Therefore, some scholars have adopted the principal component analysis (PCA) method in the water source discriminant analysis and obtained better analysis results [26]. PCA can extract and compresses the information of hydrochemical data of different water sources, transform original data into mutually independent new data without information superposition, and eliminate the effects caused by information superposition between indicators so that characteristics of different water sources can be described more effectively [27].

In this paper, a new promising method (partial least squares-discriminate analysis (PLS-DA)) is presented, and this method can effectively solve the problem of multicollinearity between multiple variables and is reliable especially when there is a high degree of correlation between them. PLS-DA is a supervised multivariate statistical method that integrates the basic functions of PCA, canonical correlation analysis and multiple regression analysis [28, 29]. Similar to PCA, PLS-DA is also a multidimensional vector analysis method based on dimensionality reduction. However, different from PCA, the PLS-DA method performs orthogonal decomposition of the measurement matrix while also performing orthogonal decomposition of the response matrix. In other words, PLS-DA can preset classifications and add grouping variables for supervised analysis to further strengthen the differences between groups [29]. Its advantage is that it can remove the influence of uncontrolled variables on data analysis as much as possible, further mine the information in the data, and quantify the degree of component difference caused by characteristic ions [30]. Barker and Matthew [30] used statistical theory to show that PLS-DA performed good classification. In recent years, PLS-DA has been widely used for screening pharmaceutical ingredients; tracing the origins of wine, meat, etc.; and identifying and classifying tea and navel oranges [31]. However, few studies have used it to identify water inrush in mines. Yan et al. [32] used laser-induced fluorescent (LIF) technology to obtain the fluorescence spectrum of inrush water sources and used it as an indicator for PLS-DA discrimination with good effect. However, it is difficult to obtain the fluorescence spectrum of the inrush water sources using this technology for all mines, and the test cost is relatively high. This paper used the conventional ion compositions of the inrush water as indicators to establish a PLS-DA discriminant model and further broaden the application range of PLS-DA in identifying mine water inrush sources. Seven conventional water chemical compositions from water samples of three aquifers were used as indicators in this study and the hierarchical clustering analysis method was used to eliminate the unqualified water samples. The PCA method was used to analyze the remaining water sample data, and then the PLS-DA discrimination model based on chemical compositions of inrush water was established. Permutation and external verification tests demonstrated model stability and discriminative ability.

2. Description of the Study Area

The Renlou Coal Mine is located in the Linhuan mining area of northern Anhui Province, China. Its geographic location is shown in Figure 1. The mine field is located in the middle of the Huaibei Plain, and the terrain is flat. There is only a small, artificially dredged seasonal river in the mine field and its flow is controlled by rainfall. The average annual rainfall in the study area is 820 mm, mostly concentrated from June to September, and the maximum rainfall in July is 268.5 mm. The annual average temperature is 14.3°C, with the lowest temperature in January of -23.2°C and the highest in July of 41°C. The maximum evaporation occurs from June to August, with a multiyear average evaporation of 1774 mm.

The Renlou Coal Mine is located in the southeast wing of the Tongting Anticline, and the stratigraphic occurrence in the area is relatively gentle, generally 13°~20°. At present, the primary mines are No. 7_2 coal, No. 7_3 coal, and No. 8_2 coal. Water inrush is an important threat to the safe production of the Renlou Coal Mine, where 21 inrushes occurred from January 1989 to February 2013. The water inrush duration at some points was long with a large amount of water. For example, during the excavation process of working face $7_2$22, the maximum instantaneous water influx reached 34570 m^3/h due to the karst collapse column connected to other aquifers, causing the entire well to be flooded. Therefore, accurate identification of water inrush sources is very important for the prevention and control of water disasters in the Renlou Coal Mine.

There are multiple groundwater aquifer layers in the minefield. From top to bottom, there are loose pore aquifers, coal-measure formation sandstone fractured aquifers, Taiyuan formation limestone karst fractured aquifers, and Ordovician karst fractured aquifers. Of these, the fourth aquifer in the loose layer (referred to as the "fourth aquifer") may enter the mine through a crack or a vertical guide channel and affect production, and it is also the main hidden water hazard in shallow coal mining. The sandstone fractured aquifer (referred to as the "coal-bearing sandstone aquifer") is mainly stored in the structural fissures of sandstone layers as static reserves. Due to the influence of geological structure, the fissures are unevenly developed. When the

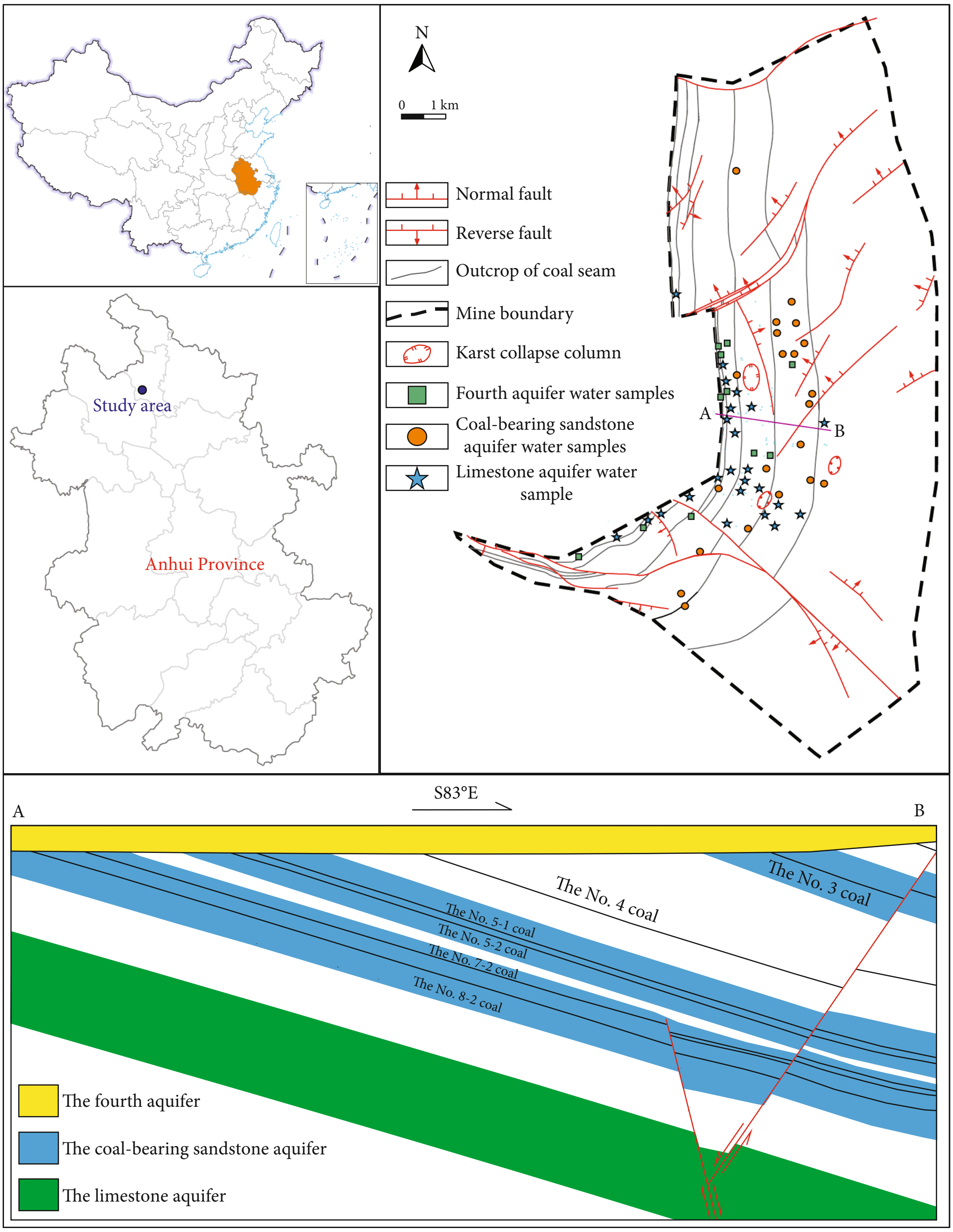


Figure 1: Map showing the location of the study area, with the distribution of faults, folds, sampling sites, and hydrogeologic section within the Renlou Coal Mine.

fissures are developed or connect with other aquifers, the water output will increase, causing other aquifers to become indirect water sources for the main coal seams. The limestone karst fissure aquifer of the Taiyuan Formation (referred to as the “limestone aquifer”) is the main water source for the mine. The average distance between the aquifer and the No. 8_2 coal seam floor is about 140 m. However, due to the development of hidden karst water-conducting subsidence columns and water-conducting faults in the minefield, these passages may cause limestone aquifer water to enter the mine. The water content of the Ordovician karst fissure aquifer is very rich, but the aquifer is nearly 290 m away from the No. 8_2 coal floor. Therefore, it does not have hydraulic connections with the mine under normal conditions and does not directly threaten the safety of the mine.

3. Materials and Methods

3.1. Sampling and Testing. In this study, 54 original water samples were collected from three main water inrush aquifers

Table 1: Test data for water samples.

Num	Ca^{2+} (mg/L)	Mg^{2+} (mg/L)	$Na^{+}+K^{+}$ (mg/L)	Cl^{-} (mg/L)	SO_4^{2-} (mg/L)	HCO_3^{-} (mg/L)	TDS (mg/L)	Water source type
Training set samples								
1	17.88	20.77	195.36	118.9	112.8	302.66	630	FA
2	349.2	109.6	488.24	1072	585.6	319.25	2765	FA
3	268.2	105	396.27	823.2	515.7	323.14	2270	FA
4	272.5	111.8	335.66	709.1	544.4	370.1	2159	FA
5	9.4	0.74	290.18	432	8.26	30.63	765	FA
6	143.2	3.99	449.98	912.2	6.05	72.31	1552	FA
7	55.73	92.33	381.66	763.3	230.5	28.74	1543	FA
8	45.69	43.67	118.29	181.9	90.55	225.29	602	FA
9	4.35	18.11	228.89	104.7	121.7	243.71	667	FA
10	8.24	3.83	479.27	250.3	20.17	785.92	1193	CBSA
11	7.01	3.89	867.41	685.4	16.81	1028.2	2661.6	CBSA
12	10.62	6.8	788.32	739.7	24.47	762.13	2377.7	CBSA
13	31.45	3.62	910.1	967.5	131.7	508.68	2298.7	CBSA
14	5.97	2.52	967.95	580.9	11.38	1384.4	2363	CBSA
15	13.35	2.52	710.35	559.2	48.03	854.46	2219	CBSA
16	4.05	0.98	755.27	395.5	3.65	1213.4	1829	CBSA
17	141.2	253.1	563.09	999.3	561.8	539.16	2788	CBSA
18	1.29	4.33	526.03	267.7	5.76	918.03	1278	CBSA
19	1.48	3.07	608.81	235.8	9.06	1146.8	1465	CBSA
20	11.36	24.56	812.22	572.5	135	1129.8	2121	CBSA
21	2.86	4.33	655.2	514.9	4.12	835.46	1618	CBSA
22	1.91	4.33	635.42	442.1	6.59	897.63	1560	CBSA
23	3.19	4.37	845.18	682.5	19.21	1075.8	2630.2	CBSA
24	5.65	2.68	851.11	621.5	311.1	754.33	2582.1	CBSA
25	6.27	2.18	732.16	613.4	12.54	788.99	2212.8	CBSA
26	110.6	62.38	242.51	355.1	247.4	367.7	1202	LA
27	341	97.16	505.66	1019	621.4	324.9	2746	LA
28	402.6	64.75	516.53	1050	638.7	305.12	2825	LA
29	375.9	91.2	262.78	1194	361.5	76.27	2432.5	LA
30	374.9	66.52	323.01	978.3	273.8	301.44	2167.3	LA
31	378	81.23	340.48	1079	262.2	271.54	2412.5	LA
32	369.5	81.35	363.7	1089	318	220.28	2331.5	LA
33	296.8	116.5	278.18	972.4	293.9	169.64	2047.6	LA
34	321.6	114.6	523.08	952.5	558.1	243.62	2591.7	LA
35	235.1	98.84	460	996.2	225	334.39	2182	LA
36	324.2	77.34	467.62	908.3	594	298.33	2670	LA
37	369.5	81.35	363.7	1089	318	220.28	2331.5	LA
38	4.15	7.61	129.02	55.42	37.32	150.78	360	LA
39	297.2	113.9	366.67	897.4	488.4	238.47	2306	LA
40	268.5	92.24	531.6	981.1	598.9	244.3	2594	LA
41	66.25	112.7	680.75	964.6	324.3	409.47	2353	LA
42	277.4	120.9	514.82	1079	531.9	270.32	2666	LA
43	231.4	140.5	380.9	729.8	447.5	308.09	2089	LA
44	62.49	202	375.22	641.3	416.1	291.44	1843	LA
45	77.8	210.5	406.89	697.8	440.8	349.73	2009	LA

Table 1: Continued.

Num	Ca^{2+} (mg/L)	Mg^{2+} (mg/L)	$Na^{+}+K^{+}$ (mg/L)	Cl^{-} (mg/L)	SO_4^{2-} (mg/L)	HCO_3^{-} (mg/L)	TDS (mg/L)	Water source type
Validation set samples								
Y1	117.4	23.97	471.5	905.8	23.63	39.66	1562	FA
Y2	12.1	15.19	160.59	35.5	165.9	218.02	523	FA
Y3	16.88	11.82	759.57	631.5	96.89	658.19	1972.8	CBSA
Y4	8.42	4.86	860.25	496.3	11.53	1139.2	2021	CBSA
Y5	3.68	4.33	699.52	596.4	2.88	818.46	1735	CBSA
Y6	335.7	76.85	451.06	1096	330.9	250.13	2439.3	FA
Y7	274.3	86.8	563.71	1026	538.3	317.55	2647	FA
Y8	292.7	93.61	464.6	933.1	593.8	205.09	2495	FA
Y9	311.7	82.94	518.77	915.2	688	296.25	2665	FA

FA: fourth aquifer water; CBSA: coal-bearing sandstone aquifer water; FA: limestone aquifer water.

in the Renlou Coal Mine. Of these, 45 samples were used to establish the water source discrimination model and nine were used to validate the model. Of the 45 water samples used for modeling, nine were from the fourth aquifer, 16 were from the coal-bearing sandstone aquifer, and 20 were from the limestone aquifer. Of the nine samples used for validation, two were from the fourth aquifer, three were from the coal-bearing sandstone aquifer, and four were from the limestone aquifer. The water sample locations are shown in Figure 1. Samples were collected through underground drainage holes or surface hydrological observation holes. The underground drainage holes were directly collected from the mine, and the surface hydrological observation holes were collected with a self-made deep-water sampler. When collecting samples, a 2.5 L polyethylene bucket was rinsed with sampling water three times before a sample was taken. After that, the samples were kept in a clean place to prevent contamination and placed in a low-temperature environment to inhibit the oxidation-reduction reaction and biochemical effects. Water sample chemical testing included $K^{+}+Na^{+}$, Ca^{2+}, Mg^{2+}, Cl^{-}, SO_4^{2-}, HCO_3^{-}, and TDS. HCO_3^{-} was tested using the acid-base titration method, Cl^{-} and SO_4^{2-} were tested using ion chromatography, Ca^{2+} and Mg^{2+} were tested using EDTA titration, $K^{+}+Na^{+}$ was tested by flame atomic absorption spectrophotometry, and TDS was calculated according to the mass concentration of each component. The water sample test data are shown in Table 1.

3.2. Hierarchical Clustering Analysis.

3.2. Hierarchical Clustering Analysis. Hierarchical clustering analysis is an unsupervised identification method that can group samples based on the data itself without known category information. The basic idea is to first treat n samples as n classes, and then specify the distance between samples and the distance between classes. Then, select the two classes with the smallest distance, merge them into a new class, and calculate the distance between the new class and other classes. Continue reducing the number of classes in this way, until all samples are clustered into one class, to obtain a classification system from small to large that can reflect the close relationships between individuals or groups and use a cluster dendrogram to represent them [33]. Classes with stronger correlations are therefore merged, and then the degree of affinity between a new merged class and other classes is considered, and then merged, so that differences within categories are as small as possible and differences between categories are as large as possible.

Generally, R-type and Q-type cluster analyses are used. The R-type cluster analysis classifies variables, and the Q-type cluster analysis classifies samples [34]. If you are interested in the mathematical basis of hierarchical clustering analysis, you can find it in the literature [35, 36]. In this paper, the Wald method was used to perform Q-type clustering analysis on the original water samples, and the square Euclidean distance was used as the metric to determine the relationships between them by using the statistical software IBM SPSS Statistics 26. Finally, the cluster dendrogram of the original water samples was obtained. The data were screened to eliminate water samples that did not meet the requirements.

3.3. Partial Least Squares-Discriminate Analysis. PLS-DA is a supervised multivariate statistical method that integrates the basic functions of principal component analysis, canonical correlation analysis, and multiple regression analysis [37] and is capable of compressing data and extracting characteristic information. The principle of PLS-DA is to separately train the characteristics of different samples, generate a training set, and test the reliability of the training set. This method can group the required observation variables in advance and perform statistical analysis on the data according to the nature of the groups, and the key variables that affect the grouping can be learned [38].

Based on the PLS regression, PLS-DA inputs class member information provided by the auxiliary matrix in the form of code when constructing the factors, uses the independent variable matrix X and the categorical variable Y from the training set samples to establish a regression model, and determines the sample category based on its predicted PLS value. It also reduces the dimensionality of the high-dimensional data matrix to a lower-dimensional space. Similar to PCA, the new variables obtained are also not related to each other, but the difference is that PLS needs to introduce the information from category matrix Y into matrix X while decomposing the independent variable matrix X, and then perform orthogonal decomposition. This processing can effectively eliminate any useless noise in the independent variable matrix X and any useless information in category matrix Y. The use of this method to analyze mine water inrush water sources can eliminate overlapping parts from water chemical information to solve the multiple correlation problem and make the data more accurate and reliable to ensure the best calibration model [30]. If you are interested in the mathematical basis of partial least squares (PLS) regression, you can find it in the literature [39]. The specific steps of the PLS-DA analysis method are as follows and this method can be completed by SIMCA 14.1 software.

(1) Establish categorical variables of training set samples

(2) Decompose the independent variable matrix X and the category matrix Y at the same time, and ensure their principal components are linearly correlated to the highest degree. The model can be expressed as

$$\begin{aligned} X &= TP^T + E, \\ Y &= UQ^T + F, \end{aligned} \tag{1}$$

where T and U are respective score matrices of X and Y; P and Q are respective load matrices of X and Y; E and F are respective fitting residual matrices of X and Y.

(3) Conduct linear regression on T and U to obtain regression factor B

$$\begin{aligned} U &= TB, \\ B &= \left(T^T\right)^{-1} T^T U. \end{aligned} \tag{2}$$

(4) According to the load matrix P, obtain the score vector t_{test} of the sample x_{test} to be tested during prediction, and then obtain the predicted value Y_P according to the following formula

$$Y_P = t_{\text{test}} BQ. \tag{3}$$

(5) Determine the type of sample to be tested according to the following rules

When $Y_P > 0.5$ and deviation $D < 0.5$, it belongs to this category; when $Y_P < 0.5$ and deviation $D < 0.5$, it does not belong to this category; when deviation $D \geq 0.5$, it is uncertain [40].

4. Results and Discussion

4.1. Processing and Analysis of Water Sample Data

4.1.1. Water Sample Screening Using the Q-Type Cluster Analysis Method. With either method, the water samples have to be screened before establishing the recognition model to identify water sources [39]. There were three main reasons for screening the original water samples. First, in order to reduce the impact of external human factors (such as possible contamination of water samples during sampling, storage and testing, measurement deviations during testing, etc.), it was necessary to screen the original water samples to eliminate unqualified samples and avoid large errors in the discrimination results [33].

Second, although the water chemical composition within an aquifer may be significantly different due to different hydrogeological conditions, it should maintain a dynamic balance through a series of physical and chemical reactions. Therefore, samples from the same aquifer generally have the same water chemistry characteristics. Due to the influence of factors such as hydraulic connections between different aquifers and groundwater movement, however, the water chemical composition from the same aquifer sometimes differ greatly. Abnormal water chemical compositions in the same aquifer cannot reflect the hydrochemical characteristics of underground water in this aquifer. We therefore had to identify the water sample that best represented the aquifer water chemical composition and establish a high-precision water inrush water source discrimination model [25].

Third, PLS-DA model performance may have deteriorated due to the presence of abnormal sample values. In order to reduce the influence of abnormal samples on the PLS-DA model, the original water samples were also screened [41].

Before screening the water samples, we performed the Piper trilinear diagram analysis on 45 original water samples, as shown in Figure 2. It can be seen from Figure 2 that among the 45 original water samples of three different water sources in the study area, some of the water samples of the same type were scattered and significantly deviated from the formation center in the Piper trilinear diagram and these samples should be regarded as abnormal water samples and excluded. Hierarchical clustering analysis is a commonly used unsupervised agglomerative clustering analysis method that can be used for this task [33]. In this paper, the ion contents of 45 original water samples were used as the analysis variable, and the Q-type cluster analysis of the original water sample was completed by SPSS software. A clustering dendrogram for the samples was obtained (Figure 3). According to the distance of each original water sample in the dendrogram, each original water sample category was compared, and water samples with numbers 2, 3, 4, 17, 26, 38, and 41 were excluded. The remaining 38 original water samples were kept for subsequent analysis and modeling, as shown in Table 1.

4.1.2. Correlation Analysis of Water Chemical Compositions. Ion concentrations in the water samples of each aquifer reveal the chemical characteristics of different groundwater sources and are the basis for distinguishing water from each aquifer. These kinds of hydrochemical components are not completely independent in groundwater, but are related to each other to a certain degree. However, most prior studies have not considered this connection [42].

This paper used Python software to draw heat maps of the correlation coefficients between water chemical compositions of water samples from three aquifers (Figure 4). There

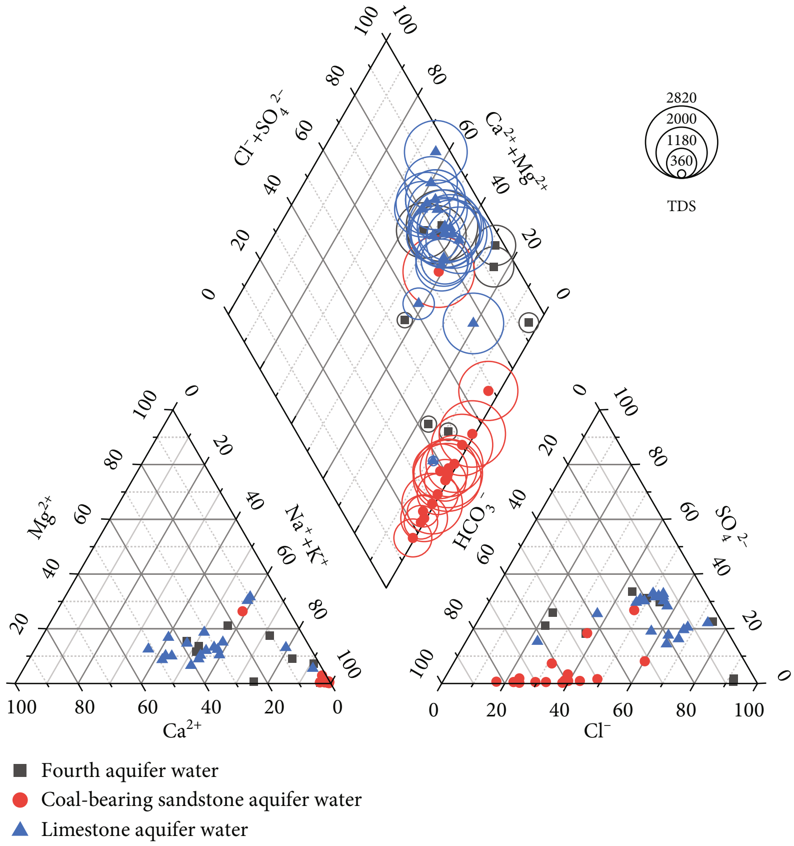


Figure 2: Piper trilinear diagram of the water samples from different water sources.

were both positive and negative correlations as well as strong and weak correlations between the evaluated ions, and some ions had strong correlations with each other. For example, the correlations between Mg^{2+} and SO_4^{2-}, as well as between Cl^- and Na^++K^+, Ca^{2+}, HCO_3^-, and TDS values in the water samples from the fourth aquifer were all greater than 0.8 (Figure 4(a)). The correlations between Na^++K^+ and TDS, as well as between Cl^- and Na^++K^+ and TDS values in the water samples from the coal-bearing sandstone aquifer were all greater than 0.8 (Figure 4(b)). Finally, the correlations between Mg^{2+} and Cl^-, as well as between Ca^{2+} and Mg^{2+} and Cl^- in the water samples from the limestone aquifer were all greater than 0.8 (Figure 4(c)). This indicated that the hydrogeological information reflected between water chemical compositions overlapped. If this kind of information overlap was not considered in water source identification, it would cause information redundancy, which can cause serious multicollinearity, affect the accuracy of the mine water inrush source identification model, and lead to poor judgment [43].

4.2. PCA of the Training Samples. PCA is an unsupervised multivariate statistical method, which is one of the most commonly used dimensionality reduction methods. Through orthogonal transformation, multiple indicator data are converted into a set of linear and uncorrelated few new comprehensive variables. PCA is helpful to analyze hydrochemical data and can be considered on hydrochemical data to screen the variation between composition and sample variation [44].

This study imported the water chemistry data of 38 water samples into the SIMCA 14.1 software for principal component analysis. The analysis results show that the eigenvalues of the first two principal components were greater than 1 (the first and second principal components were 3.85 and 2.53, respectively), and the cumulative contribution rate reached 91.1%, which means that the selection of two principal components can fully reflect the hydrochemical information of the training samples [33]. Therefore, using the first and second principal components as the abscissa and ordinate, respectively, the PCA score plot (Figure 5) and PCA loading plot (Figure 6) of the three different water sources were obtained. The PCA score plots can explain the variation among sample sources, and loading plots can explain the variation among compositions.

It can be seen from the PCA score plots (Figure 5) that the first principal component scores of the water samples of the fourth aquifer water, coal-bearing sandstone aquifer water, and limestone aquifer water ranged from -56.32 to -5.54, -47.56 to 8.22, and 2.00 to 50.60, respectively. The second principal component scores of the water samples of the fourth aquifer water, coal-bearing sandstone aquifer water, and limestone aquifer water ranged from -47.64 to -22.45, -3.00 to 56.67, and -23.24 to 0.20, respectively. Therefore, PCA can roughly divide water samples from three different water sources into three categories. However, it was

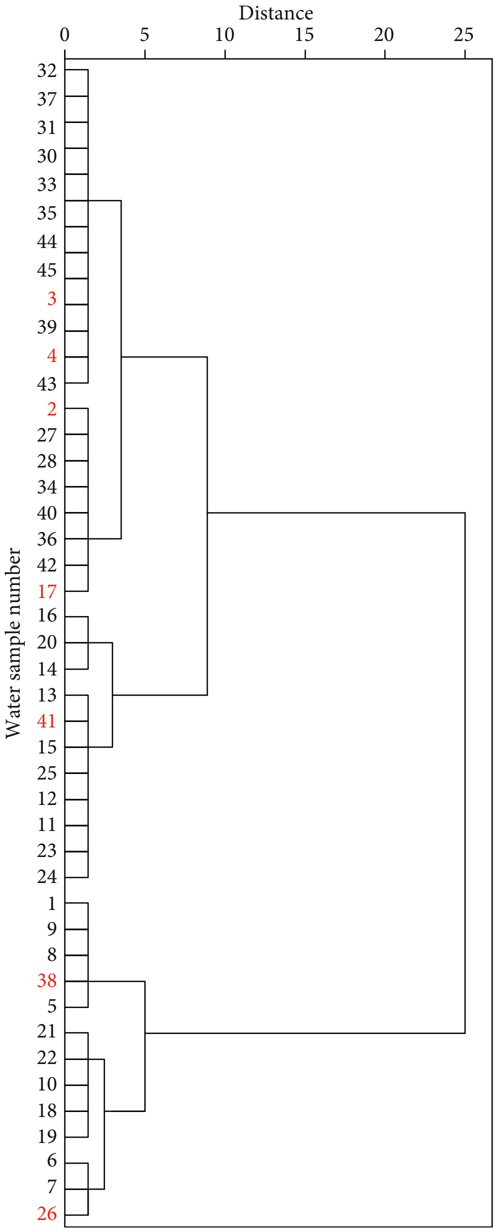


Figure 3: Q-type cluster analysis results of original water samples.

impossible to distinguish the three water sources based on the first or second principal component alone.

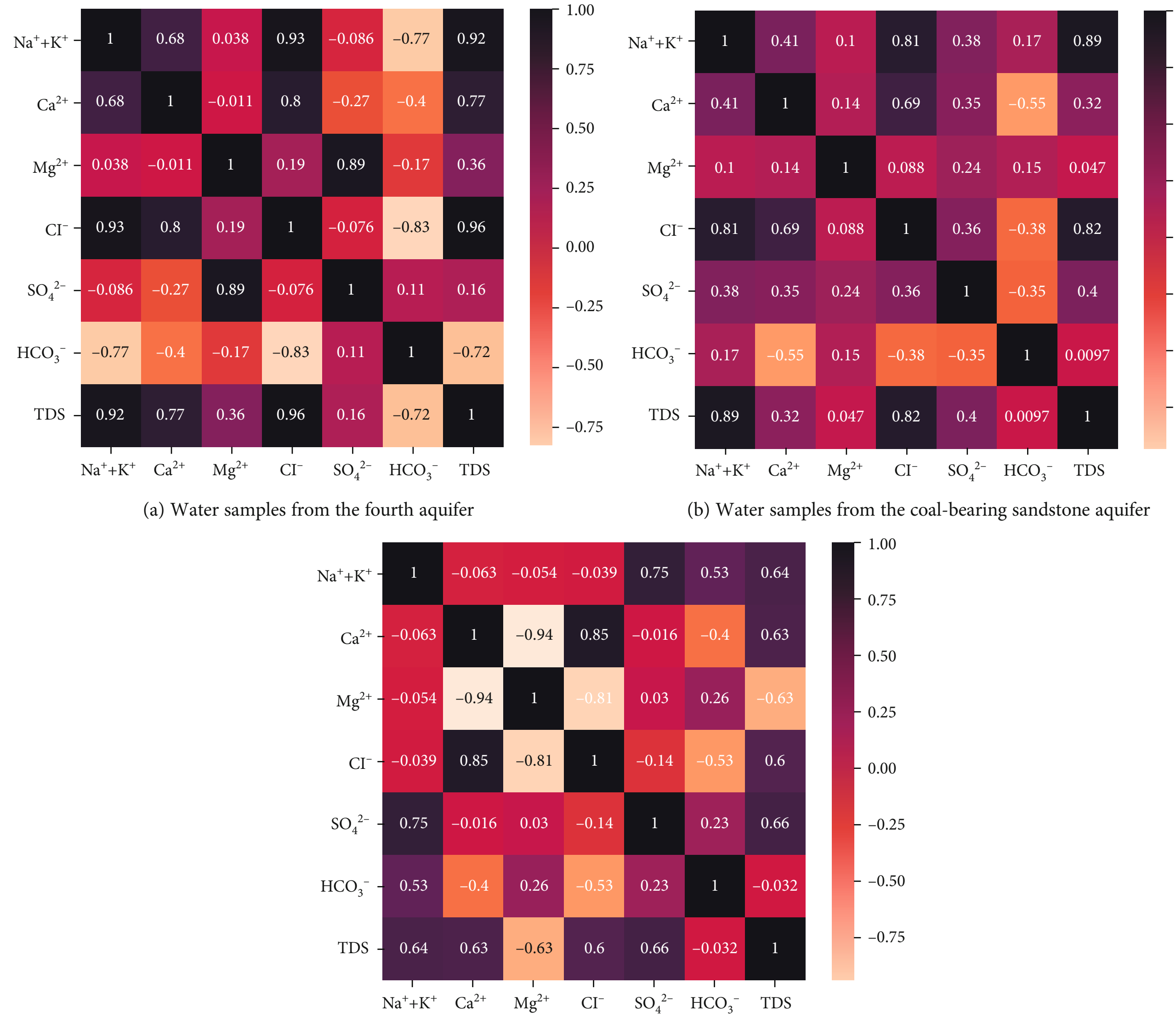
It can be seen from the PCA loading plot (Figure 6) that HCO_3^-, Na^++K^+, and TDS were far from the origin, indicating that these three water chemical composition variables played a greater role in water source identification. The first principal component was mainly composed of TDS, SO_4^{2-}, and Cl^-, and the second principal component was mainly composed of HCO_3^-, Na^++K^+, and TDS.

4.3. PLS-DA Discriminant Model Establishment Based on Water Chemical Compositions for Mine Water Inrush Sources. On the basis of principal component analysis, the PLS-DA method was used to further analyze the water chemistry data of different water sources to discover and screen out the characteristic water chemical components, and establish the PLS-DA discrimination model based on water chemical compositions for mine water inrush sources.

4.3.1. Determining Classification Variable Values. The PLS-DA discriminant model for mine water inrush sources is a PLS-based regression model between the classification variables and the ion component content of the water samples. This paper used SIMCA 14.1 software to establish and analyze the PLS-DA model. Taking 38 water samples as the training set, first the classification variable values of the training set samples were assigned. The classification variable group Y was manually set according to the water sample category, as shown in Table 2. Then, the PLS method was used to perform regression analysis on the content of the seven ion components for the training set samples and the classification variable Y, and a model of the ion components and the classification variable Y was established.

4.3.2. Determining the Number of Principal Components. When modeling, the appropriate number of principal components must be determined. Generally speaking, increasing the number of principal components can extract more information, but using too many principal components will introduce some redundant information [40]. When selecting the number of principal components, therefore, the cumulative explanatory power (expressed by R2X(cum)) and the prediction accuracy of the model (expressed by cumulative cross-validity Q2(cum)) should be considered [40, 45].Table 3 shows the relevant statistical results when the number of principal components was used for modeling. When there were five principal components, the cumulative cross-validity value Q2(cum) began to decrease, so the prediction accuracy of the model decreased. Therefore, the appropriate number of principal components was four.

4.3.3. Analysis of Discriminant Model Results. The model quality parameter R2X(cum) is 0.979, R2Y(cum) is 0.889, and Q2(cum) is 0.848, indicating good model fit [46]. In the model space, the first and second principal component scores for the water samples are shown in Figure 7. Each point in the PLS-DA model score map represents a water sample, and the degree of aggregation reflects the similarity between them. The results of PLS-DA analysis were consistent with the results of PCA analysis. All data points were within the 95% confidence interval, and the water samples of the three aquifers had obvious clustering. However, the number of water samples in the fourth aquifer was relatively small, and the distribution dispersion was relatively large. At the first principal component $t[1]$, the limestone aquifer water samples were easily distinguished from fourth aquifer water samples and coal-bearing sandstone aquifer water samples, but it was impossible to further accurately distinguish the water samples between the fourth aquifer and coal-bearing sandstone aquifer. At the second principal component $t[2]$, the fourth aquifer samples were easily distinguished from coal-bearing sandstone aquifer water samples and limestone aquifer water samples, but it was impossible to further accurately distinguish between the sandstone and limestone samples. The schematic diagram of the PLS-DA



(a) Water samples from the fourth aquifer

(b) Water samples from the coal-bearing sandstone aquifer

(c) Water samples from the limestone aquifer

Figure 4: Heat maps of hydrochemical composition correlation coefficients.

model in three-dimensional space using the first principal component $t[1]$, the second principal component $t[2]$, and the third principal component $t[3]$ of the water samples (Figure 8) showed that the three-dimensional space diagram could significantly distinguish the water samples from the three different sources.

Figure 9 shows the loading scatter plot of the PLS-DA model, which clearly demonstrates the relationship between the characteristic variable X and the categorical variable Y, reflecting the contribution of each water chemical composition variable on the score plot. Blue dotted DA (100), DA (010), and DA (001) in Figure 9 represent the positions of the Y values for the three water source categories in the scatter plot, and each green point represents an ion variable. The farther the point is from the origin, the greater the weight value, or the greater the effect of determining the sample difference [47]. It can be seen in Figure 6 that TDS, HCO$_3^-$, and Na$^+$+K$^+$ were far from the origin, indicating that these three water chemical composition variables played a greater role in water source identification. On the first principal component $t[1]$, the loading values of HCO$_3^-$, SO$_4^{2-}$, and Cl$^-$ were larger, and on the second principal component $t[2]$, the loading values of TDS, HCO$_3^-$, and Na$^+$+K$^+$ were larger. Therefore, the first principal component mainly reflected the content characteristics of HCO$_3^-$, SO$_4^{2-}$, and Cl$^-$ in different water sources, and the second principal component reflected the content characteristics of TDS, HCO$_3^-$, and Na$^+$+K$^+$ in different water sources.

Compared with PCA, PLS-DA has the function of quantifying the difference of different water sources caused by the characteristic chemical compositions of water. To further analyze the effect of each water chemical composition variable X on the categorical variable Y, a VIP score plot was created (Figure 10). It summarizes the importance of the variables to explain X and correlate to Y. VIP scores can quantify the contribution of each variable in the PLS-DA

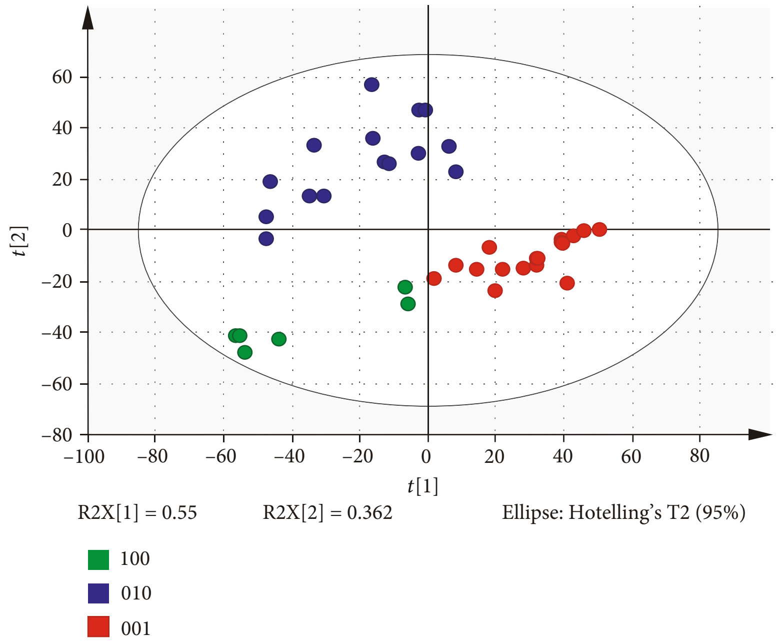


Figure 5: Score plot of PCA. 100: fourth aquifer water; 010: coal-bearing sandstone aquifer water; 001: limestone aquifer water.

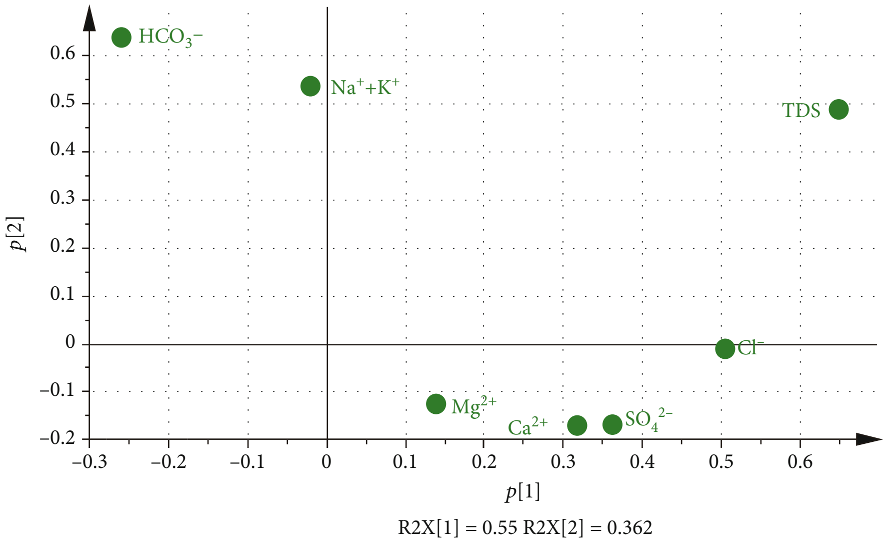


Figure 6: Loading scatter plot of PCA. 100: fourth aquifer water; 010: coal-bearing sandstone aquifer water; 001: limestone aquifer water.

Table 2: Classification variables of different water source types.

Water source type	Fourth aquifer water	Coal-bearing sandstone aquifer water	Limestone aquifer water
Classification variables	[100]	[010]	[001]

model to the classification. The larger the VIP value, the more obvious the difference of the variable in different water source categories. When the VIP value of a variable is greater than 1.0, it indicates a higher than average contribution of the variable to the overall model with a statistically significant impact on the water sample classification, which can be used as the difference marker composition [47]. When the value is less than 0.5, it indicates that the variable is unimportant in the process of model classification and discrimination. The interval between 1 and 0.5 is a gray area, where the importance level depends on the size of the data set. Figure 7 shows that in the explanatory water chemical composition variable X, there were two variables with VIP scores greater than 1, followed by HCO_3^- and TDS, indicating that HCO_3^- and TDS played an important role in distinguishing three different types of water sources. The VIP scores for Cl^-, Na^++K^+, and SO_4^{2-} were between 0.9 and 1.0, indicating that these three ion variables played roles in distinguishing three

Table 3: Relevant statistical results of modeling with different principal component numbers.

Component	R2X	R2X(cum)	Eigenvalue	R2Y	R2Y(cum)	Q2	Limit	Q2(cum)
1	0.516	0.516	3.61	0.494	0.494	0.482	0.05	0.482
2	0.395	0.911	2.76	0.3	0.794	0.577	0.05	0.781
3	0.0399	0.951	0.279	0.0781	0.872	0.277	0.05	0.842
4	0.0278	0.979	0.195	0.0168	0.889	0.0389	0.05	0.848
5	0.0152	0.994	0.106	0.00415	0.893	-0.0745	0.05	0.836
6	0.00521	0.999	0.0365	0.000976	0.894	-0.0658	0.05	0.826
7	0.00102	1	0.00717	0.00307	0.897	-0.018	0.05	0.823

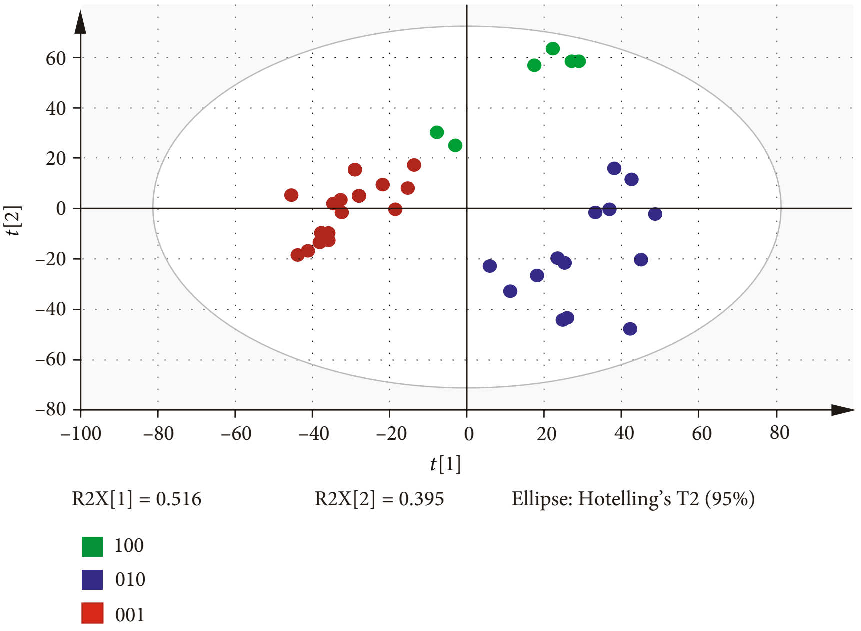


Figure 7: Score plot of PLS-DA. 100: fourth aquifer water; 010: coal-bearing sandstone aquifer water; 001: limestone aquifer water.

different water source types to some degree. The VIP score for Mg^{2+} was the lowest among the seven hydrochemical components, indicating that Mg^{2+} played the least important role in discrimination.

The statistical results of the PLS-DA discriminant model for 38 water samples are shown in Table 4 and Figure 11. Table 4 shows good correlations between the water sample hydrochemical composition variables and the categorical variables established by PLS regression. The correlation coefficients R between the actual values of the categorical variables and the predicted values of the model were 0.8876, 0.9608, and 0.9778, respectively. Root mean square error of estimation (RMSEE) is an index to predict the average error of training set samples by using the model built by training set samples. Root mean square error of cross validation (RMSEcv) is an important parameter in internal cross validation, which is used to measure the accuracy of prediction results of training set samples. The discriminant rate of all training set samples was 100%, indicating that the model fit well.

Figure 11 shows regression curves for the PLS predicted values and actual values of classification variables for all training samples. The straight lines are the regression curves of the model prediction and classification results. The three models clearly distinguished the three types of water source samples: water sample points scattered on the line where the actual values were equal to 1 and the other two water source water points on the line where the actual values were equal to 0 were obviously separated. The PLS-DA model established in this study had high reliability and can be used to test and discriminate new water samples.

4.4. PLS-DA Discriminant Model Validation for Mine Water Inrush Sources

4.4.1. Permutation Test. Statistical inference analysis was used to further validate the built PLS-DA discriminant model. Two hundred permutation tests were performed to iteratively analyze the predictor variable Y based on the known measured data variable X and obtain statistics on these variables (Figure 12). By examining the intercept of the fitting line formed by the calculated values of R2 and Q2 corresponding to all samples on the Y coordinate axis, the reliability of the model and the degree of overfitting were determined. The larger the value of Q2, the better the predictive ability of the model, and the larger the value of R2, the

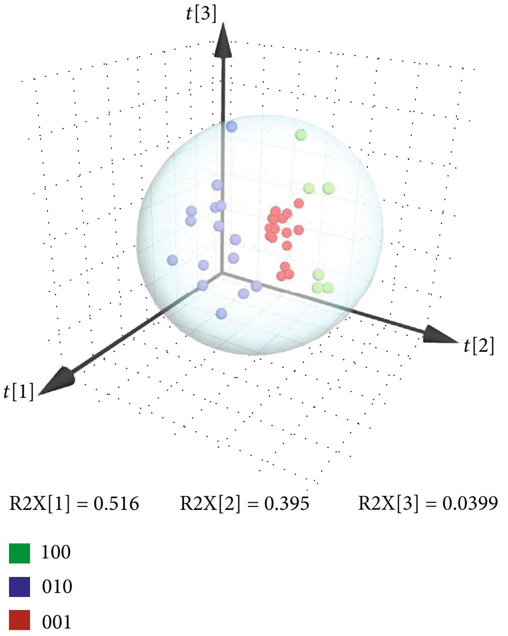


Figure 8: Schematic diagram of the PLS-DA model in three-dimensional space. 100: fourth aquifer water; 010: coal-bearing sandstone aquifer water; 001: limestone aquifer water.

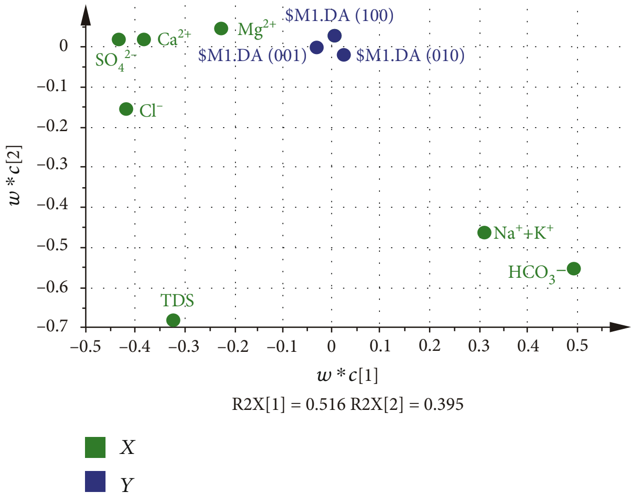


Figure 9: Loading scatter plot of the PLS-DA model.

stronger the explanatory ability [48]. The permutation tests (Figure 12) showed that all R2 and Q2 values (Y-axis data) on the left were lower than the R2 and Q2 values on the far right. The intercepts of the Q2 regression line were all negative, indicating that although there were differences in predictability, the three PLS-DA discriminant models established were not overfitted and all had good predictive ability [49]. Therefore, they can be used for discriminant analysis of various types of water sources.

4.4.2. External Validation. The actual predictive ability of the model was further checked by an external validation set test. The validation set was composed of nine water samples that were not involved in the modeling, including two fourth aquifer water samples, three coal-bearing sandstone aquifer water samples, and four limestone aquifer water samples. The categorical variable predicted values Y_p of the verification set water samples were calculated using SIMCA 14.1 software, and the prediction results were evaluated according to rules described above. The results are shown in Table 5. The accuracy of this model for the validation set water samples was 100%.

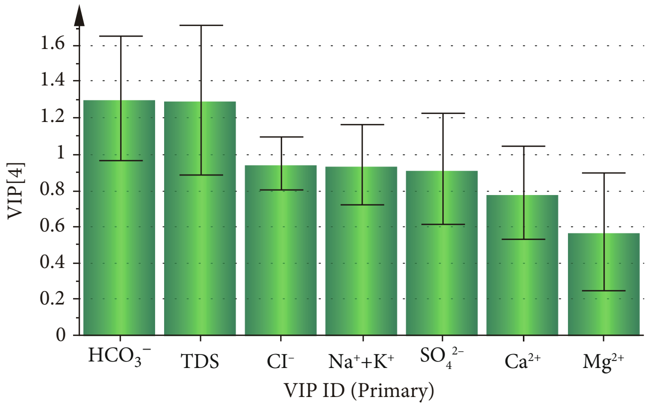


Figure 10: VIP score plot of PLS-DA.

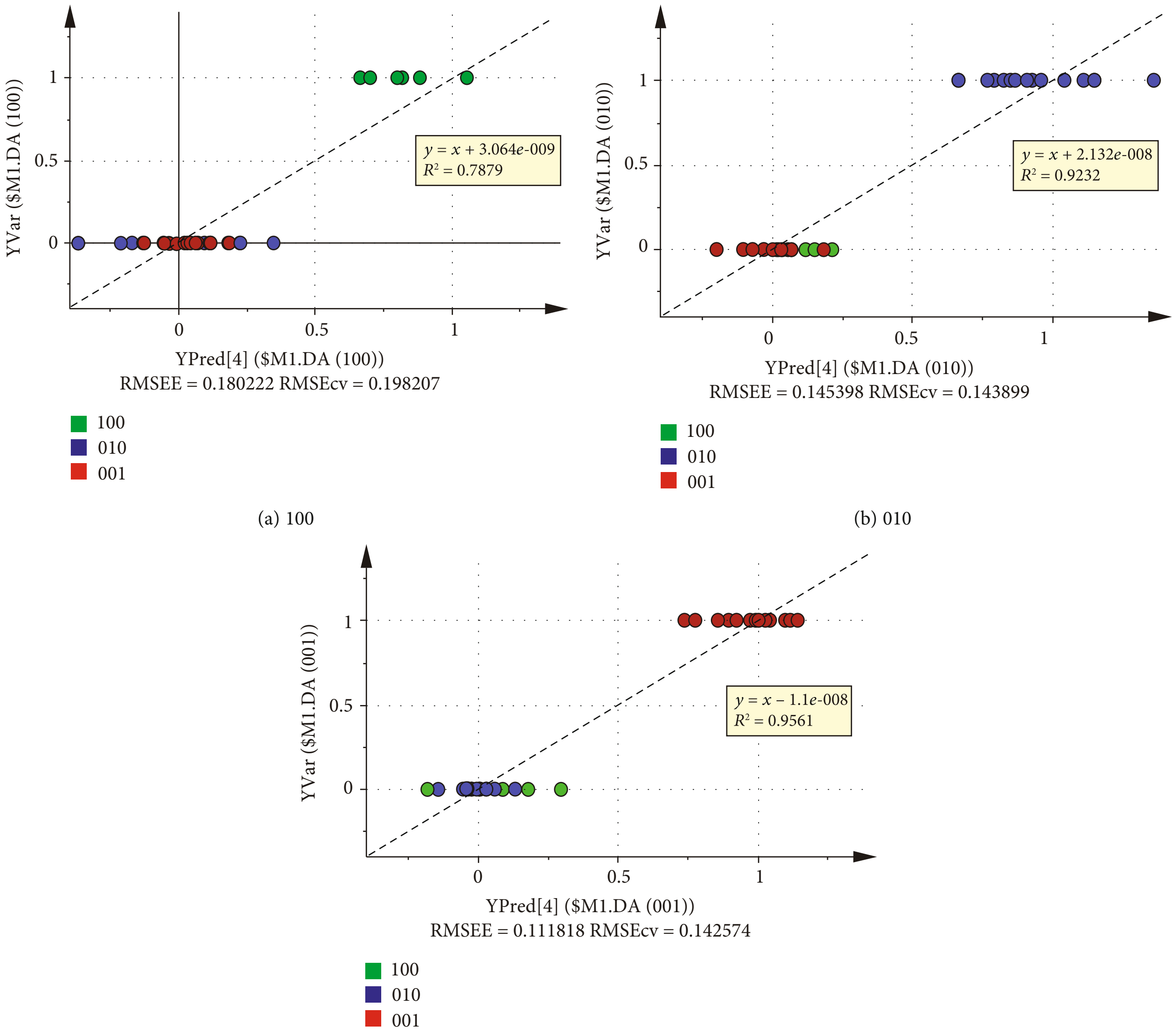
5. Discussion

This study innovatively combined the PLS-DA method with the water chemical compositions to establish a discriminant model to be used in the identification of water inrush sources in mines, which effectively solved the problem of low discrimination accuracy caused by not considering the overlapping information between hydrochemical identification indexes. The water sample data were processed by PCA and PLS-DA methods, and the results showed that both methods can classify water samples from three different water sources (as shown in Figures 5 and 7). However, with regard to the two principal components extracted by the PCA method, it was impossible to distinguish any one of the three water sources based on the first or second principal component alone. For the PLS-DA results, the limestone aquifer water samples were easily distinguished from fourth aquifer water samples and coal-bearing sandstone aquifer water samples based on the first principal component, and the fourth aquifer samples were easily distinguished from coal-bearing sandstone aquifer water samples and limestone aquifer water samples based on the second principal component. This showed that PLS-DA has better data processing and analysis capabilities than PCA. The reason is that PLS-DA is a supervised discriminant analysis method, which artificially adds grouping variables, further excavates the information in the water sample data, strengthens the difference of water chemical composition between different water sources, and makes up for the deficiency of the PCA method [28].

In addition, compared with the PCA method, PLS-DA has the function of quantifying the degree of difference between different water sources caused by the characteristic water chemical composition. The loading scatter plots of PCA and PLS-DA both showed that TDS, HCO_3^-, Na^++K^+,

Table 4: Discrimination results of the PLS-DA model of the water sample training set.

Types	Number of water samples	Correlation coefficient R^2	RMSEE	RMSEcv	Discrimination accuracy
Fourth aquifer water	6	0.7879	0.180222	0.198207	100%
Coal-bearing sandstone aquifer water	15	0.9232	0.145398	0.143899	100%
Limestone aquifer water	17	0.9561	0.111818	0.142574	100%



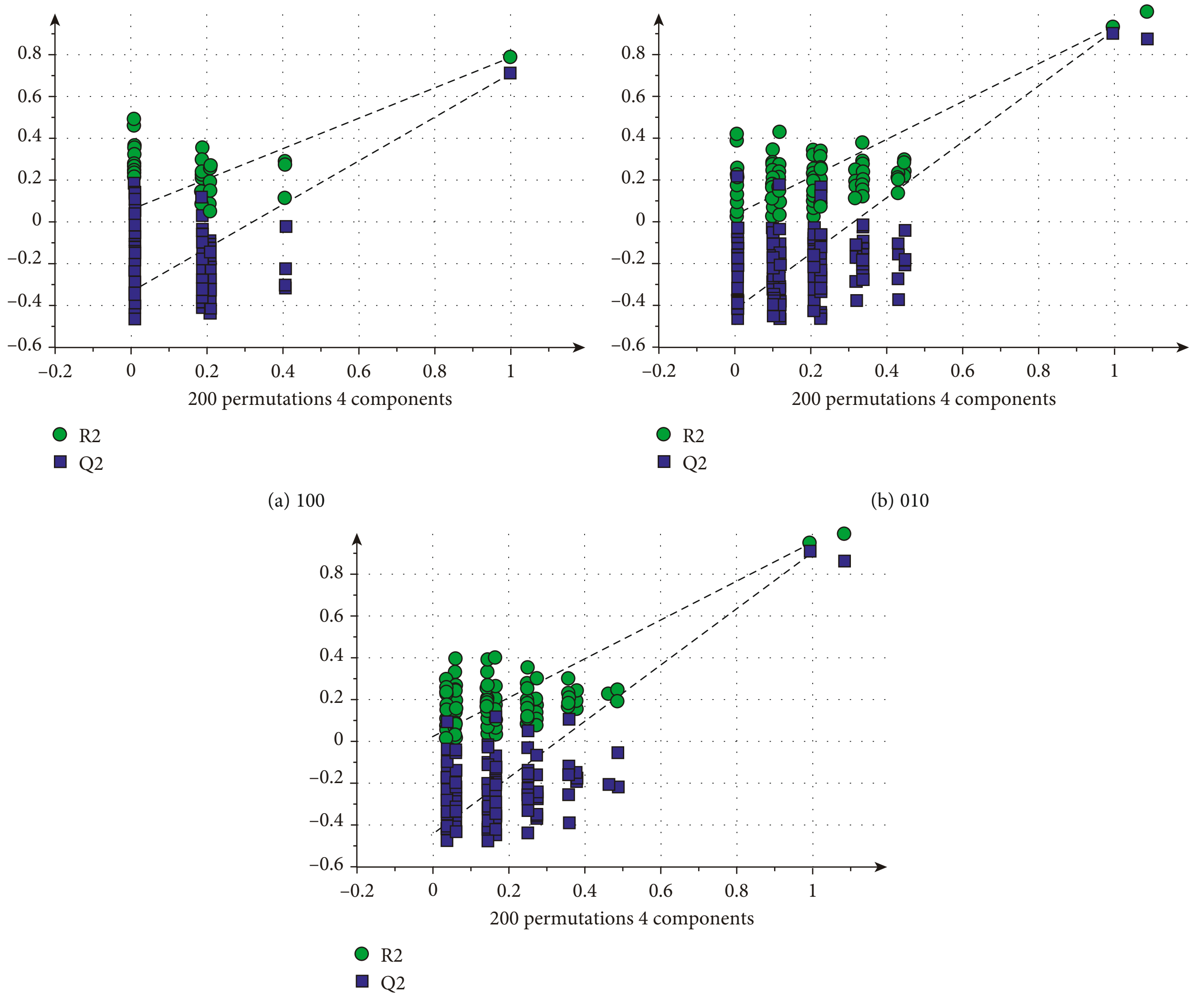
(a) 100

(b) 010

(c) 001

Figure 11: The regression diagrams for PLS predicted and actual values of the categorical variables from the PLS-DA model training set. 100: fourth aquifer water; 010: coal-bearing sandstone aquifer water; 001: limestone aquifer water.

Cl^-, and SO_4^{2-} were factors that had a greater impact on the discrimination results of the three different water sources (as shown in Figures 6 and 9), but the degree of influence of each factor could not be accurately determined. Through the VIP scores in the PLS-DA method, we can accurately screen out the characteristic water chemical components that cause differences in different water sources. Through the analysis of the VIP scores, two difference marker compositions were found from the seven water chemical compositions, followed by HCO_3^- and TDS, indicating that HCO_3^- and TDS were the main marker compositions that distinguished the difference between the fourth aquifer water, the coal-bearing sandstone aquifer water, and the limestone aquifer water in the study area [49]. The influence of other water chemical components on the water source identification results were Cl^-, Na^++K^+, SO_4^{2-}, Ca^{2+}, and Mg^{2+},



(a) 100

(b) 010

(c) 001

Figure 12: Permutation test validation of the PLS-DA model. 100: fourth aquifer water; 010: coal-bearing sandstone aquifer water; 001: limestone aquifer water.

Table 5: Discrimination results of the PLS-DA model for the validation set water samples.

Number	Predicted value of categorical variables			Types		Discrimination accuracy
	Y_{100}	Y_{010}	Y_{001}	Actual type	Predicted type	
V1	0.7270	0.2027	0.0704	FA	FA	100%
V2	0.9309	0.0562	0.0129	FA	FA	
V3	0.2902	0.7845	-0.0747	CBSA	CBSA	100%
V4	-0.0872	1.1500	-0.0627	CBSA	CBSA	
V5	0.2095	0.8554	-0.0649	CBSA	CBSA	
V6	0.0329	0.0664	0.9007	LA	LA	100%
V7	-0.0077	0.1252	0.8825	LA	LA	
V8	0.0875	-0.0403	0.9528	LA	LA	
V9	-0.0167	0.0106	1.0061	LA	LA	

FA: fourth aquifer water; CBSA: coal-bearing sandstone aquifer water; LA: limestone aquifer water.

respectively. Quickly determining the iconic ionic components in each aquifer not only is conducive to accurately and quickly identifying mine water inrush sources but also furthers research on the formation and evolution of aquifers. However, considering the complexity of hydrogeological conditions in different mines, the difference marker ions will vary. Therefore, more ionic components should be tested, and factors such as water temperature, isotopes, and trace elements should also be considered in future studies to perfect the discriminant model as much as possible, so that it can be better used to identify the source of water inrush in mines [26].

Furthermore, as a supervised model, the PLS-DA model has the disadvantage of overfitting, so the model can distinguish samples well but performs poorly when used to predict new sample sets. Therefore, for the supervised classification model, we need to verify the reliability of the model [40]. In this study, only seven water chemical compositions were tested and used as the identification index combined with the PLS-DA method to establish a mine water inrush source discrimination model. A good discrimination effect was achieved with discrimination rates for the training and validation sets as high as 100%, which indicated that the PLS-DA discriminant model for mine water inrush sources performed better in identifying water samples. We used permutation testing to judge the reliability of the model. The permutation test randomly scrambles the classification mark of each sample, and then remodels and predicts. The Q2 of a reliable model should be significantly greater than the Q2 obtained by randomly scrambling the data. The results of the permutation test showed that the model had no overfitting and was reliable [46], which indicated that the established water source recognition model was successful.

6. Conclusions

Based on the hydrogeological conditions of the study area, water chemical compositions of water inrush samples from three aquifers were tested. The water samples were screened using the hierarchical clustering analysis method, and some unqualified samples were removed. The PCA and PLS-DA methods were used to analyze and process the remaining water sample data. On the basis of PCA analysis, a PLS-DA discriminant model for mine water inrush sources was established. According to the results, the following conclusions were obtained:

(1) Hierarchical clustering analysis was used to screen the 45 original water samples and eliminate seven unqualified samples to reduce errors, so the remaining 38 samples well represented the water chemical compositions of the aquifers. The 38 samples were used to establish a discriminant model and avoid the influence of abnormal water samples

(2) Correlation analysis was carried out on the water chemical compositions of the water samples from three aquifers. The results showed that there were strong correlations between some water chemical compositions, indicating that the hydrogeological information reflected between the water chemical compositions had a significant overlap. It would cause information redundancy, which could lead to multicollinearity

(3) The PCA and PLS-DA methods were used to analyze and process the remaining water sample data, and the results showed that both methods can distinguish water samples from different water sources; however, the classification effect of PLS-DA was better than PCA. The reason is that PLS-DA is a supervised discriminant analysis method, which artificially adds grouping variables, further excavates the information in the water sample data, strengthens the difference of water chemical composition between different water sources, and makes up for the deficiency of the PCA method

(4) The PLS-DA discriminant model for mine water inrush sources was established. The correct discrimination rate of the PLS-DA discriminant model was as high as 100%, and permutation tests showed that the model was not overfit. External validation found that the model had good stability and discrimination

(5) PLS-DA has the function of quantifying the degree of difference between different water sources caused by the characteristic water chemical composition. VIP scores were used to identify the most important difference marker compositions that affected the discrimination results of the three different water source types, followed by HCO_3^- and TDS, while Mg^{2+} had little effect in distinguishing them

(6) The discriminant model established in this study combined the advantages of principal component analysis and multiple regression analysis, and had a high discrimination accuracy. Thus, it can meet the needs of modern mine water inrush source identification, and can be applied to other mines as well

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

The study was supported by the National Natural Science Foundation of China (Grant No. 41272278) and the Scientific Research Platform Innovation Team Construction Project in Universities of Anhui (Grant No. 2016-2018-24). The authors would also like to thank Dr. Ju for his valuable comments and suggestions for the improvement of the manuscript.

References

[1] W. Zeng, Z. Huang, Y. Wu, S. J. Li, R. Zhang, and K. Zhao, "Experimental investigation on mining-induced strain and failure characteristics of rock masses of mine floor," *Geomatics Natural Hazards and Risk*, vol. 11, no. 1, pp. 491–509, 2020.

[2] D. K. Zhao, Q. Wu, F. P. Cui et al., "Using random forest for the risk assessment of coal-floor water inrush in Panjiayao Coal Mine,Northern China," *Hydrogeology Journal*, vol. 26, no. 7, pp. 2327–2340, 2018.

[3] Y. Wu, Z. Huang, K. Zhao, W. Zeng, Q. X. Gu, and R. Zhang, "Unsteady seepage solutions for hydraulic fracturing around vertical wellbores in hydrocarbon reservoirs," *International Journal of Hydrogen Energy*, vol. 45, no. 16, pp. 9496–9503, 2020.

[4] J. C. Wei, G. H. Li, D. L. Xie, G. Y. S. Yu, X. Q. Man, and J. Wang, "Discrimination of mine water-inflow sources based on the multivariate mixed model and fuzzy comprehensive evaluation," *Arabian Journal of Geosciences*, vol. 13, no. 17, pp. 741–749, 2020.

[5] X. Y. Wang, H. Y. Ji, Q. Wang et al., "Divisions based on groundwater chemical characteristics and discrimination of water inrush sources in the Pingdingshan coalfield," *Environmental Earth Sciences*, vol. 75, no. 10, 2016.

[6] L. Ma, J. Z. Qian, W. D. Zhao, Z. Curtis, and R. G. Zhang, "Hydrogeochemical analysis of multiple aquifers in a coal mine based on non-linear PCA and GIS," *Environmental Earth Sciences*, vol. 75, no. 8, pp. 1–14, 2016.

[7] X. L. Duan, F. S. Ma, H. J. Zhao et al., "Determining mine water sources and mixing ratios affected by mining in a coastal gold mine, in China," *Environment and Earth Science*, vol. 78, no. 10, pp. 1–16, 2019.

[8] P. Négrel, E. Petelet-Giraud, J. Barbier, and E. Gautier, "Surface water- groundwater interactions in an alluvial plain: chemical and isotopic systematics," *Journal of Hydrology*, vol. 277, no. 3, pp. 248–267, 2003.

[9] W. H. Sui, J. Y. Liu, S. G. Yang, Z. S. Chen, and Y. S. Hu, "Hydrogeological analysis and salvage of a deep coalmine after a groundwater inrush," *Environmental Earth Sciences*, vol. 62, no. 4, pp. 735–749, 2011.

[10] H. T. Zhang, G. Q. Xu, X. Q. Chen, J. Wei, S. T. Yu, and T. T. Yang, "Hydrogeochemical characteristics and groundwater inrush source identification for a multi-aquifer system in a coal mine," *Acta Geologica Sinica-English Edition*, vol. 93, no. 6, pp. 1922–1932, 2019.

[11] L. Mou, "Application of dynamic curve prediction method in discriminating water-bursting source," *Coal Geology & Exploration*, vol. 44, no. 3, pp. 70–79, 2016.

[12] Q. Wu, W. P. Mu, Y. Xing et al., "Source discrimination of mine water inrush using multiple methods: a case study from the Beiyangzhuang Mine,Northern China," *Bulletin of Engineering Geology and the Environment*, vol. 78, no. 1, pp. 469–482, 2019.

[13] X. L. Zhang, Z. X. Zhang, and S. P. Peng, "Application of the second theory of quantification in identifying gushing water sources of coal mines," *Journal of China University of Mining and Technology*, vol. 32, no. 3, pp. 251–254, 2003.

[14] Y. Wang, W. F. Yang, M. Li, and X. Liu, "Risk assessment of floor water inrush in coal mines based on secondary fuzzy comprehensive evaluation," *International Journal of Rock Mechanics and Mining Sciences*, vol. 52, pp. 50–55, 2012.

[15] L. Gong and C. L. Jin, "Fuzzy Comprehensive Evaluation for Carrying Capacity of Regional Water Resources," *Water Resources Management*, vol. 23, no. 12, pp. 2505–2513, 2009.

[16] M. Qiu, L. Q. Shi, C. Teng, and Y. Zhou, "Assessment of water inrush risk using the fuzzy Delphi analytic hierarchy process and grey relational analysis in the Liangzhuang coal mine, China," *Mine Water & the Environment*, vol. 36, no. 1, pp. 1–12, 2016.

[17] S. Y. Zhang, Y. B. Hu, and S. P. Xing, "Discrimination of the mine water inrush source based on principal component analyses-theory of gray relational degree," *Hydrogeology & Engineering Geology*, vol. 45, no. 6, pp. 36–41, 2018.

[18] Y. Wu and Z. C. Yu, "Application of neural network in water source distinguishing of mine water inrush," *Industry and Mine Automation*, vol. 37, no. 10, pp. 60–62, 2011.

[19] P. H. Huang and J. S. Chen, "Fisher indentify and mixing model based on multivariate statistical analysis of mine water inrush sources," *Journal of China Coal Society*, vol. 36, no. S1, pp. 131–136, 2011.

[20] J. Z. Qian, Y. Tong, L. Ma, W. D. Zhao, R. D. Zhang, and X. R. He, "Hydrochemical characteristics and groundwater source identification of a multiple aquifer system in a coal mine," *Mine Water and the Environment*, vol. 37, no. 3, pp. 528–540, 2017.

[21] X. Y. Wang, T. Xu, and D. Huang, "Application of distance discriminance in identifying water inrush resource in similar coalmine," *Journal of China Coal Society*, vol. 36, no. 8, pp. 1354–1358, 2011.

[22] Z. H. Jiang, Y. B. Hu, Q. D. Ju, L. Zhou, and S. Y. Zhang, "A discrimination method of mine water inrush source," *Industry and Mine Automation*, vol. 46, no. 4, pp. 28–33, 2020.

[23] L. S. Shao and X. C. Li, "Indentification of mine water inrush source based on MIV-PSO-SVM," *Coal Science and Technology*, vol. 46, no. 8, pp. 189–196, 2018.

[24] Y. Wang, M. R. Zhou, P. C. Yan et al., "A rapid identification model of mine water inrush based on extreme learning machine," *Journal of China Coal Society*, vol. 42, no. 9, pp. 2427–2432, 2017.

[25] P. H. Huang, X. Y. Wang, and C. Federico, "Piper-PCA-Fisher recognition model of water inrush source: a case study of the Jiaozuo mining area," *Geofluids*, vol. 2018, 10 pages, 2018.

[26] B. Li, Q. Wu, and Z. J. Liu, "Identification of mine water inrush source based on PCA-FDA: Xiandewang coal mine case," *Geofluids*, vol. 2020, 8 pages, 2020.

[27] I. M. Farnham, K. J. Stetzenbach, A. K. Singh, and K. H. Johannesson, "Deciphering groundwater flow systems in Oasis Valley, Nevada, using trace element chemistry, multivariate statistics, and geographical information system," *Mathematical Geology*, vol. 32, no. 8, pp. 943–968, 2000.

[28] L. Guo, Q. Jiao, D. Zheng, A. P. Liu, Q. Wang, and Y. G. Zheng, "Quality evaluation of Artemisiae Argyi Folium based on fingerprint analysis and quantitative analysis of multicomponents," *China journal of Chinese materia medica*, vol. 43, no. 5, pp. 977–984, 2018.

[29] P. M. Izquierdo Cañas, E. García Romero, S. Gómez Alonso, and M. L. L. Palop Herreros, "Changes in the aromatic composition of Tempranillo wines during spontaneous malolactic fermentation," *Journal of Food Composition and Analysis*, vol. 21, no. 8, pp. 724–730, 2007.

[30] M. L. Barker, "Partial least squares for discrimination in fMRI data," *Magnetic Resonance Imaging*, vol. 30, no. 3, pp. 446–452, 2012.

[31] B. Q. Liu, X. Q. Chen, X. G. Wu, W. Zheng, and Z. H. Wang, "Study of Pu'er raw materials grade classification by PCA and PLS-DA," *Journal of Tea Science*, vol. 35, no. 2, pp. 179–184, 2015.

[32] P. C. Yan, M. R. Zhou, Q. M. Liu, R. Wang, and J. Liu, "Research on the source identification of mine water inrush based on LIF technology and PLS-DA algorithm," *Spectroscopy and Spectral Analysis*, vol. 36, no. 9, pp. 2858–2862, 2016.

[33] H. Zhang, H. F. Xing, D. X. Yao, L. L. Liu, D. R. Xue, and F. Guo, "The multiple logistic regression recognition model for mine water inrush source based on cluster analysis," *Environmental Earth Sciences*, vol. 78, no. 20, pp. 1–15, 2019.

[34] X. C. Cao, J. Z. Qian, and X. P. Sun, "Hydrochemical classification and identification for groundwater system by using integral multivariate statistical models: a case study in Guqiao Mine," *Journal of China Coal Society*, vol. 35, no. S1, pp. 141–144, 2010.

[35] G. P. Panagopoulos, D. Angelopoulou, E. E. Tzirtzilakis, and P. Giannoulopoulos, "The contribution of cluster and discriminant analysis to the classification of complex aquifer systems," *Environmental monitoring and assessment*, vol. 188, no. 10, p. 591, 2016.

[36] P. Villegas, V. Paredes, T. Betancur, and L. Ribeiro, "Assessing the hydrochemistry of the Urabá Aquifer, Colombia by principal component analysis," *Journal of geochemical exploration*, vol. 134, pp. 120–129, 2013.

[37] J. Tang, X. Liao, H. Tong, and J. Gao, "GC-MS combined with PLS-DA to discriminate the varieties of XinJiang lavender essential oil," *Computers and Applied Chemistry*, vol. 31, no. 6, pp. 701–704, 2014.

[38] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[39] H. Zhang and D. X. Yao, "The Bayes recognition model for mine water inrush source based on multiple logistic regression analysis," *Mine Water and the Environment*, vol. 39, pp. 1–14, 2020.

[40] H. W. Wang, Z. B. Wu, and J. Meng, *Partial Least Squares Regression-Linear and Nonlinear Methods*, National Defense Industry Press, Beijing, China, 2006.

[41] C. F. So, K.-S. Choi, J. W. Y. Chung, and T. K. S. Wong, "An extension to the discriminant analysis of near-infrared spectra," *Medical Engineering & Physics*, vol. 35, no. 2, pp. 172–177, 2013.

[42] H. Zhang, D. X. Yao, H. F. Lu, N. N. Zhu, and L. Xue, "Application of principal component analysis and Bayes discrimination approach in water source identification," *Coal Geology & Exploration*, vol. 45, no. 5, pp. 87–93, 2017.

[43] Q. H. Deng, J. Y. Cao, L. P. Zhang, Y. X. Lin, and D. D. Zhang, "The Bayesian discrimination model for sources of mine water inrush based on principal components analysis," *Hydrogeology & Engineering Geology*, vol. 41, no. 6, pp. 20–25, 2014.

[44] J. Z. Qian, L. Wang, L. Ma, Y. H. Lu, W. D. Zhao, and Y. Zhang, "Multivariate statistical analysis of water chemistry in evaluating groundwater geochemical evolution and aquifer connectivity near a large coal mine, Anhui, China," *Environmental Earth Sciences*, vol. 75, no. 9, 2016.

[45] L. Eriksson, N. KettanehWold, J. Trygg, C. Wikström, and S. Wold, *Multi- and Megavariate Data Analysis. Part I: Basic Principles and Applications*, Umetrics AB, Sweden, 2006.

[46] F. Xiang, J. F. Ye, and Q. S. Hou, "Study on HPLC fingerprint of Hyperici Perforati Herba at different growth stage based on PCA and PLS-DA," *Chinese Journal of Pharmaceutical Analysis*, vol. 40, no. 3, pp. 568–576, 2020.

[47] B. Worley and R. Powers, "Multivariate analysis in metabolomics," *Current Metabolomics*, vol. 1, no. 1, pp. 92–107, 2013.

[48] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit et al., "Assessment of PLS DA cross validation," *Metabolomics*, vol. 4, no. 1, pp. 81–89, 2008.

[49] D. D. Cui, L. J. Zeng, J. L. Huang, X. Y. Feng, X. Y. Zhang, and K. Feng, "Study on quality grade evaluation of Andrographis Herba based on principal component clustering and PLS regression," *Chinese Traditional and Herbal Drugs*, vol. 50, no. 13, pp. 3200–3206, 2019.