Study on Multivariate Regression Analyzing and BP ANN Combination Method for Groundwater quality Forecasting

Ping Yang[1, a],Laijun Lu [1,b], Xinmin Wang[1,c]

[1]College of Earth Science, Jilin University, Changchun, 130000, China

[a]936342317@qq.com, [b]lulj1956@163.com, [c]wxm@jlu.edu.cn

Keywords: the quantification theory, multivariate linear regression, data dimensionality, BP artificial neural network, groundwater quality forecasting

Abstract. In this paper, the quantification theory model I based on the theory of multivariate linear regression was used as a preprocessing tool to reduce data dimensionality in 13 factors influenced groundwater quality. Then BP ANN groundwater quality forecasting model was created and 8 important characteristic factor is used as nodes of input layer . The simulation results covered most of the existing experimental data in LiGuanPu area of Shenyang city. The results showed that the method was more precise and accorded with actual instance.

Introduction

Groundwater quality are generally affected by human activities, water and rock interaction processes, soil, geological structure of aquifer and other natural factors. The research is related to geological characteristics, geochemistry, geophysics, remote sensing, hydrogeological geology and other massive geosciences data. Thus reducing data dimensionality, discovering the essence of the data constitute the main content of massive data analysis in hydrogeological geology.

Many models have been proposed about dimensionality reduction in the past few decades. For example, there are Factor Analysis model, Principal Component Analysis, Discriminant Analysis model and so on.

In this paper, the quantification theory model I based on the theory of multivariate linear regression was used to find the contribution of each factor. As a result, the important factors are selected and dimension of input variables is reduced through quantification theory I. Finally, an BP ANN model is created by learning and training with the mass sample data of monitoring wells. Overall, the combined model can be used to predict the pollutant concentration in the groundwater for a period of time in the future. The effectiveness of this method has been confirmed effective by the predicting application of groundwater pollution in LiGuanPu, where is a water source area in Shenyang.

The background of Study Area

The water source area - LiGuanPu is located in the south-west of Shenyang, along the north bank of Hunhe river. Geographical coordinates: longitude 123 ° 15 '00 " ~ 123 ° 30 ′ 00 ″, latitude 41 ° 42' 00" ~ 41 ° 47 ′ 30 ″. The study area covers an area of 26.5 square kilometers. There are 62 wells in the study area and conductivity of aquifer is better. Specific capacity can reach 30L/(s.m).

The basic theory, model and calculation

Dimensionality reduced based on the quantification theory model . In this paper, Cl^-, NO_3^-, SO_4^{2-}, K^+, Ca^{2+}, Mg^{2+}, pH values, time t, the geological characteristic of aquifer, the composition of aqueous media - all the influence factors - are used as the project and related categories in the quantification theory model I. The concentration of ammonia in the groundwater is used as reference variable. As shown in Table 1,13 items (8 quantitative items,5 qualitative items) and 21 categories are listed.

Table 1. Impact factors of Pollutant Concentrations in Groundwater - Item- Category

Factor	Item No	Item Title	Category No	Category Title/Unit
Quantitative	Item-1	Concentration of Cl^-	C1	[mg/L]
	Item-2	Concentration of NO_3^-	C2	mg/L
	Item-3	Concentration of SO_4^{2-}	C3	mg/L
	Item-4	Concentration of K^+	C4	mg/L
	Item-5	Concentration of Ca^{2+}	C5	mg/L
	Item-6	Concentration of Mg^{2+}	C6	mg/L
	Item-7	Concentration of HCO^-	C7	mg/L
	Item-8	Sequence Number of Time t (1998/01 as base datum line.value=1; 1998/02 value=2;and so on.)	C8	N/A
Qualitative	Item-9	the salinity of aquifer	C9	Sand Cobble gravel(SC)
			C10	Sand gravel(sg)
			C11	Pebbly grit(Pg)
	Item-10	the structure of aquifer	C12	Single Layer
			C13	Double Layer
			C14	Multilayer
	Item-11	Structural faults	C15	Yes
			C16	No
	Item-12	Structural fold	C17	Yes
			C18	No
	Item-13	pH value of groundwater	C19	Acidity
			C20	alkaline
			C21	Neutral

As is described in Table 2, there are 8 quantitative and 5 qualitative variables and 9 samples is observed.

Table2. The sample partly data of 9 drilling wells for Items-Categories

Wells No / Item-Category		2#	3#	15#	25#	30#	38#	39#	43#	44#
Cl^-		121.3	94.22	87.03	118.1	93.7	89.1	112.3	97.9	95.2
NO_3^-		5.23	10.2	5.01	7.20	2.03	7.14	3.16	3.60	7.0
SO_4^{2-}		94.8	60.9	68.6	98.3	74.7	65.4	93.9	91.6	114.8
Sequence No of Time		1	0	1	0	1	1	0	1	1
the salinity of aquifer	SC	0	0	1	1	0	1	0	1	0
	Sg	0	1	0	0	1	0	0	0	1
	pg	1	0	0	0	0	0	1	0	0
the structure of aquifer	S	0	0	0	0	0	0	1	0	0
	D	1	1	1	1	1	1	0	0	0
	M	0	0	0	0	0	0	0	1	1
…	…	...	...	...	...	...	...	...	...	...
Structural faults	Yes	0	0	1	0	0	1	0	0	0
	No	1	1	0	1	1	0	1	1	1
reference variable (ammonia)		0.22	0.64	0.02	0.25	0.09	0.02	0.02	0.04	0.02

For qualitative item, $d_i(j,k)$ is the reaction of Category k of Item j in Sample i.

$$d_i(j,k)=\begin{cases}1 & \text{There is response in the j}^{th}\text{the type of the k}^{th}\text{item in the i}^{th}\text{sample}\\ 0 & \text{No response in the j}^{th}\text{the type of the k}^{th}\text{item in the i}^{th}\text{sample}\end{cases}$$

For quantitative item, the sample data is $x_i(u)(u=1,2,...,8;\text{i}=1,2,...,9)$,so Reaction Matrix is:

$$X=\begin{bmatrix} x_1(1), & ..., & x_1(8), & d_1(1,1), & ..., & d_1(1,r_1), & ..., & d_1(5,r_5)\\ x_2(1), & ..., & x_2(8), & d_2(1,1), & ..., & d_2(1,r_1), & ..., & d_2(5,r_5)\\ ... & ... & ... & ... & ... & ... & ... & ...\\ x_9(1), & ..., & x_9(8), & d_n(1,1), & ..., & d_n(1,r_1), & ..., & d_n(5,r_5)\end{bmatrix}$$

In quantification theory I, The react data of reference variable, item and category can be expressed by the following linear regression model[1,3]:

$$\mathrm{y_i}=\sum_{u=1}^{h}b_u x_i(u)+\sum_{j=1}^{m}\sum_{k=1}^{r_j}d_i(j,k)b_j k+e_i, i=1,2,...,\mathrm{n} \qquad (1)$$

In Eq. 1, $b_j k$ is constant coefficients or regression coefficient which depends on category-k of item-j, e_i is random error in the i sample, y_i is actual observations in the i sample.

As similar as linear regression analysis, least squares estimation value b_u and b_{jk} calculated according to the principle of least square method. It is been proved that the least squares estimation value is minimum variance unbiased linear estimates. From the react matrix- X, it is been concluded:

$$X^T Xb=X^T Y$$

$$Y=[y_1,y_2,...y_n]^T$$

$$b=\left[b_1,b_2,...,b_h,b_{11},...,b_{1r_1},b_{21},...,b_{m1},...,b_{mr_m}\right]^T$$

The accuracy of the quantification theory is the same as the accuracy of multiple linear regression prediction model. It can be presented as complex correlation, which is a correlation coefficient related to predicted value and measured value.

$$r_{\hat{y}y}=\sqrt{a}=\frac{s_{\hat{y}}}{s_y} \qquad a=\frac{s^2_{\hat{y}y}}{s^2_y s^2_{\hat{y}}}$$

The partial correlation coefficient –k is used to measure the contribution of each item for reference variable[4].

$$k=\frac{-r_{iy}}{\sqrt{r_{ii}\times r_{yy}}}$$

After the parameter b_u and b_{jk} are calculated, the equation of predicting groundwater quality is presented:[5]

$$\begin{aligned}\hat{y}=&0.182x_1(u)+0.273x_2(u)-0.240x_3(u)+0.142x_4(u)+\\ &0.267x_5(u)+0.226x_6(u)+0.213x_7(u)+0.154x_8(u)+0.127d(1,1)+\\ &0.087d(1,2)+0.131d(1,3)-0.024d(2,1)+0.234d(2,2)-0.113d(2,3)+\\ &0.019d(3,1)+0.198d(3,2)+0.242d(4,1)+0.003d(4,2)+0.109d(5,1)+\\ &0.164d(5,2)+0.122d(5,3)\end{aligned} \qquad (2)$$

In Eq. 2,while regression coefficient- $b_j k$ is solved, partial correlation coefficient –k of each item is calculated ,as is shown in Table 3.

Table 3. Item-Parameter relationship table

Item No- Title	partial correlation coefficient -k	Variance ratio	range
1- the concentration of Cl^-	0.272	0.152	0.017
2- the concentration of NO_3^-	0.454	0.280	0.028
3- the concentration of SO_4^{2-}	0.151	0.142	0.011
4- the concentration of K^+	0.053	0.006	0.004
5- the concentration of Ca^{2+}	0.051	0.013	0.001
6- the concentration of Mg^{2+}	0.018	0.006	0.004
7- the concentration of HCO^-	0.250	0.106	0.013
8- Sequence Number of Time t	0.311	0.206	0.013
9- the structure of aquifer	0.141	0.137	0.008
10 - the salinity of aquifer	0.113	0.105	0.001
11 - Structural faults	0.051	0.004	0.004
12 - Structural fold	0.051	0.002	0.001
13 - pH value of groundwater	0.101	0.102	0.002

From the Table 3,it can be seen that the values of three parameters for items are basically consistent. According to the contribution on reference variables, All items are sorted as Concentration of NO_3^-, Time t, Concentration of Cl^-, Concentration of HCO^-, concentration of SO_4^{2-}, the structure of aquifer, the salinity of aquifer, pH value of groundwater…Structural fold . From the point of view of variance ratio, the cumulative sum of the first eight items account for 95.72% of the total. It is consistent with the current situation of research area where Permanganate, Ammonia Nitrogen exceeded seriously in recent years because of environment pollution. In addition, it can be concluded that the factor about space(such as the structure of aquifer, the salinity of aquifer) and the factor about time affect the transporting of groundwater contamination greatly[6-7].

It has been proved that 8-dimensional data instead of the original 13-dimensional data not only simplifies the input layer of subsequent BP artificial neural network model ,but also reflects the actual state of groundwater system structure.

The prediction of groundwater quality based on BP ANN model. In the field of system simulation and prediction about groundwater resources, BP neural network has been more applications. BP neural network is simple ,the speed of training is fast ,and any function can be fitted at arbitrary precision, especially which is suitable for solving problem of classification and prediction.

In this paper, a BP ANN model is created to predict the pollutant concentration. In the BP ANN structure, there are 8 nodes of input layer: Concentration of NO_3^-, Time t, Concentration of Cl^-, Concentration of HCO^-, concentration of SO_4^{2-}, the structure of aquifer, the salinity of aquifer, pH values, There are 6 nodes of hidden layer, because according to the experience, the number of hidden layer nodes is generally 75% of the number of nodes in the input layer[8]. There is one node of out layer: concentration of Ammonia. The forward calculation of the neural network is as follows:

$$o_j = \sum_{i=1}^{n} w_{ij} x_i - q_j, \quad n = 1,2,...,8; j = 1,2,...,6 \tag{3}$$

$$B_j = f_1(o_j), \ j = 1,2,...,l \tag{4}$$

$$L_u = \sum_{j=1}^{l} v_{ju} B_j - g_u, \quad u = 1 \tag{5}$$

$$C_u = f_2(L_u), \quad u = 1 \tag{6}$$

In Eq. 3 - Eq. 6, $x_i (i = 1,2,...8)$ is the node of input layer. $w_{ij} (i = 1,2,...8; j = 1,2,...6)$ is the connection weight of neuron of input layer and neuron of hidden layer. $q_j (j = 1,2,...6)$ is the threshold

of hidden layer. $f_1(x)$ is the motivation function of hidden layer. $B_j(j=1,2,...6)$ is the output of hidden layer. $v_{ju}(j=1,2,...6;u=1)$ is the connection weight of neuron of hidden layer and neuron of output layer. $g_u(u=1)$ is the threshold of output layer. $f_2(x)$ is the motivation function of out layer. $C_u(u=1)$ is the output of output layer.

$$E=\frac{1}{2}\sum_{k=1}^{T}\left(C^{(k)}-Y^{(k)}\right)^2 \tag{7}$$

In Eq. 7, C is the actual output of output layer. Y is the expected output of output layer. C is called Mean square error, which can guide us to correct the connection weight value[9]. The reverse learning process is as follows:

$$\begin{cases}\Delta v_{ju}=v_{ju}-a\dfrac{\partial}{\partial v_{ju}}E\\ v_{ju}^{'}=v_{ju}+\Delta v_{ju}\end{cases} \tag{8}$$

$$\begin{cases}\Delta w_{ij}=w_{ij}-a\dfrac{\partial}{\partial w_{ij}}E\\ w_{ij}^{'}=\Delta w_{ij}+w_{ij}\end{cases} \tag{9}$$

In Eq. 8- Eq. 9, E is mean square error, a is learning rate.

In this paper, 180 sample data of 9 wells in the dry season and the wet season of 10 years is selected from materials of borehole data in LiGuanPu – water sources– from 1998 to 2007 to be trained in the BP network.60 sample data from 2008 to 2010 is selected as verification sample.

Verify the result of prediction. In this paper, the forecasting program of BP neural network for single well is developed by C++ builder6 tools. The 8 feature factors affected the quality of groundwater pollution concentration of 60 sample in 2008,2009,2010 is inputted to BP Neural Networks. As is shown in Table 4.

Table 4. 60 sample data about 8 feature factors for 9 wells in 2008-1010

Year	Season	No	NO_3^-	Cl^-	HCO^-	SO_4^{2-}	t	aquifer structure	aquifer salinity	pH value
2008	Dry	2#	2.8	112.3	80.9	29.2	120	S	SC	AC
2008	Wet	2#	5.6	98.2	100.5	43.2	120	S	SC	N
2009	Dry	2#	3.6	87.7	78.2	28.2	132	D	SG	AC
2009	Wet	2#	7.6	108.3	120.9	34.2	132	D	PG	AC
2010	Dry	2#	2.5	93.3	76.9	26.2	144	D	SG	AC
2010	Wet	2#	8.3	98.4	88.2	45.5	144	M	SC	N
2008	Dry	13#	8.6	99.3	98.2	43.1	120	M	SC	AL
…	…	…	…	…	…	…	…	…	…	…
2010	Wet	44#	9.8	100.8	121	35.1	144	S	PG	AL

After calculated by above combined model, the values of groundwater pollutant concentration of 2008-2010 are obtained. As is shown is Table 5,Relative error of result is -2.5%,5.1%,-2.0%...4.5%,the average relative error is 3.75%. So it could be concluded that the accuracy of combined model for prediction is very high and could meet the actual requirement.

Table 5.The prediction error of BP Neural Network

Year	Season	No	Actual value （Ammonia）	Estimate value （Ammonia）	Absolute error	Relative error
2008	Dry	2#	0.0513	0.050	-0.0013	-2.5%
2008	Wet	2#	0.1713	0.1801	0.0088	5.1%
2009	Dry	2#	0.2645	0.2592	-0.0053	-2.0%
2009	Wet	2#	0.4922	0.5084	0.0162	3.2%
2010	Dry	2#	0.1412	0.1501	0.0089	6.3%
2008	Wet	13#	0.3313	0.3221	-0.0092	-2.7%
…	…	…	…	…	…	…
2010	Wet	44#	0.0247	0.0258	0.0011	4.5%

Conclusions

(1) By the quantification theory I, the feature factor is extracted while regression model about hydrogeological data is established and data dimensionality is reduced. Moreover, not only qualitative data can be processed, but also the hybrid data(qualitative and quantitative) can be processed. It is more consistent with the actual situation in geology.

(2)After reducing the data dimension, BP neural network model is used for groundwater quality prediction. The simulation results of model covers most of the existing measured data. The method demonstrated the effectiveness and superiority in the field of prediction of groundwater resources, and It provides a new direction about the research of groundwater contaminant migration and has some promotional value.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No.51278065).

References

[1] Na Li, Weibo Zhou: ANFIS Model and Its Application in Groundwater Prediction in Irrigation District Based On Principal Component Analysis. YELLOW RIVER,Vol.36(2014) ,p.80-83

[2] Tao Sun, Xianfeng Chu,Shibing Pan:Determination of Hydrogeological Point Parameters Based on Quantification Theory I.Journal of Earth Sciences and Environment,Vol.29(2007),p.285-288

[3]Sulin Xiang,Zhanmeng Liu,:Study on Multivariate Linear Regression Analyzing Model for Groundwater Discharge Forecasting,*JOURNAL OF CHINA HYDROLOGY,*Vol.26(2006), p.36-37

[4]Dongan Li, Junrui Ning:Reservoir prediction with integerated information based on artificial neural network technology and geostatistics,*OIL&GAS GEOLOGY,*Vol.31(2010) ,p.493-498

[5]Zhenyu Zhao,Shicheng Wang:Information processing of mineral resources prediction,*GOLD GEOLOGY,*Vol.9(2003), p.50-54

[6]Qian Wang,Guangjie Li:Application of quantification theory in forecasting debris flows,*The Chinese Journal Of Geological Hazard and Control,* Vol.18(2006), p.85-88

[7]Guangya Zhou,Wenquan Dong: On The Methematical Models Of Quantification Theories I&II,*Journal of JiLin University,*Vol.1(1979),p.11-18.

[8]Zhaotai Chen,Ruzeng Gao:Priliminary,application of quantification theory in South-Sichuan hydrocarbon decision,*OGP,*Vol.28(1991),p.57-66.

[9]Lin Lin,JinZhong,Bin Zhang:A Simplified Numerical Model Of 3-D Groundwater And Solute Transport At Large Scale Area, *Journal of Hydrology* ,Vol.22(2010),p.319-32