Performance of Classification Methods to Evaluate Groundwater (Case Study: Shoosh Aquifer)

Mohamad Sakizadeh

Assistant Professor, Faculty of Sciences, Shahid Rajaee Teacher Training University, Tehran, Iran

ABSTRACT The objective of this study was to classify the Shoosh Aquifer to several zones with different water quality in Khuzestan Province, Iran. In this regard, the performance of classification methods (Discriminant function and Cluster analysis) for the classification of groundwater based on the level of pollution with an emphasis on the problem of over-fitting in training data were considered. An over-fitted model will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data. Cluster Analysis(CA) was adopted to spatially explain the similarity of sampling stations with respect to measured parameters. Three methods for variable selection were used including regularized discriminant analysis, principal component analysis and Wilks's lambda method. The best algorithm for variable selection was Wilks'lambda which resulted in reducing the generalization error of the test sample to 0.1 for leave-one-out and 4-fold cross-validation. The second best performed algorithm was regularized discriminant function with 0.167 and 0.133 misclassification error for the two above-mentioned methods, respectively. Principal component analysis did not proved to be a promising algorithm for variable selection in the classification methods.

Key words: *Cluster analysis, Discriminant function, Groundwater quality, Over-fitting, Variable selection*

1 INTRODUCTION

Groundwater is generally a very good source of drinking water, because of the purification properties of soils; it is also used for irrigation and where surface water is scarce, for industrial purposes as well (Fried, 1975). Most of the area of Iran is located in arid and semi-arid zones so, it is the main source of water in these areas accordingly. Although an aquifer is more protected than surface waters, groundwater appears to be subject to pollution which isdefined as a modification of physical, chemical and biological properties of water,

restricting or preventing its use in the various applications where it normally plays a role (Fried,1975). In this regard, groundwater quality dataare characterized by high variability because of a variety of natural and anthropogenic influences. The best approach to avoid misinterpretation of environmental monitoring data is the application of multivariate statistical methods for environmental data classification and modeling (Reisenhofer *et al.* 1996). Discriminant analysis (DA) is one of these multivariatetechniques; however its application in environmental sciences is not as common as

Corresponding author: Assistant Professor, Faculty of Sciences, Shahid Rajaee Teacher Training University, Tehran, Iran,Tel: +98 212 297 0005, Email: msakizadeh@gmail.com

that of the other methods. It belongs to classification methods which are fundamental chemometric techniques designed to find mathematical models able to recognize the membership of each object to its proper class on the basis of a set of measurements (Sun, 2009). Among traditional classifiers, discriminant analysis is probably the most known method (McLachlan, 2004) and can be considered the first multivariate classification technique. Classification methods find mathematical relationships between a set of descriptive variables (e.g. groundwater quality variables) and a qualitative variable (i.e. the membership to a defined category). The basic mathematical framework of this method is out of the scope of this paper and has been given in other references (e.g. McLachlan, 2004). The goal of discriminant analysis application in environmental researches is different. Here, our purposes were to predict (with a reasonable misclassification error) the level of groundwater pollution based on measured groundwater quality variables and to identify those variables which have contributed to the separation of these stations.

In the earlier studies on the use of discriminant analysis for the prediction of pollution level (Feio *et al.,* 2009; Carroll *et al.,* 2009; Zhuang and Dai, 2007), re-substitution error (e.g. the difference between the response training data and the predictions the classifier makes of the response based on the input training data) has been utilized which does not guarantee good predictions for new data. Re-substitution error is often an overly optimistic estimate of the predictive error on new data. If the re-substitution error is high, you can't expect the predictions of the classifier to be good. In these situations, it is said that the classifier has over fitted the training dataset. In reality, one of the main problems of using discriminate analysis is to overlook the danger of over-fitting especially when working with a small dataset. Over-fitting occurs when a forecasting model has too few degrees of freedom. In other words, it has relatively few observations in relation to its parameters and therefore it is able to memorize individual points rather than learn the general patterns (Baum and Haussler, 1989).

To the best of our knowledge, in none of the earlier studies on the use of discriminant analysis in for treating environmental data, the problem of over-fitting has been considered extensively by researchers. Alberto *et al.* (2001) have pointed out the problem of over-fitting in their paper, however even though they have just used a holdout method to consider its effect on the generalization of the developed model. Thus, the main objectives of this study were (1) to classify the aquifer to several zones with different water quality (2) to extract the most important parameters in assessing variations in ground water quality of different zones (3) to reduce the risk of over-fitting (which is a common problem when working with small data set)during the assignment of each sample to the respected group for the test data set.

2 MATERIALS AND METHODS
2.1 Study area
Khuzestan province, located in southwest of Iran, has an area equal to 63213 square kilometers. Shoosh with a population of 202762 inhabitants is one of the northern cities of this province. The Shoosh aquifer is the primary source of groundwater, supplying nearly 100% of the total drinking water for people living in the region. According to the statistics gathered by Khuzestan Meteorological Department, average annual temperature and precipitation in the study area in 2012 have been 24.8°C and 102mm, respectively. The local economy depends largely upon farming. Tourism and manufacturing also contribute to the area's economy. Farms occupy over 70% of the study area, and the main agricultural crops are wheat,

corn and sugar-cane. Average groundwater-level fluctuations are very low; about 0.5-1 m between dry and wet seasons because of continues recharge with Dez and Karkhe rivers. The general direction of groundwater flow is southward (Babaei *et al.,* 2006).

2.2 Discriminant and cluster analysis

In this study, we used 18 groundwater quality variables (Table 1) which have been collected from 30 wells in Shoosh Plain (Figure 1), located in the northwest of Khuzestan Province in Iran, during an 8-year time period (from 2006 to 2013). The mean of parameters was used for statistical analysis. The flow diagram of this study has been illustrated in Figure 2. At the first step, Cluster Analysis (CA) (belongs to the class of data-analysis tools employed for unsupervised pattern recognition), was adopted to spatially explain the similarity of sampling stations with respect to measured parameters, attempting to minimize the sum of squares of any two clusters that could be formed at each step (Burden *et al.,* 2004).

Table1 Descriptive statistics of groundwater quality variables used in this study

Groundwater quality variables	Mean	Standard deviation	Min.	Max.
Electrical Conductivity (EC) ($\mu S\ cm^{-1}$)	895.05	2366.67	650.00	456.16
Total Dissolved Solids (TDS) (mgl^{-1})	297.97	1185.00	325.00	224.36
Turbidity (NTU)	5.16	15.00	0.25	3.21
pH	224.68	7.80	7.60	0.07
Total Hardness (TH) ($mg\ l^{-1}$)	264.74	1200.00	255.00	189.21
Ca ($mg\ l^{-1}$)	70.52	213.00	50.00	38.58
Mg ($mg\ l^{-1}$)	220.79	162.99	25.02	24.33
Sulfate ($mg\ l^{-1}$)	197.91	1125.00	250.00	197.18
Nitrate ($mg\ l^{-1}$)	3.84	14.31	0.00	3.17
Nitrite ($mg\ l^{-1}$)	0.34	0.46	0.02	0.14
Fluoride ($mg\ l^{-1}$)	25.30	0.83	0.41	0.11
Chloride ($mg\ l^{-1}$)	25.12	84.49	17.23	17.88
Phosphate ($mg\ l^{-1}$)	0.43	0.47	0.08	0.09
Cl residue	0.42	2.25	0.00	0.44
Fe ($mg\ l^{-1}$)	0.26	0.27	0.05	0.07
Mn ($mg\ l^{-1}$)	0.56	2.67	0.07	0.56
Cu ($mg\ l^{-1}$)	0.40	2.91	0.06	0.81
Cr (VI) ($mg\ l^{-1}$)	0.05	0.08	0.03	0.01
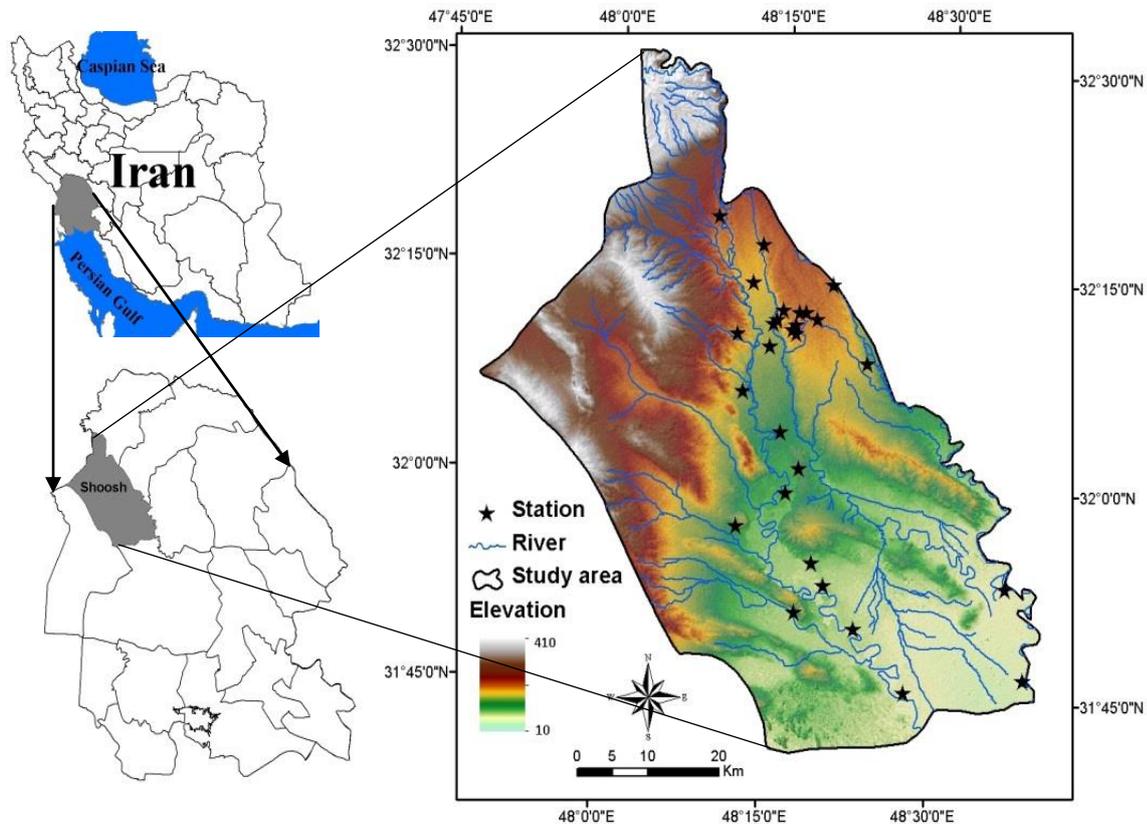
Figure 1 A view of sampling wells in Shoosh Aquifer, Iran

CA was applied on experimental data normalized to zero mean and unit variance (standardized data) in order to avoid misclassifications arising from the different orders of magnitude of both numerical value and variance of the parameters analyzed (Alberto *et al.,* 2001). Euclidean distance was used to compute the distance between objects in the data matrix. In order to verify the dissimilarity among objects (e.g. sampling stations), Cophenetic correlation coefficient was utilized to measure how well the cluster tree reflects the original data. The closer this coefficient is to one, the better is the cluster solution (Trauth, 2006). Two discriminant methods were used in this study. Fisher's linear discriminant analysis (Johnson and Wichern, 2007) and Canonical discriminant analysis. The earlier one was adopted to identify variables in terms of their discriminating power whereas the first one was used along with feature selection methods to consider the over-fitting of the developed function (s).
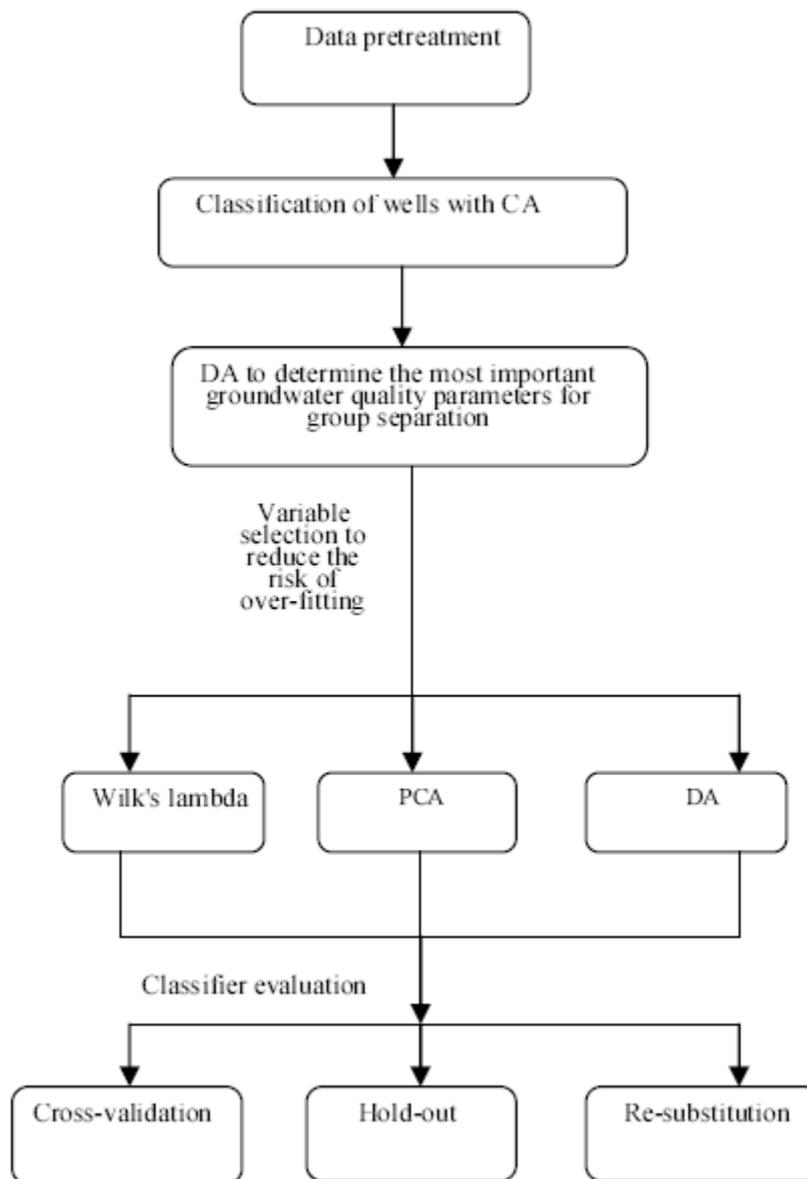
Figure 2 Flow diagram of the research for the classification of groundwater sampling stations in Shoosh Aquifer, Iran

2.3 Dimensionality reduction by feature selection

If in the formation of a discriminant rule, the number p of available feature variables is large relative to the total sample size, it will probably increase the risk of over-fitting which lead to poor out-of sample performance. In this study, we had 18 variables (e.g. groundwater quality parameters) and 30 observations (e.g. sampling wells) so, there was a high risk of over-fitting of our model to the training data therefore, the number of variables must have been reduced with a variable selection method to decrease the danger of over-fitting of the model to the training data set.

The aim of variable selection in this study was to identify those feature variables that are most useful in describing differences among the

possible groups and to reduce the risk of over-fitting as mentioned above. As a whole, three methods for variable selection were adopted including gregularized discriminant analysis (Friedman, 1989), principal component analysis (Jolliffe, 1972 and1973) and Wilks's lambda method (Mardia *et al.*1979; Ouardighi *et al.,* 2007). In order to examine the suitability of the data for principal component analysis Kaiser-Meyer-Olkin (KMO) was performed.

For variable selection with Wilk's lambda method, we used the probability of *F* as the criteria for enter of each variable into the discriminant model. A variable is entered into model if the significance level of its *F* value is less than the entry value and is removed if the significance level is greater than the removal value. The maximum significance of *F* to enter was 0.05 while the minimum significance of *F* to remove was 0.1 respectively.

2.4 Classifier evaluation

In normal practice, we only have a data set *S* with *n* samples available. The problem arises of how to divide the available cases into training set and test set. We have used three methods for this purpose: 1) re-substitution method: in which the whole set S is used for training, and for testing the classifier.2) hold-out method: in which the available n samples of S are randomly divided into two disjointed sets, S_d and S_t used for design and test, respectively.3) Partition or cross-validation method (Raudys and Jain, 1991).

The two cross-validation methods used in this study were leave-one-out and 4-fold cross-validation. For the leave-one-out method, 25% of the original dataset were randomly selected and used as the test sample, while the rest of the data were used as the training data for the model development. The misclassification error of the training set was regarded as re-substitution error while that of the test dataset was considered as the generalization error. All

of the required computations in this study were done with MATLAB (R2013b) and SPSS Statistics 17softwares.

3 RESULTS AND DISCUSSION

A first exploratory approach was the use of CA on the standardized data matrix, sorted by monitoring area (Figure 3). CA renders a dendrogram where 30 sampling wells were grouped into three statistically significant clusters. According to this figure, we can separate three groups base on the level of pollution. The cluster 1 (Shahrak Denial (1), Horrreyahi (3), Abozar Ghafari (4), Jarieh seyed mohamad (5), Koy Salman Farsi (10), Rahahan Station (12), Hamid Abad (13), Jarieh Seyed Razi (15), Tabatabayi Street (20), Sakhi (21), Behdasht Center (26)) correspond to relatively less polluted (LP) sites. Cluster (2) Shahid Beheshti (21), Radadeh (6), Tarvij (7), Tassisat Haj Abid (8), Aljazayer (9), Seyed Adnan (11), Tassisat Nader (14), Tassisat Mojahedin (16), Fahadbalkoh (17), Banader (23), Jarieh Seyed Mosa (24), Tassisat Seyed Abbas (27), Tassisat Zobeydat (29), Sorkhe Azadi (30)), correspond to moderately polluted (MP) sites and cluster 3 (Habaireh Sadat (18), Asabhad (19), Ankoosh (22), Tassisat Khalifeh Heydar (28)) corresponds to highly polluted sites (HP).The resultant Cophenetic correlation coefficient was 0.863 indicating that the clustering solution reflects the original data accurately. With respect to the results of Cophenetic correlation, average linkage was the best among the tested methods.
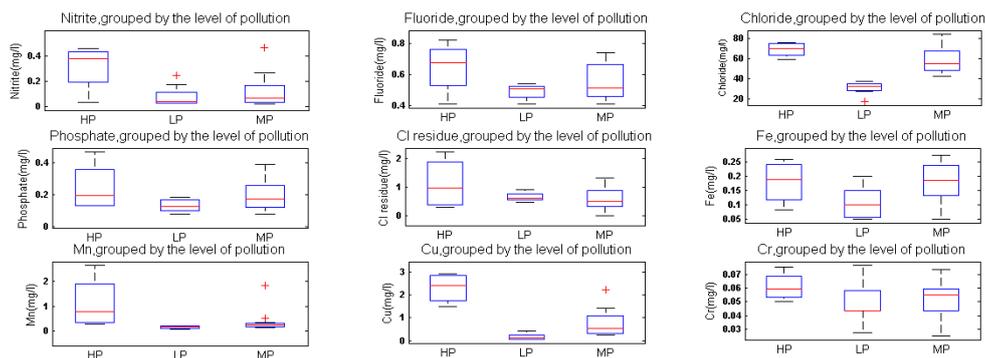
Box plots of discriminating parameters identified by spatial DA were constructed to evaluate different patterns associated with spatial variations in groundwater quality data (Figure 4a,b).
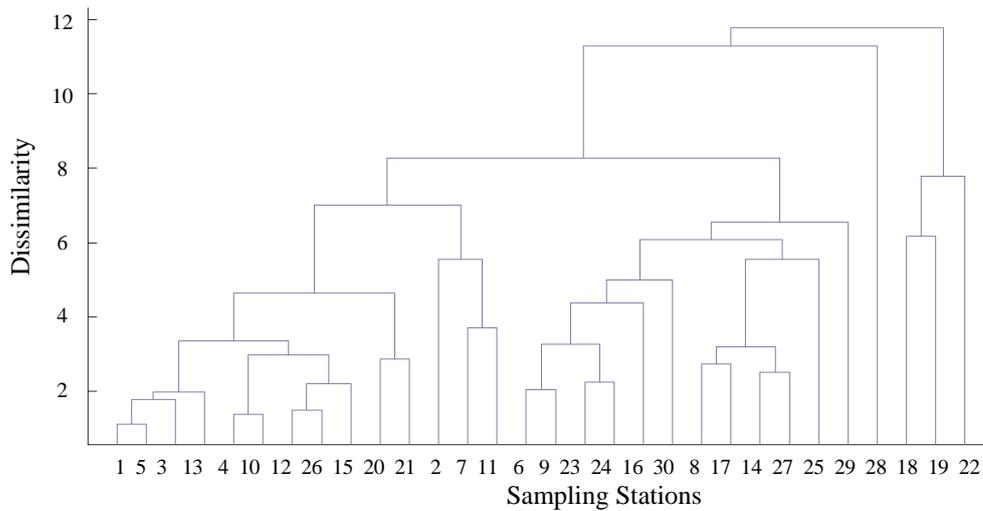
Figure 3 Dendrogram showing clustering of sampling sites
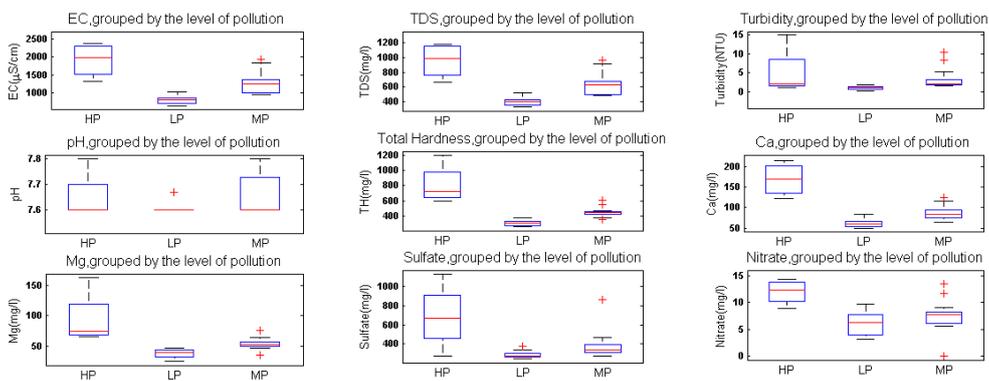
(a)

(b)

Figure 4 Box plot of groundwater quality variables of EC, TDS, Turbidity, pH, Totalhardness, Ca, Mg, Sulfate and Nitrate (a) and Nitrite, Floride, Chloride, Phosphate, Clresidue, Fe, Mn, Cu and Cr (b) base on the level of pollution determined by cluster analysis

603

Considering these figures, it is obvious that cluster analysis have been successful in the separation of available data into three groups with respect to the level of pollution. This is because, for most of the variables the median concentration corresponds to the level of pollution. The only exception is for Cl residue which contains higher median in LP in comparison with MP groups. Thus, we can assign each station to its respective group for discriminant analysis.

In discriminant analysis we are trying to predict a group membership, so firstly we examined whether there are any significant differences between groups on each of the independent variables using group means and ANOVA results data. The group statistics and tests of equality of group means tables provide this information. If there are no significant group differences, it is not worthwhile proceeding any further with the analysis. The results of equality of group means have been given in Table2.

Considering this table, the mean of all of the groundwater quality parameters other than pH, phosphate, Cl residue and Cr are different base

on the 5 percent significant level. In this field, Ca, Total Hardness, Chloride, Cu, TDS and EC produced very high value F's and low Wilks' lambda indicating that they might be potential variables with high discriminating power among groups. Wilks' lambda is the ratio of within-groups sums of squares to the total sums of squares. This is the proportion of the total variance in the discriminant scores notexplained by differences among groups. A lambda of 1.0 occurs when observed group means are equal (all the variance is explained by factors other than difference between those means), while a small lambda occurs when within-groups variability is small compared to the total variability. A small lambda indicates that group means appear to differ. Here, the value of Wilks' lambda distribution as shown in Table3 are 0.005 and 0.120 for the first and second discriminate function, respectively which are significant according to chi-square distribution. Therefore the discriminant model can be considered to be good enough for developing a discriminant function.

Table 2 Test of equality of group means

Groundwater quality Parameters	Wilks'lambda	F	Sig
EC	0.366	23.376	0.000
TDS	0.354	24.686	0.000
Turbidity	0.795	3.491	0.045
pH	0.834	2.682	0.087
Total Hardness	0.279	34.825	0.000
Ca	0.222	47.392	0.000
Mg	0.467	15.415	0.000
Sulfate	0.591	9.341	0.001
Nitrate	0.637	7.689	0.002
Nitrite	0.688	6.117	0.006
Floride	0.788	3.638	0.040
Chloride	0.301	31.406	0.000
Phosphate	0.818	3.010	0.066
Cl residue	0.834	2.678	0.087
Fe	0.791	3.561	0.042
Mn	0.683	6.273	0.006
Cu	0.301	31.368	0.000
Cr	0.916	1.235	0.307

Table 3 Wilks'lambda distribution

Test of Function(s)	1	2
Wilks'Lambda	0.005	0.120
Chi-square	99.435	39.219
df	36	17
sig	0.000	0.002

Table 4 Standardized canonical discriminant function coefficients

Groundwater quality parameters	Function	
	1	2
EC	-5.210	-0.615
TDS	4.894	0.157
Turbidity	0.360	1.126
pH	0.209	0.433
Total Hardness	5.201	-3.117
Ca	-0.593	0.132
Mg	-2.891	0.778
Sulfate	-0.341	1.567
Nitrate	-0.157	-0.370
Nitrite	0.241	0.487
Floride	-0.141	1.362
Chloride	-0.473	1.768
Phosphate	0.837	0.539
Cl residue	0.931	-1.779
Fe	0.125	0.530
Mn	0.405	-1.087
Cu	0.461	0.608
Cr	-0.208	-0.195

Standardizing the variables ensures that scale differences between the variables are eliminated. When all variables are standardized, absolute weights (i.e. ignore the sign) can be used to rank variables in terms of their discriminating power, the largest weight being associated with the most powerful discriminating variables. Variables with large weights are those which contribute mostly to differentiating the groups. Table4 shows the relative strength of the variables selected in the discriminant model on the basis of their discriminating power. Here, EC, Total hardness, TDS, and Mg had a high discriminating power for the first discriminant function whereas total harness had a high discriminating power for the second discriminant function. These variables with large standardized coefficients stand out as those that strongly predict pollution level of groundwater resources in the study area. To some extent, Cl residue, chloride, sulfate, turbidity, fluoride and Mn had the same role for the second discriminant functionas well.

Table 5 Misclassification error (MCE) of discriminant function using different methods for data partition and different algorithms for variable selection

Methods	Misclassification error			
	Resubstitution	Holdout	Leave-one-out cross-validation	4-fold cross-validation
Original data	0	0.22	0.20	0.30
Reduced data with RDA	0.045	0	0.167	0.133
Reduced data with PCA	0.045	0.303	0.267	0.333
Reduced data with Wilks'lambda	0.045	0	0.100	0.100

The results of Fisher's linear discriminant analysis on the original dataset have been given in Table5. According to this table, the re-substitution error of the original dataset was 0 while that of the holdout method was 0.22 indicating that 22 percent of the test sample has been misclassified. The amounts of leave-one-out and 4-fold cross-validation misclassification error for this method were 0.20 and 0.30, respectively. This shows that the developed discriminant function has over-fitted the training dataset despite the use of holdout and cross-validation methods because there is a high degree of freedom in our dataset (e.g. the number of feature is high in comparison with the number of observations). Thus, we have to use some methods to obviate the problem. One of these methods was regularized discriminant analysis which was applied on the developed discriminant function. As mentioned earlier (refer to classifier evaluation), one of the methods for classifier evaluation is to consider error rate with respect to the number of predictors. Figure 5shows these two parameters after the initial development of discriminant function. Considering this figure, there is a reasonable tradeoff between lower number of predictors (e.g. groundwater quality variables) and lower error and the minimum error rate obtained when the number of predictors reduced to nine predictors.

The values of Gamma and lambda that gave this minimal error were 0.333 and 1.366, respectively. Having assigned these values to our model, the final retained variables were Total harness, Ca, Nitrite, Fluoride, chloride, Phosphate, Cl residue, Mn and Cu. Therefore, sampling wells were re-clusters using theses reduced variables and the error values were worked out for each method accordingly. This time, the re-substitution error was augmented to 0.045 however even though the error for holdout, leave-one-out and 4-fold cross-validation were reduced to 0, 0.167 and 0.133, respectively. This indicates that with just 50% of the original dataset we have been able to significantly decrease the generalization error of our classifier.

The results of variable selection using PCA has been given in Table 6. In this field, the amount of yielded KMO value was 0.577 indicating the suitability of the data for conducting PCA. The size of the KMO value has no statistically critical point, but, according to empirical experience, the larger the KMO value, the more common factors suitable for PCA analysis there are. If the KMO value is greater than 0.8, this indicates that the data set is fit for PCA, but if the KMO value is smaller than 0.5, PCA is not suitable (Wu and Kuo, 2012).
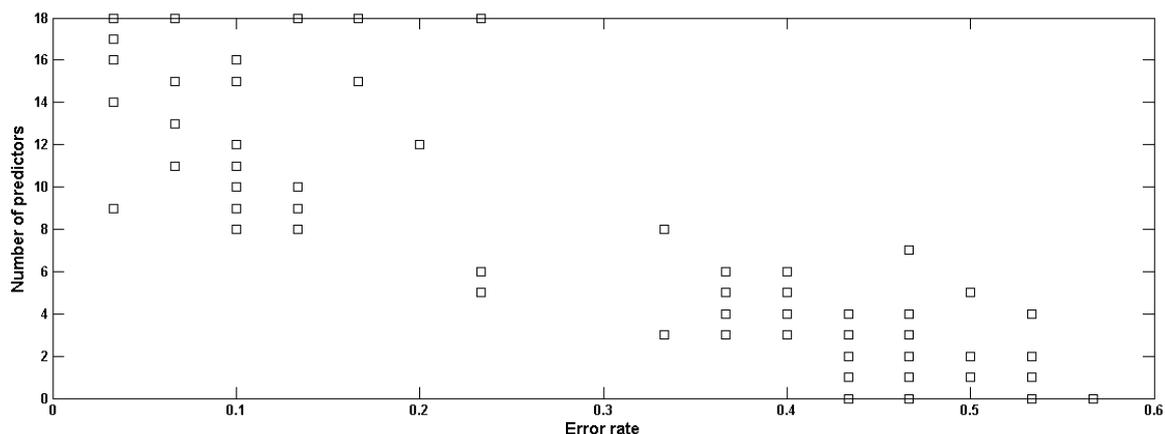
Figure 5 Error rate with respect to the number of predictors

Table 6 The results of PCA on the original groundwater dataset including the number of principal components, eigenvalues and the percent of variance explained by each PC

Number of Principal Components	Eigenvalue	Percent of variance explained
1	7.996	44.420
2	2.973	16.514
3	1.556	8.646
4	1.200	6.667
5	1.073	5.961
6	0.837	4.652
7	0.666	3.699
8	0.533	2.961
9	0.384	2.132
10	0.251	1.393
11	0.194	1.078
12	0.118	0.655
13	0.084	0.466
14	0.063	0.352
15	0.042	0.235
16	0.026	0.144
17	0.003	0.019
18	0.001	0.005

Moreover, considering the cut-off value of $\lambda=0.5$, this table shows that seven PCs can be retained accounting for 90.56 percent of the total variance. The associated variables for these PCs were Turbidity, Nitrite, Chloride, Phosphate, Fe, Cu and Cr. After re-clustering the sampling stations with respect to these parameters and development of linear discriminant function, the re-substitution error did not change but that of holdout, leave-one-

out cross validation and 4-fold cross validation increased to 0.303,0.267 and 0.333, respectively showing that there is a significant over-fitting in our training sample.

On the other hand, the results of variable selection using Wilks' lambda (Table7) indicate that at the first step the groups differ most on Ca (Λ drops to .222 if Ca is entered) and "Sig. of *F* to enter" is less than .05, so, that predictor is entered first. The following variables entered into the model were Chloride, Cl residue, Fe and Mn respectively. In the last step, Wilks' lambda reduced to the minimum value of 0.03 and no variable was entered into the model thereafter. The reduction of this statistic indicates an increase of the variability between groups, what leads to a better discrimination and thereafter a better classification. At this point, no variable already in meets the criterion for removal and no variable out meets the criterion for entry, so the analysis stops. The generalization error of this method after re-clustering the sampling stations (Table5) showed the best results among the others namely 0 for holdout and 0.100 for leave-one-out and 4-fold cross-validation. However, the re-substitution error has not changed in comparison with other variable selection methods.

In order to study the effect of variable selection on over-fitting, the method of Qiao *et al.* (2008) was followed. We did a Kruskal-Wallis one-way analysis of variance on variables and calculated the *p*-value for each one as a measure of how effective it is at separating groups. These *p*-values were ordered and plotted against misclassification error for training and test samples (Figure 5).

Considering this figure, it is clear that over-fitting happens when the number of significant

variables is more than four variables. In the other words, when the number of significant features is equal to 4, the misclassification error (MCE) for the test sample is the minimum, on the other hand, when the number of significant features is equal to 9, again the MCE reach its minimum. These two cut-off points are the number of variables retained by Wilks' lambda and regularized discriminant analysis, respectively. When the number of significant variables increased to 10, the MCE of the test set increased to 0.25 indicating that it has over-fitted the training data set. On the contrary, the re-substitution MCE stayed unchanged and even reached its minimum when the number of significant features was higher than 13 variables confirming the over-fitting in the training data set.

As a whole, as mentioned earlier, Wilks'lambds method outperformed that of the other methods for feature selection. Ouardighi *et al.* (2007) in their study on the comparison of five algorithms for feature selection (including Wilks'lambda) concluded that the feature selection by the Wilk's lambda statistic can lead, in general, to an improvement of the classifier performances. This filtering procedure is efficient but suffers from the significant number of features selected.

Principal component analysis did not prove to be a promising algorithm when applied with classification methods. This is because unsupervised learning methods such as PCA do not make use of the response variable, and hence may exclude components with little variance but a great deal of group information. That is to say, some small principal components that might be essential for classification are thrown away after PCA step (Zhuang and Dai, 2007; Tian *et al.*, 2010).

Table 7 Stepwise variable selection with Wilks'lambda method

Step	Variable entered	Wilks'lambda statistic	F	Sig
1	Ca	0.222	47.392	0.000
2	Chloride	0.104	27.362	0.000
3	Cl residue	0.058	26.416	0.000
4	Fe	0.040	24.057	0.000
5	Mn	0.030	22.107	0.000
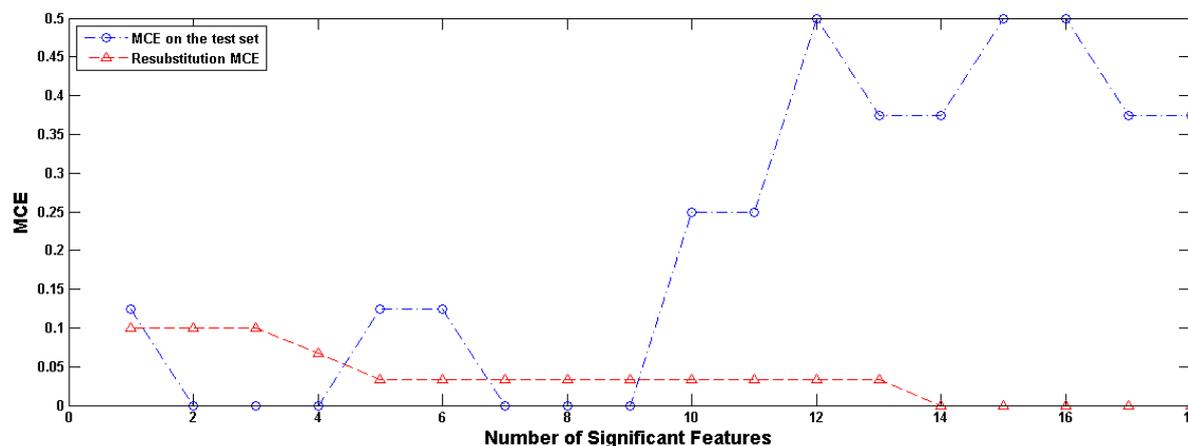
Figure 5 Misclassification error against number of significant features for training and test samples

In addition, when the number of variables exceeds the sample size, the within-class covariance matrix is singular. In these situations, regularized LDA (rLDA for short) can be used to circumvent the singularity problem as in ridge regression (Qiao *et al.,* 2009). This method outperformed that of PCA however its performance was less than that of Wilks'lambda. It might be due to the fact that the number of retained variables were higher (nine variables) than Wilks'lambda method which might have contributed to the over-fitting in the training dataset.

4   CONCLUSION

In order to study the over-fitting of classification methods, misclassification error of a developed discriminant function on groundwater quality dataset belonging to an eight years time period in Khuzestan province was considered using four algorithms and four methods. The results showed that when using discriminant analysis on the original dataset with high number of variables in comparison with the number of observations (18 variables and 30 observations in this case), the re-substitution error is low while its generalization error is high indicating that the original data have over-fitted the developed function. On the contrary, if we use variable selection methods along with discriminant analysis to select variables with the highest discriminating power, the generalization error decreases significantly. All in all, the best results were obtained with Wilks'lambda algorithm, while that of PCA method was not reasonable. That is due to the

fact that, PCA belongs to unsupervised learning methods which are not suitable for classification problems.

5   REFERENCES

Alberto, W.D., Pilar, D.M.D., Valeria, A.M., Fabiana, P.S., Cecilia, H.A. and Angeles, B.M.D.L. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. a case study: Suquia river basin (Cordoba-Argentina), Water Res., 2001; 35(12): 2881-2894.

Babaei, A.A., Mahvi, A.H., Nouri, J., Ahmadpour, E. and Mohsenzadeh, F. An experimental study of macro and micro elements in groundwater, Biotechnology, 2006; 5 (2): 125-129.

Baum, E.B. and Haussler, D. What size net gives valid generalization? Neurocomputing, 1989; 6:151-160.

Belhumeur, P.N., Hespanha, J.P. and Kriegman, D.J. Eigenfacesvs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 1997; 19: 711-720.

Burden, F.R., Donnert, D., Godish, T. and Mckelvie, I. Environmental monitoring handbook. 2004; McGraw-Hill Handbooks.

Carroll, S.P., Dawes, L., Hargreaves, M. and Goonetilleke, A. Faecal pollution source identification in an urbanizing catchment using antibiotic resistance profiling, discriminant analysis and partial least squares regression. Water Res., 2009;43: 1237-1246.

Feio, M., Almeida, S.F.P., Craveiro, S.C. and Calado, A.J.A comparison between biotic indices and predictive models in stream water quality assessment based on

benthic diatom communities, Ecol. Indic, 2009; 9: 497-507.

Fried, J.J. Groundwater Pollution. Developments in Water Science Series, 4 Elsevier, Amsterdam, 1975; 312 P.

Friedman, J.H. Regularized discriminant analysis. Jam. Statist. Assoc., 1989; 84: 165-175.

Jennrich R.J. Stepwise discriminant analysis. In: Statistical Methods for Digital Computers, John Wiley and Sons, 1977; NewYork.

Johnson, R.A. and Wichern, D.W. Applied multivariate statistical analysis, sixth edition, Pearson Prentice Hall, 2007; New Jersey.

Jolliffe, I.T. Discarding variables in principal component analysis. I: Artificial data. Appl. Statist., 1972; 21: 160-173.

Jolliffe, I.T. Discarding variables in principal component analysis. II: Real data. Appl. Statist., 1973; 22: 21-31.

Mardia KV, Kent, J.T. and Bibby, J.M. Multivariate Analysis. London, 1979; Academic Press.

McLachlan, G. Discriminant analysis and statistical pattern recognition, John Wiley and Sons, INC., Publication, 2004; New Jersy.

Ouardighi, A.E., Akadi, A.E. and Aboutajdine, D. Feature Selection on Supervised Classification Using Wilks Lambda Statistic, ISCIII'07.2007; International Symposium on Computational Intelligence and Intelligent Informatics.

Qiao, Z., Zhou, L. and Huang, J. Z. Effective linear discriminant analysis for high dimensional, low sample size data, Proceedings of 2008 World Congress of

Engineering (WCE 2008), 2008; 1070-1075.

Qiao, Z., Zhou, L. and Huang, J.Z. Sparse linear discriminant analysis with application to high dimensional low sample size data, Int. J. Appl. Math., 2009; 39: 48-60.

Raudys, S.J. and Jain, A.K. Small sample size effects in statistical pattern recognition: Recommendations for practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991; 13: 252-264.

Reisenhofer, E., Adami, G. and Favretto, E. Heavy metals andnutrients in coastal, surface seawaters (Gulf of Trieste, Northern Adriatic Sea): an environmental study by factor analysis. Fresenius J. Anal. Chem., 1996; 354: 729-734.

Sun, D.W. Infrared spectroscopy for food quality analysis and control, Academic press in an imprint of Elsevier, 2009; 51-82.

Tian, T.S., Wilcox, R.R. and James, G.M. Data reduction in classification: A simulated annealing based projection method. Statistical and Analytical Data Mining, 2010; 3 (5): 319-331.

Trauth, M.H. Matlab recipes for earth sciences, Springer, 2006; USA.

Wu, E.M.Y. and Kuo, S.L. Applying a multivariate statistical analysis model to evaluate the water quality of a watershed, Water Environ. Res., 2012; 84: 2075-2085.

Zhuang, X.S. and Dai, D.Q. Improved discriminate analysis for high-dimensional data andits application to face recognition, Pattern Recognition, 2007; 40: 1570-1578.

Zhou, F., Guo, H., Liu, Y. and Jiang, Y. Chemometrics data analysis of marine water quality and source identification in Southern Hong Kong, Mar. Pollut. Bull., 2007; 54: 745-756.

کارایی روش‌های طبقه‌بندی در ارزیابی آب‌های زیرزمینی
(مطالعه موردی: آبخوان شوش)

محمد ساکی‌زاده

استادیار، دانشکده علوم پایه، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران

چکیده هدف از انجام این مطالعه طبقه‌بندی آبخوان شوش به نواحی مختلف در استان خوزستان می‌باشد. بدین منظور روش‌های مختلف کلاسه بندی (تابع تشخیص و تحلیل خوشه‌ای) برای طبقه بندی آب زیرزمینی براساس میزان آلودگی، با تاکید برمشکل بیش برازش مدل به داده‌های تیمار مورد استفاده قرار گرفتند. مدلی که دارای مشگل بیش برازش باشد معمولا دارای قدرت پیش‌بینی پایینی می باشد. تحلیل خوشه‌ای به منظوربررسی تشابه میان ایستگاه‌های نمونه‌برداری براساس پارامترهای اندازه‌گیری شده مورد استفاده قرار گرفت. سه روش انتخاب متغیر شامل تابع تشخیص تنظیم شده، آنالیز مولفه‌های اصلی وروش ویلک لامبدا به کاربرده شدند. بهترین روش از بین اینها روش ویلک لامبدا بود که منجر به کاهش خطای داده‌های آزمون به ۰/۱ شد. دومین روش با بالاترین کارایی روش تابع تشخیص تنظیم شده بود که منجر به خطای ۰/۱۶۷ و۰/۱۳۷ برای داده‌های تیمار گردید. این در حالی است که آنالیز تجزیه به مولفه‌ها روش مناسبی برای انتخاب متغیرها تشخیص داده نشد.

کلمات کلیدی: انتخاب متغیر، آنالیز خوشه‌ای، بیش برازش، تابع تشخیص، کیفیت آب زیرزمینی