Environmental Research Journal
MEDWELL PUBLICATIONS

The Use of Environmetric Techniques Combined with Sensitivity Analysis for the Discrimination of Groundwater Quality Parameters

Usman Nasiru Usman and Hafizan Juahir
East Coast Environmental Research Institute, 21300 Kuala Terengganu, Malaysia

Key words: Sensitivity analysis, cluster analysis, discriminant analysis, groundwater quality, groundwater pollution

Corresponding Author:
Usman Nasiru Usman
East Coast Environmental Research Institute, 21300 Kuala Terengganu, Malaysia

Page No.: 73-78
Volume: 13, Issue 3, 2019
ISSN: 1994-5396
Environmental Research Journal

Abstract: Understanding the most effective pollutants affecting groundwater quality is of utmost importance in promoting sustainable development of groundwater resource. The study was performed to reduce the less significant parameter and give a preliminary judgment on the most significant water quality parameters discriminating the groundwater regions based on ANN model. This, study shows the use of sensitivity analysis combined with environmetric techniques such as Cluster Analysis (CA), Discriminant Analysis (DA). The water quality data was obtained from 10 different wells, over the period of 6 years (2006-2011) using 24 water quality parameters. Sensitivity analysis was carried out for nine models (ANN-R-AP, ANN-R-Na^+, ANN-R-Ca^+, ANN-R-HCO_3, ANN-R-Cl-, ANN-R-SiO_2, ANN-R-TDS, ANN-R-pH, ANN-R-EC). Percentage of contribution and R^2 was used for model performance evaluation criterion. The CA allowed the formation of two clusters between the sampling wells. The Low Contaminant Level as LCL and moderate contaminant level as MCL reflecting differences on water quality at different locations. DA as a data reduction techniques was used to evaluate the spatial variability in water quality as it uses 6 parameters (SO_4^-,Cl^-, As, Mn, NO_2 and total dissolved solid) affording 90.00% correct assignation to discriminate between the clusters using forward stepwise mode from the original 24 parameters. The sensitivity analysis reveals that Na^+, HCO_3, SiO_2 and EC are the four most effective parameters for discriminating groundwater quality regions with a percentage of contribution of 17.49, 17.50, 17.57 and 17.46%, respectively. This study reveals the significance of sensitivity analysis and multivariate techniques for the use of less parameter for understanding the most effective pollutant in water resource management, since, its time and cost consuming.

INTRODUCTION

The quality of groundwater is mainly influenced by both natural processes (lithology of the area, weathering and mineralization) and anthropogenic activities (municipal wastewater, industries and agriculture). Sources of groundwater contamination are widespread and include accidental spills, landfills, storage tanks, pipelines and agricultural activities, among many other sources (Bedient *et al*., 1994). Disposal of wastewater generated from municipal, industrial and agricultural sources with little or no treatment prior to discharge is a common practice in many developing countries including Malaysia (Juahir *et al*., 2008) and this can be harmful to living organism, not only human being but to the microorganism, wild life and plants.

Nevertheless, groundwater modeling has a great importance for society and particularly for public health aspect (Alagha *et al*., 2004). Knowledge of the most significant parameters contributing to the contamination of groundwater is of great importance, so as to control the activities related to the discharge of the pollutants. Therefore, protecting groundwater in the aspect of qualitative and quantitative aim is so important. Analytical techniques such as cluster analysis and discriminant analysis combined with sensitivity analysis were used to determine the most significant water quality parameters that best discriminate the two regions created by the cluster analysis and contributed to the water pollution. CA was employed to examine the spatial groupings of the sampling wells. It is a common method to classify variables into cluster (Massart and Kaufmann, 1983). The main objective of DA is to discriminant between two or more groups in term of the discriminating variables. Artificial Neural Network (ANN) sensitivity analysis with leave-one-out technique was employed, aimed to give relative significance of the input variable that contributed most in discriminating the regions. Sensitivity analysis is a tool for ranking the importance of model input-variable by assessing their contribution to the variability of the model output (Manache and Melching, 2008). During the last decade, process based groundwater modeling techniques were the default groundwater modeling tools (Javadi and Al-Najjar, 2007). These techniques have become very popular and effective for modeling complicated hydrological process using relatively less cost, effort and data (Iliadis and Maris, 2007; Chen *et al*., 2007; Dixon, 2005). Therefore, this study aim to investigate the most significant parameters discriminating the groundwater quality and provide the best input parameters that contribute most in discriminating the groundwater quality of the regions.

MATERIALS AND METHODS

Study area: Terengganu is situated in the North-Eastern Peninsular of Malaysia and it is bordered to the North-West by Kelantan and to the South-West by Pahang and to the East by South China Sea with a total area of land of 13035 km^2 and the maximum elevation of the state is 1507 m.

Terengganu has a population of 1,015.776 people as of 2006, Malay make up 94.7% of the population and Chinese 2.6% while Indians 0.2%. Other ethnic group raises the remainder 2.4%. The state population was only 48.7% urban; the majority lived in the rural areas of the state.

The study area has a strong tropical monsoon climate, relatively uniform temperature within 21 and 32°C range, January till April; the weather is dry and warm with humidity in the lowland consistently high between 82-86% annually. The annual average rainfall is 2,032-2540 mm with the most it, falling between Novembers till January.

Data collection: The water quality data in this study were obtained from ten monitoring wells by the department of mineral and geosciences, Terengganu. All the ten monitoring wells were observed and identified based on the availability of recorded data from the period of 2006-2011. The ten wells are: PT002, PT017, PT021, PT116, PT117, PT123, PT164, PT267, PT284 and PT300. Even though there are 50 water quality parameters but only 24 consistently sampled parameters were selected and a total of 60 samples and 1440 observation were used for the analysis. All the statistical analyses were performed using Microsoft excel 2007, JMP 2011 and XLSTAT 2014 Versions.

Cluster Analysis (CA): This is a group of environmetric techniques which primarily classify (Massart and Kaufmann, 1983) variables or cases (observation or samples) into cluster with high homogeneity level within the class and high heterogeneity level between classes to minimized their number and present it in a configuration of a tree-like structure with different branches (Dendogram) which provide visual summary of the clustering process. Branches that have linkage closer to each other indicate a stronger relationship.

In this present study, CA was applied for the grouping of ten monitoring wells using standard mode. The ward's linkage method (Ward, 1963) was used in the analysis. A classification scheme using Euclidean distance (straight line distance between two point in C-dimensional space define by C variable) for similarity measurement together with Ward method for linkage produces the most distinctive groups where each member within groups is more similar to its fellow member than to any member outside the group (Guler *et al*., 2002).

Discriminant Analysis (DA): The main objective of DA is to discriminant between two or more groups in term of the discriminating variables. It was applied to determine whether the group differ with regard to the mean of the

variables and use that variable to predict group membership. It was performed on the data set based on three different modes, i.e., standard mode, forward stepwise and backward stepwise modes to construct the best Discriminant Functions (DFs) to confirm the two clusters determined by means of CA and to evaluate spatial variation in portable water quality in Terengganu. The discriminant functions can be expressed as follows:

$$f(Gi) = ki + n\sum j = 1\ wij \times pij$$

Where:

I = The number of groups (G)
ki = The constant inherent to each group
n = The number of parameters used to classify a set of data into a given group
wij = The weight coefficient assigned by DF analysis (DFA) to a given parameter (pij)

In forward stepwise mode, variables are included step-by-step beginning with the more significant until no changes are obtained, whereas, in backward stepwise mode, variables are removed step-by-step beginning with less significant until significant changes are obtained. The membership of a well in a cluster 1 and 2 was the dependent variables whereas all the measured parameters constituted the independent variables.

Sensitivity analytical technique: The model developed in this study uses eight significant parameters obtained by the means of DA. It was performed in order to give a prefatory judgment on the importance of each water quality parameter on the groundwater regions using ANN which include the leave one-out method in order to understand which parameter contribute most in the two groundwater regions (Juahir *et al*., 2004). The study reveals the use of sensitivity analysis based on ANN to evaluate the significance of each parameter on the groundwater regions.

Sensitivity analysis was carried out for nine models. The first model was run using all parameters as input variable and named as Artificial Neural Network-Regions-All Parameters (ANN-R-AP) which served as a reference model. Two performance evaluation criterions were used to evaluate and compare model each other. These are correlation of coefficient (R^2) and percentage of contribution. Percentage of contribution of each input variable was obtained by using this formula:

$$\text{Contribution\%} = \frac{R^2_{ref} - R^2_{LP}}{\sum \Delta R^2}$$

where, R^2_{ref} is the correlation of coefficient (R^2) reference which was obtained by running all parameters as input variables and served as a reference model. R^2_{LP} is the R^2 leave-out parameters of each variable and $\Sigma\Delta R^2$ is the summation of change in R^2. Change of R^2 (ΔR^2)of each input variable was obtained by subtracting R^2 of leave-out parameter from the R^2 reference as shown in the equation below:

$$\Delta R = R^2_{ref} - R^2_{LP}$$

The second model was developed named as artificial neural network-regions-leave Na^+(ANN-R-Na^+) which means that Na^+ is excluded in forecasting the regions value. The third model is artificial neural network-regions-leave Ca^+ (ANN-R-Ca^+). The forth model is artificial neural network-regions-leave HCO_3(ANN-R-HCO_3). The fifth model is artificial neural network-regions-leave Cl^-(ANN-R-Cl^-). The sixth model is artificial neural network-regions-leave SiO_2 (ANN-R-SiO_2). The seventh model is artificial neural network-regions-leave TDS (ANN-R-TDS). The eighth model is artificial neural network-regions-leave-pH (ANN-R-pH) and the ninth model is artificial neural network-regions-leave EC (ANN-R-EC). A total of 540 observations from the year 2006-2011 were selected as data set and all models were run using JMP11 software.

RESULTS AND DISCUSSION

Cluster analysis: CA was carried out on the water quality data set to classify and evaluate the spatial variability among the monitoring wells. This analysis resulted in the grouping of the monitoring wells into two groups as shown in Fig. 1, Cluster 1 include four wells (PT002, PT017, PT021 and PT164) and are presented as Low Contaminant Level (LCL) while Cluster 2 contains six wells (PT116, PT117, PT123, PT267, PT284 and PT300) which represent the Moderate Contaminant Level (MCL). The reason behind this classification is that, Cluster 2 scored the highest mean of most of the pollutant

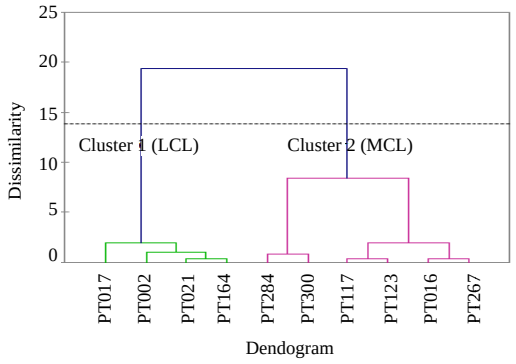

Fig. 1: A Dendogram showing the two regions of the cluster

Table 1: Mean by class of some of the major pollutants concentration

Variables/Class	LCL	MCL
Na^+ (mg L^{-1})	5.979	15.525
Ca^+ (mg L^{-1})	5.463	28.731
HCO_3 (mg L^{-1})	29.646	109.083
Cl^- (mg L^{-1})	5.688	17.167
SiO_2 (mg L^{-1})	8.965	17.297
TDS (mg L^{-1})	43.833	148.889
PH	6.492	7.464
EC (µS cm^{-1})	58.163	233.667

Table 2: Classification matrix for DA of spatial variation of the groundwater in Terengganu

	Region assigned by DA			
Sampling regions	Correct (%)	LCL	MCL	Total
Standard mode				
LCL	91.67	22	2	24
MCL	91.67	3	33	36
Total	91.67	25	35	60
Forward stepwise				
LCL	95.83	23	1	24
MCL	86.11	5	31	36
Total	90.00	28	32	60
Backward stepwise				
LCL	87.50	21	3	24
MCL	88.89	4	32	24
Total	83.33	25	35	60

concentration while Cluster 1 scored the least of the mean concentration. For instance, the mean concentration of MCL for EC is 233.667 µS cm^{-1} while for LCL is 58.163 µS cm^{-1}. Table 1 shows the details of the mean of some of the major pollutants concentration by class.

The outcome indicates that, onlyone well in each cluster is needed to represent a logical, accurate spatial distribution of the water quality for the whole network. The CA techniques shorten the need for numerous sampling stations, monitored from two monitoring wells that represent two different regions are sufficient. Figure 2 shows the two regions given by CA.

Discriminant analysis: In order to confirm the spatial variation of groundwater quality among different wells, DA was employed and it was performed using original data of 24 parameters after classification into two major clusters obtained from the CA. Groups (MCL and LCL) were run as dependent variables while water quality parameters were treated as independent variables. DA was carried out via standard mode, forward stepwise and backward stepwise modes. Wilk's lambda for each discriminant function of standard mode, forward stepwise and backward stepwise modes varied from 0.089, 0.307 and 0.139 at $p<0.0001$, respectively, suggesting that spatial DA was credible and effective.

Classification matrices and discriminant function obtained from standard mode, forward and backward stepwise modes are shown in Table 2 and 3, respectively. The accuracy of spatial classification using standard mode, forward stepwise, backward stepwise modes discriminant functions were 91.67, 90.00 and 83.33%, respectively (Table 2).

Table 3: Classification function coefficient resulting from discriminant analysis

Variables	Standard	Forward	Backward
Turbidity (NTU)	0.058		
color (HU)	0.467		
Na^+ (mg L^{-1})	0.007		0.058
K^+ (mg L^{-1})	0.582		
Ca^{2+} (mg L^{-1})	0.000		0.757
Mg^{2+} (mg L^{-1})	<0.0001		
Fe^{2+} (mg L^{-1})	0.407		
So_{4-} (mg L^{-1})	0.757	0.757	
CO_3 (mg L^{-1})	0.205		
F (mg L^{-1})	0.107		
P (mg L^{-1})	0.409		
HCO_3 (mg L^{-1})	<0.0001		0.181
Cl^- (mg L^{-1})	0.022	0.022	0.022
NO_3 (mg L^{-1})	0.294		
As (mg L^{-1})	0.181	0.181	
NH_4 (mg L^{-1})	0.960		
Mn (mg L^{-1})	0.025	0.025	
Zn (mg L^{-1})	0.058		
SiO_2 (mg L^{-1})	0.027		0.025
Total Solid (mg L^{-1})	0.513		
TDS (mg L^{-1})	<0.0001	< 0.0001	
NO_2 (mg L^{-1})	0.002	0.002	
PH()	<0.0001		0.002
EC (µS cm^{-1})	<0.0001		<0.0001

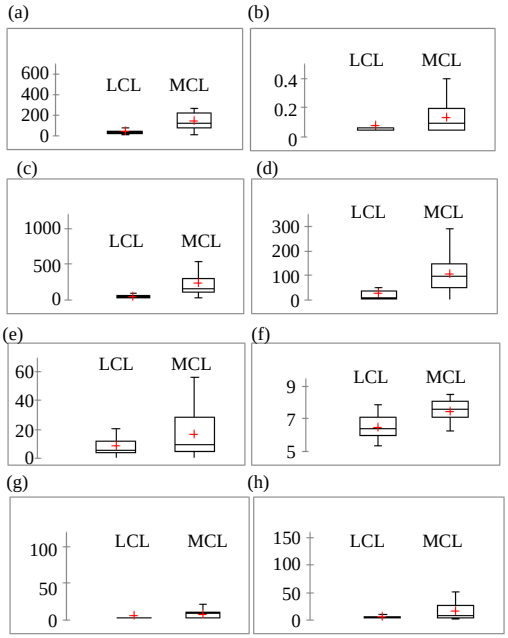

Fig. 2(a-h): Box and Whisker plot of discriminating parameters, (a) Box plots of dissolve solid, (b) Box plots of Mn, (c) Box plots of EC, (d) Box plots of HCO_3, (e) Box plots of SiO_2, (f) Box plots of PH, (g) Box plots of SO_4 and (h) Box plots of Cl

In forward stepwise mode, 6 parameters (Table 3) were found to be the most significant variable that best discriminate the clusters (SO_4, NO_{2-}, Cl^-, Mn, As and

Table 4: Result of sensitivity analysis for best input contributors in discriminating groundwater regions

Leave-out parameter (LP)	R^2_{LP}	R^2	Percentage of contribution	Highest contributors (%)
Na^+	0.6602	0.2392	17.4866	17.49
Ca^+	0.7487	0.1507	11.0168	
HCO_3	0.6601	0.2393	17.4939	17.50
Cl^-	0.8486	0.0508	3.7137	
SiO_2	0.6591	0.2403	17.5670	17.57
TDS	0.8438	0.0556	4.0646	
pH	0.7461	0.1533	11.2069	
EC	0.6607	0.2387	17.4501	17.46
Total		Σ1.3679		70.02

R^2_{ref} = 0.8994

Dissolve solid) which means that these parameters accounted for the most expected spatial variation in the groundwater quality. Backward stepwise mode on the other hand yielded seven parameters (Cl^-, EC, pH, Na^+, Ca^+, SiO_2 and HCO_3) to discriminate the two clusters (Table 3). The forward stepwise mode was proven to be a useful tool in recognizing the discriminant parameters in the spatial variation of potable water quality; this is because in forward stepwise mode, variables include step by step beginning with the more significant variables until no significant changes are obtained. The spatial DA suggest that SO_4, NO_{2-}, Cl^-, Mn, As and dissolve solid were the most significant parameters for discriminating among the cluster yielded by CA and accounted for most of the expected spatial variation in portable water quality. Thus, DA is a method that can determine the classification into predetermined group.

Box and Whisker plot of discriminating parameters identified by spatial DA (forward and backward modes) were constructed to evaluate different pattern associated with spatial variation in groundwater quality and presented in Fig. 2.

Determination of best input parameters in discriminating groundwater regions: Table 4 shows the overall result of nine ANN-R Models developed for sensitivity analysis. The Model ANN-R-AP was used as a reference to other models developed. ANN-R-AP Model show goodness of accuracy and present minimum residual error compared to other models with R^2 = 0.8994 and served as an R^2 reference and were used to predict the best input parameter that contribute most in discriminating the groundwater quality of the regions.

A slight reduction of R^2 value was noticed when excluding SiO_2 and HCO_3 parameters in predicting the best input contributors in discriminating groundwater quality regions. This shows that SiO_2 and HCO_3 are the highest contributors in discriminating groundwater regions. This also suggests that the model fitness was decrease and high residual error occur. Another lower value of R^2 was noticed for the model ANN-R-LNa^+ and ANN-R-LEC, 0.6602 and 0.6607, respectively Table 4. This indicates the significant of Na^+ and EC as important parameters in discriminating groundwater quality regions. These four models contributed up to 70% and served as the best models in discriminating the groundwater quality regions. However, models of ANN-R-LCl^-, ANN-R-LCa^+, ANN-R-LTDS and ANN-R-LpH demonstrate the less residual error and contributed less in discriminating groundwater quality regions.

CONCLUSION

The study has examined water quality of groundwater in Terengganu, Malaysia. The groundwater is classified as LCL and MCL by means of CA which indicates that water quality is varied smoothly and such spatial variation is likely due to natural hydrogeological environment and multipurpose nature of land use of the area. Thus, cluster analysis has confirmed the spatial variability of the groundwater. Nevertheless, DA gives a supportive result by providing the important parameters to discriminate the sampling wells affording correct assignation of 91.67, 90.00 and 83.33% for standard mode, forward stepwise, backward stepwise modes, respectively. The most significant variable that best discriminate the clusters (SO_4, NO_2^-, Cl^-, Mn, As and TDS) which means that these parameters accounted for the most expected spatial variation in the groundwater quality. Thus, discriminant analysis has determined the discriminant parameters associated with spatial pattern of groundwater. A sensitivity analysis helped to identify the effectiveness of the input parameters in discriminating groundwater quality regions. It has been found that Na^+, HCO_3, SiO_2 and EC are the four most effective parameters for discriminating groundwater quality regions with the total percentage of contribution up to 70% residual error. The study was performed to reduce the less significant parameter; therefore, the less important input parameters such as Cl^-, Ca^+ and TDS should be removed to simplify the model.

REFERENCES

Badient, P.B., H.S. Rifa'I and C.J. Newell, 1994. Groundwater Contaminant: Transport and Remediation. Prentice Hall, New Jersey, USA., Pages: 540.

Chen, K., J.J. Jiao, J. Huang and R. Huang, 2007. Multivariate statistical evaluation of trace elements in groundwater in a coastal area in Shenzhen, China. Environ. Pollut., 147: 771-780.

Dixon, B., 2005. Applicability of neuro-fuzzy techniques in predicting ground-water vulnerability: A GIS-based sensitivity analysis. J. Hydrol., 309: 17-38.

Guler, C., G.D. Thyne, J.E. McCray and A.K. Turner, 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J., 10: 455-474.

Iliadis, L.S. and F. Maris, 2007. An Artificial Neural Network model for mountainous water-resources management: The case of Cyprus mountainous watersheds. Environ. Model. Software, 22: 1066-1072.

Javadi, A.A. and M.M. Al-Najjar, 2007. Finite element modeling of contaminant transport in soils including the effect of chemical reactions. J. Hazard. Mat., 143: 690-701.

Juahir, H., S.M. Zain, M.E. Toriman, M. Mokhtar and H.C. Man, 2004. Application of artificial neural network models for predicting water quality index. J. Kejuruteraan Awam, 16: 42-55.

Juahir, H., T.M. Ekhwan, S.M. Zain, M. Mokhtar, J. Zaihan and M.J. Ijan Khushaida, 2008. The use of chemometrics analysis as a cost-effective tool in sustainable utilisation of water resources in the Langat River Catchment. Am. Eurasian J. Agric. Environ. Sci., 4: 258-265.

Manache, G. and C.S. Melching, 2008. Identification of reliable regression-and correlation-based sensitivity measures for importance ranking of water-quality model parameters. Environ. Modell. Software, 23: 549-562.

Massart, D.L. and L. Kaufman, 1983. The Interpretation of Analytical Chemical Data by the use of Cluster Analysis. Wiley, New York, Pages: 237.

Ward, Jr. J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc., 58: 236-244.