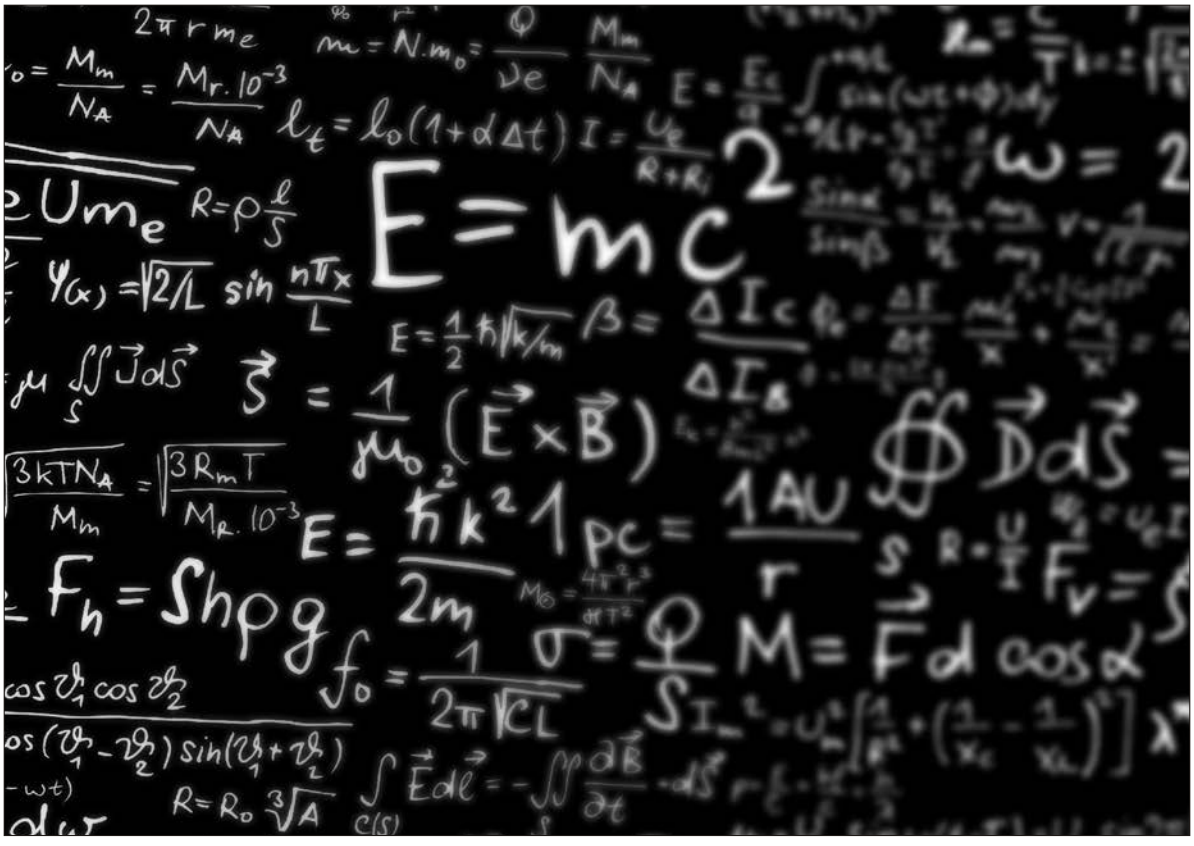

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 5 (4): 121-131, 2012
ISSN 1994-5418 / DOI: 10.3923/ajms.2012.121.131

Multiple Regression Models up to First-order Interaction on Hydrochemistry Properties

[1,2]Aminatul Hawa Yahaya, [1]Noraini Abdullah and [1]H.J. Zainodin

[1]Mathematics with Economics Programme, School of Science and Technology, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

[2]Malaysian Institute of Marine Engineering and Technology, Universiti Kuala Lumpur, 33200 Lumut, Perak, Malaysia

Corresponding Author: Aminatul Hawa Yahaya, School of Science and Technology, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

ABSTRACT

This study illustrated the procedure in selecting the best model in estimating the Electrical Conductivity (EC) levels based on the hydrochemistry properties and nature effecting factors using multiple regressions. The six independent variables and two dummy variables considered in this data set. The Multiple Regression (MR) models were involved up to first-order interaction and there were 57 possible models considered. This study is the extension of prior research which had generated 63 possible models, by using the same technique but no interaction involved between the independent variables. In this study, the process of getting the best model from the total of 120 possible models had been illustrated. The backward elimination of variables with the highest p-value was employed to get the selected model. The best model includes the combination of single and first order interaction (Li, Mg, Na-SO_4, Na-Li, Na-Mg and SO_4-Mg). The best model obtains then being verified by the Mean Absolute Percentage Error (MAPE) calculation to measure the models' relative overall fit.

Key words: Multiple regression, first-order interaction variables, dummy variables, backward elimination

INTRODUCTION

Groundwater is one of the major water resources especially in small tropical islands. Small tropical islands have limited sources of freshwater, no surface water and fully reliant on rainfall and groundwater recharge. The distortion caused by over exploitation of freshwater using pumping well have created an imbalance in the recharge-discharge equilibrium and resulting in the drawdown of the water table or upcoming of the saltwater intrusion (Kristie, 2007). This is often exacerbated by insufficient recharge to the freshwater aquifer which can occur in times of deficiency. The freshwater aquifers afloat on top of the saltwater at the interface due to density differences in the two respective water sources. The saltwater tends to form a lodge under the freshwater that extends inland. As saltwater intrusion occurs, this lodge extends further inland and is seen at shallower depths. The result is that wells that previously produced freshwater can see an increase in chloride concentration that makes the well unusable for irrigation or potable uses (Baharuddin *et al*., 2009).

The intrusion of saltwater has been the factor of saline water penetration due to the "landward and upward displacement of the freshwater-saltwater interface in coastal aquifers (Knighton *et al*., 1991) and as the invasion of fresh or brackish surface water or groundwater by water with higher salinity. Salinity can be explained by the total of all non-carbonate salts dissolved in water. Salinity is a capacity of the total salt concentration, comprised mostly of Na^+ and Cl^- ions. Even though, there are smaller quantities of other ions in seawater (e.g., K^+, Mg^{2+} or SO_4^{2-}), sodium and chloride ions represent about 91% of all seawater ions (Al-Naeem, 2008). Salinity is an important measurement in seawater where freshwater mixes with salty water (Abdullah *et al*., 2011). Chloride, an ion of the element chlorine, is naturally abundant within sea water. High chloride concentrations are often used as an indicator that seawater intrusion is occurring at a well but it is not a conclusive confirmation (Naeem *et al*., 2007).

In general, the Total Dissolved Solids (TDS) concentration is the amount of the cations (positively charged) and anions (negatively charged) ions in the water. Thus, salinity of the water can be determined by the TDS concentration (Mitra *et al*., 2007). Electrical Conductivity (EC) is a useful indicator of TDS because the conduction of current in an electrolyte solution is primarily dependent on the concentration of ionic species. EC is proportional to the sum of cations and anions and roughly equivalent to TDS in water. Solids can be found in nature in a dissolved form. Salts that dissolve in water break into positively and negatively charged ions. Conductivity is the capability of water to conduct an electrical current and the dissolved ions are the conductors (Alslaibi *et al*., 2011). The best method of monitoring mixture of fresh water and saline water can be done by measuring the electrical conductivity. Monitoring is conducted for separating stream hydrographs and geophysical mapping of contaminated groundwater. Examples of EC for distilled water should typically have an EC of less than 0.3 $\mu S\ cm^{-1}$ compared to groundwater, EC values greater than 500 $\mu S\ cm^{-1}$ indicate that the water may be polluted. The EC value of drinking water should be no more than 2500 $\mu S\ cm^{-1}$. Water with a higher TDS may have water quality problems and be unpleasant to drink (Thirumalini and Joseph, 2009). Factors of existing major chemical elements such as Na^+, Cl^-, K^+, Mg^{2+} and SO_4^{2-} contribute a significant role in the process of classifying and assessing groundwater quality. These ionic chemical elements have the ability to carry an electric current. The more dissolved ionic solutes in water, the greater it's EC, because the conduction of current in an electrolyte solution is primarily dependent on the concentration of ionic species. Warm weather in small tropical islands can increase the water salinity because of the evaporation process. This will leads to more widespread and severe problems in groundwater quality in small tropical islands.

MATERIALS AND METHODS

Study area: Currently, small tropical island which have been known as tourist attraction depends entirely on shallow aquifer groundwater supply mainly for domestic usage and washing purposes. Increased demands of the residents and tourists impose great pressure on the available groundwater resources. Dug wells are used to extract groundwater from the sandy aquifer. Groundwater is pumped routinely using water pump with water level meter integrated. Thus, aquifer in this tropical small island will becomes gradually vulnerable to seawater intrusion.

Data collection: A total of 59 groundwater samples were collected monthly from October 2008 to March 2009 in Manukan Island, a small tropical island in Sabah, West Malaysia located in South China Sea. The groundwater samples were collected from 10 boreholes which were drilled, using hand auger, to align perpendicularly from the sea. The depth of the boreholes ranged from

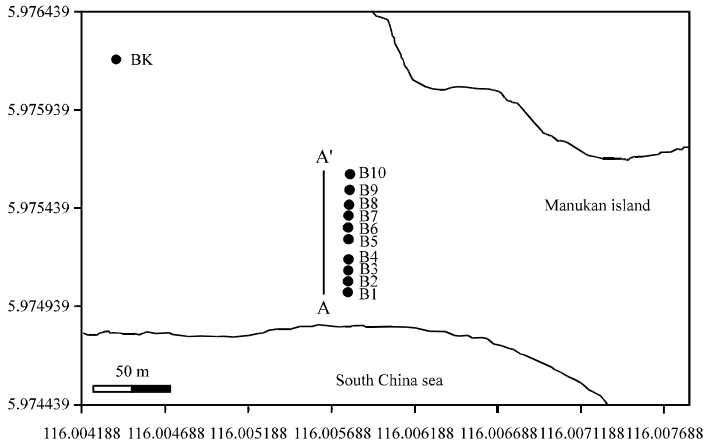

Fig. 1: Cross section (A-A') of boreholes installed

1.5-3 m. Cross section (A-A') of boreholes were installed perpendicularly from the coastline towards inland over a distance of 130 m and the proximity to the sea is in the order of B1 to B10, as depicted in Fig. 1.

The groundwater was extracted from boreholes using a portable vacuum pump interconnected with 0.3 inch polyethylene tubing. Groundwater was allowed to run for approximately 10 min in order to purge several boreholes volumes. The main reason was to remove stagnant water and allow representative groundwater to be sampled. Prior to each sample collection, the bottles were rinsed thoroughly with the groundwater from the boreholes. In-situ parameter such as EC was measured on site. Water samples to be sent for laboratory analysis were collected in Polyethylene (PE) bottles of one liter volume for anions and cations analysis. After filling the bottles with samples, the bottles were capped tightly, labeled and stored in a cooler box.

STATISTICAL ANALYSES

Preliminary study: The dataset used in this paper had undergone a preliminary study by using the Factor Analysis (FA). Factor analysis was performed on a subset of 19 selected variables (pH, EC, Ca, Mg, Na, K, HCO_3, Cl, SO_4, H_4SiO_4, Al, Ba, Be, Fe, Li, Mn, Pb, Se and Sr), that represented the overall groundwater chemistry framework. Five factors were extracted from the rotated component matrix. High positive loadings of Na, EC, SO_4, Li, K, Mg and Cl on Factor 1 indicated that the groundwater chemical composition was largely influenced by marine signature as these ions were found to be predominant in seawater (Voudouris *et al.*, 2000). For the Multiple Regression Analysis (MRA), only parameters from Factor 1 will be used. Two dummy variables (Tides and Borehole position) have been created to be included in the process of model building as shown in Table 1.

This variables set have been analyzed by using the MRA without any interaction involved. Only the single variables have been used and the result of the final model (M63.0.6) explained that only Sodium and Borehole Position give the significant effect in EC estimation. The details of this study are explained in Lin *et al.* (2012).

Multiple regression (MR) models with interaction: Multiple regression analysis, a form of general linear modelling (Hair *et al.*, 2010) is a statistical technique that can be used to analyze the relationship between a single dependent (criterion) variable and several independent (predictor)

Table 1: Description of variable involved in the models

Variable	Description	Unit
Y	Electrical Conductivity (EC)	($\mu S\ cm^{-1}$)
X_1	Sodium (Na^+)	($mg\ L^{-1}$)
X_2	Sulphate (SO_4^{2-})	($mg\ L^{-1}$)
X_3	Lithium (Li)	($mg\ L^{-1}$)
X_4	Potassium (K^+)	($mg\ L^{-1}$)
X_5	Chlorine (Cl^-)	($mg\ L^{-1}$)
X_6	Magnesium (Mg^{2+})	($mg\ L^{-1}$)
T	Tides (High = 1, Low = 0)	
P	Boreholes position (from sea) (Near = 1, Far = 0)	

variables. The objective of regression analysis is to predict a single Dependent Variable (DV) from the knowledge of one or more Independent Variables (IV)'s. Interaction effects represent the combined effects of variables on the criterion or dependent measure. When an interaction effect is present, the impact of one variable depends on the level of the other variable. Part of the power of MR is the ability to estimate and test interaction effects when the predictor variables are either categorical or continuous. As, Pedhazur and Schmelkin (1991) had noted, the idea that multiple effects should be studied in research rather than the isolated effects of single variables is one of the important contributions of Sir Ronald Fisher. When interaction effects are present, it means that interpretation of the individual variables may be incomplete or misleading. The specific MR model that has been explained by Lind *et al*. (2005) can be stated as follows:

$$Y_i = \beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+\varepsilon_i \quad (1)$$

where, X_i is the random variable representing the ith value of DV Y. Thus, X_{1i}, X_{2i}, ..., X_{ki} are the ith value of IV for i = 1, 2, ..., n.

Models results

All possible models: In the development of the MR models for this datasets, Electrical Conductivity (EC) would be the Dependent Variable (DV) noted by Y, whereas, Na (X_1), SO_4 (X_2), Li (X_3), K (X_4), Cl (X_5) and Mg (X_6) would be the Independent Variables (IV). Tides (T) and boreholes position (P) were included as independent dummy variables included in the models. Dummy variables were executed during the calculation of the possible models but included in the models before next model building procedure was carried out. All possible models, N can be calculated by using the formula:

$$N = \sum_{j=1}^{q} j\left(C_j^q\right) \quad (2)$$

where, N is the number of possible models generated and q is the number of variables and j = 1,2,..., q.

For this study, q = 6 (excluded the 2 dummy), the possible model is:

$$N = 1\left(C_1^6\right)+2\left(C_2^6\right)+3\left(C_3^6\right)+4\left(C_4^6\right)+5\left(C_5^6\right)+6\left(C_6^6\right)=192 \quad (3)$$

Table 2: Summary of all possible models

		Interaction					
No. of variable	Single	1st	2nd	3rd	4th	5th	Total
1	6						6
2	15	15					30
3	20	20	20				60
4	15	15	15	15			60
5	6	6	6	6	6		30
6	1	1	1	1	1	1	6
Total	63	57	42	22	7	1	192

The summary of all possible models are shown in Table 2. In this study, 192 models have been considered into further analysis because the interaction with dummy variable can only be done until the first order interaction (shaded area). The total numbers of models that have been considered in this analysis is 192 = single variable (63 models) + first order interaction variable (57 models).

Selected models: Multicollinearity is the intercorrelation of IV. The higher correlation coefficient will increase the standard error of the beta coefficients and produce assessment of the unique role of each independent resulting in difficult or impossible output. Multicollinearity exist if Correlation Coefficient >0.95. Zainodin-Noraini multicollinearity remedial procedures had been applied and details are explained in Abdullah *et al.* (2011) and Zainodin *et al.* (2011). Pearson correlation analysis verifies that there is existence of multicollinearity between IV's in M116 and seven variables (X_1P, X_2X_3, X_1T, X_2X_6, P, X_3X_5, X_3X_6) have been eliminated from the models (M116.7.0).

Next, the coefficient test should be carried out as an elimination procedure of insignificant variable by using the backward elimination as shown by Abdullah *et al.* (2008). To justify the removal of the insignificant variable, Wald test (Ramanathan, 2002) should be applied to the possible models upon the completion of all the elimination procedure of insignificant variables. In this step, the total of 14 insignificant variables have been eliminated from the model M116.7.0. At the end of this phase, only six variables have been left in the model (i.e., model M116.7.14). Table 3 shows the entered variable before the elimination procedure and the remaining variable after the elimination of insignificant variables.

Eight criteria of model selection (8SC): Identification of the best model should be based on Eight Selection Criteria (8SC) as shown in Abdullah *et al.* (2011). The objective is to determine a model with the lowest value of a criterion statistic. The calculation of the criterion statistics will be based on the Sum of Square Error (SSE), number of estimated parameters and the sample size. Table 4 shows the details of each model selection criteria.

Where, n would be the number of observations, (k+1) is the number of model's parameters and SSE the sum of square of error. The Akaike Information Criterion (AIC) (Akaike, 1974) and Finite Prediction Error (FPE) (Akaike, 1970) are developed by Akaike. The Generalised Cross Validation (GCV) is developed by Golub *et al.* (1979) while the HQ criterion is suggested by Hannan and Quinn (1979). The RICE criterion is discussed by Rice (1984) and the SCHWARZ criterion is discussed by Schwarz (1978). The SGMASQ is developed by Ramanathan (2002) and the Shibata criterion is suggested by Shibata (1981).

Table 3: Model M116.7.0 with entered variable before elimination procedure of insignificant variables and model M116.7.14 with remaining variable after elimination procedure of insignificant variables

	Unstandardized coefficients			
Model	β	Std. Error	t	p-value
M116.7.0				
(Constant)	0.397	0.064	6.186	4.39E-7
X_1	-0.041	0.203	-0.201	0.8419
X_2	0.066	0.151	0.436	0.6653
X_3	0.273	0.102	2.685	0.0110
X_5	0.020	0.115	0.176	0.8615
X_6	-0.491	0.163	-3.006	0.0049
X_1X_2	0.318	0.254	1.253	0.0014
X_1X_3	-0.715	0.206	-3.479	0.2939
X_1X_5	0.287	0.269	1.066	2.05E-11
X_1X_6	1.385	0.143	9.665	0.2453
X_2X_5	-0.255	0.215	-1.182	0.2974
X_5X_6	-0.233	0.220	-1.058	0.2773
T	-0.079	0.072	-1.104	0.7105
X_2T	-0.061	0.163	-0.374	0.8498
X_3T	-0.025	0.129	-0.191	0.6627
X_5T	0.041	0.093	0.440	0.1720
X_6T	0.114	0.082	1.394	0.9496
X_2P	-0.007	0.105	-0.064	0.8693
X_3P	0.021	0.126	0.166	0.6796
X_5P	0.039	0.095	0.416	0.6334
X_6P	-0.053	0.110	-0.481	0.2184
M116.7.14				
(Constant)	0.428	0.022	19.160	2.09E-24
X_3	0.273	0.058	4.749	1.82E-5
X_6	-0.634	0.044	-14.324	3.83E-19
X_1X_2	0.416	0.075	5.544	4.04E-10
X_1X_3	-0.747	0.096	-7.788	1.58E-22
X_1X_6	1.510	0.087	17.318	1.24E-5
X_2X_5	-0.228	0.047	-4.863	1.17E-6

Table 4: Eight selection criteria (8SC) for best model identification

AIC: $\left(\frac{SSE}{n}\right)e^{\frac{2(k+1)}{n}}$	RICE: $\left(\frac{SSE}{n}\right)\left(1-\frac{2(k+1)}{n}\right)^{-1}$
FPE: $\left(\frac{SSE}{n}\right)\frac{n+k+1}{n-(k+1)}$	SCHWARZ: $\left(\frac{SSE}{n}\right)(n)^{\frac{2(k+1)}{n}}$
GCV: $\left(\frac{SSE}{n}\right)\left(1-\frac{k+1}{n}\right)^{-2}$	SGMASQ: $\left(\frac{SSE}{n}\right)\left(1-\frac{k+1}{n}\right)^{-1}$
HQ: $\left(\frac{SSE}{n}\right)(\ln n)^{\frac{2(k+1)}{n}}$	SHIBATA: $\left(\frac{SSE}{n}\right)\frac{n+2(k+1)}{n}$

From 192 possible models generated during the stage of this analysis, only 67 models have been selected with the same SSE value and number of model parameter. These models then been grouped and any models from this group can be the selected model. The best model was then chosen from the selected models by using the 8SC based on the majority of least values as shown in Table 5. The best model selected is M116.7.14.

Table 5: Value of 8SC for all selected models

Selected model	Model parameter	SSE	AIC	RICE	FPE	SCHWARZ	GCV	SGMASQ	HQ	SHIBATA
M69.0.4	6	0.705	0.0156	0.0160	0.0156	0.0182	0.0158	0.0141	0.0170	0.0153
M70.0.4	6	0.828	0.0183	0.0188	0.0183	0.0214	0.0185	0.0166	0.0199	0.0180
M71.0.6	4	1.211	0.0249	0.0252	0.0250	0.0276	0.0251	0.0233	0.0264	0.0247
M72.2.5	3	0.983	0.0195	0.0197	0.0195	0.0211	0.0196	0.0185	0.0204	0.0194
M73.0.5	5	0.806	0.0172	0.0175	0.0172	0.0196	0.0174	0.0158	0.0185	0.0170
M74.0.4	6	0.843	0.0187	0.0192	0.0187	0.0217	0.0189	0.0169	0.0203	0.0183
M75.1.5	4	0.826	0.0170	0.0172	0.0170	0.0189	0.0171	0.0159	0.0180	0.0169
M76.0.7	3	0.946	0.0188	0.0189	0.0188	0.0203	0.0189	0.0178	0.0196	0.0187
M77.1.4	5	1.119	0.0239	0.0243	0.0239	0.0272	0.0241	0.0219	0.0256	0.0236
M78.1.3	6	1.390	0.0308	0.0316	0.0308	0.0359	0.0311	0.0278	0.0335	0.0301
M79.3.7	5	0.460	0.0098	0.0100	0.0098	0.0112	0.0099	0.0090	0.0105	0.0097
M81.0.8	7	0.416	0.0095	0.0099	0.0096	0.0114	0.0097	0.0085	0.0105	0.0093
M82.2.4	9	0.156	0.0038	0.0041	0.0039	0.0048	0.0040	0.0033	0.0044	0.0037
M84.3.6	6	0.428	0.0095	0.0097	0.0095	0.0110	0.0096	0.0086	0.0103	0.0093
M85.3.4	8	0.136	0.0032	0.0034	0.0032	0.0040	0.0033	0.0028	0.0036	0.0031
M86.0.9	6	0.435	0.0096	0.0099	0.0096	0.0112	0.0097	0.0087	0.0105	0.0094
M87.0.6	9	0.159	0.0039	0.0042	0.0039	0.0049	0.0040	0.0034	0.0044	0.0038
M88.1.4	10	0.148	0.0038	0.0041	0.0038	0.0049	0.0039	0.0032	0.0043	0.0036
M89.0.8	7	0.610	0.0140	0.0145	0.0140	0.0167	0.0142	0.0124	0.0154	0.0136
M91.2.8	5	0.730	0.0156	0.0159	0.0156	0.0177	0.0157	0.0143	0.0167	0.0154
M92.0.7	8	0.556	0.0132	0.0139	0.0132	0.0162	0.0135	0.0116	0.0148	0.0128
M93.2.6	7	0.783	0.0180	0.0186	0.0180	0.0215	0.0183	0.0160	0.0198	0.0175
M95.0.10	5	0.722	0.0154	0.0157	0.0154	0.0175	0.0155	0.0142	0.0165	0.0152
M96.1.8	6	0.725	0.0160	0.0165	0.0161	0.0187	0.0162	0.0145	0.0174	0.0157
M98.1.9	5	0.819	0.0175	0.0178	0.0175	0.0199	0.0176	0.0161	0.0188	0.0172
M100.4.10	7	0.417	0.0096	0.0099	0.0096	0.0114	0.0097	0.0085	0.0105	0.0093
M101.5.7	9	0.129	0.0032	0.0034	0.0032	0.0040	0.0033	0.0027	0.0036	0.0030
M102.3.12	6	0.437	0.0097	0.0099	0.0097	0.0113	0.0098	0.0087	0.0105	0.0095
M103.3.9	9	0.159	0.0039	0.0042	0.0039	0.0049	0.0040	0.0034	0.0044	0.0038
M104.1.11	9	0.137	0.0034	0.0036	0.0034	0.0042	0.0035	0.0029	0.0038	0.0032
M105.4.11	6	0.434	0.0096	0.0099	0.0096	0.0112	0.0097	0.0087	0.0104	0.0094
M106.5.9	7	0.129	0.0030	0.0031	0.0030	0.0035	0.0030	0.0026	0.0033	0.0029
M107.5.9	7	0.124	0.0028	0.0030	0.0028	0.0034	0.0029	0.0025	0.0031	0.0028
M108.1.8	12	0.132	0.0036	0.0041	0.0036	0.0049	0.0038	0.0030	0.0043	0.0034
M109.0.14	7	0.517	0.0119	0.0123	0.0119	0.0142	0.0121	0.0106	0.0131	0.0115
M110.3.11	7	0.632	0.0145	0.0150	0.0145	0.0173	0.0147	0.0129	0.0160	0.0141
M111.3.9	9	0.649	0.0160	0.0171	0.0160	0.0201	0.0165	0.0138	0.0181	0.0153
M112.3.11	7	0.589	0.0135	0.0140	0.0135	0.0162	0.0137	0.0120	0.0149	0.0131
M113.1.15	5	0.696	0.0149	0.0151	0.0149	0.0169	0.0150	0.0136	0.0159	0.0146
M114.7.16	5	0.457	0.0098	0.0099	0.0098	0.0111	0.0098	0.0090	0.0105	0.0096
M115.9.8	11	0.103	0.0027	0.0030	0.0027	0.0036	0.0028	0.0023	0.0032	0.0026
M116.7.14	7	0.109	0.0025	0.0026	0.0025	0.0030	0.0025	0.0022	0.0028	0.0024
M117.4.15	9	0.155	0.0038	0.0041	0.0038	0.0048	0.0039	0.0033	0.0043	0.0037
M118.5.13	10	0.102	0.0026	0.0028	0.0026	0.0034	0.0027	0.0022	0.0030	0.0025
M119.4.18	6	0.624	0.0138	0.0142	0.0138	0.0161	0.0140	0.0125	0.0150	0.0135

Best model verification: By using the Wald test, the complete model (M116) was taken as initial possible model and M116.7.14 as the reduced model. The complete (C) model (M116):

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_5X_5+\beta_6X_6+\beta_{12}X_1X_2+\beta_{13}X_1X_2+\beta_{15}X_1X_5+\beta_{16}X_1X_6+\beta_{23}X_2X_3+\beta_{25}X_2X_5+\beta_{26}X_2X_6+\beta_{35}X_3X_6+\beta_{56}X_5X_6+\beta_TT+\beta_PP+\beta_{1T}X_1T+\beta_{2T}X_2T+\beta_{3T}X_3T+\beta_{5T}X_5T+\beta_{6T}X_6T+\beta_{1P}X_1X_P+\beta_{2P}X_2P+\beta_{3P}X_3P+\beta_{5P}X_5P+\beta_{6P}X_6P+\varepsilon \quad (4)$$

$$SSE_C = 0.067 \text{ and } df_C = 28$$

The reduced (R) model (M116.7.14):

$$Y = \beta_0+\beta_3X_3+\beta_6X_6+\beta_{12}X_1X_2+\beta_{13}X_1X_3+\beta_{16}X_1X_6+\beta_{25}X_2X_5+\varepsilon \quad (5)$$

$$SSE_R = 0.109 \text{ and } df_R = 49$$

Hypothesis:

H_0: $\beta_1 = \beta_2 = \beta_5 = \beta_{15} = \beta_{23} = \beta_{26} = \beta_{35} = \beta_{36} = \beta_{56} = \beta_T = \beta_P = \beta_{1T} = \beta_{2T} = \beta_{5T} = \beta_{6T} = \beta_{1P} = \beta_{2P} = \beta_{5P} = \beta_{6P} = 0$
H_0: At least one β_S is non zero

Decision:

$$F_{Cal} = \frac{(SSE_C - SSE_R)/(df_C - df_R)}{SSE_C / df_C} = 0.8358$$

The value of $F_{critical}$ value from F distribution curve = $F_{table} = F_{0.05, 28, 49} = 1.80$ and the calculated value of $F = F_{cal} = 0.8358$. Since the calculated value of F is less than F_{table}, the decision is to accept H_0. The removal of insignificant variables in coefficient test is justified.

The final phase of model building is applying the Goodness-of-Fit on the final best model. The goodness-of-fit comprises of the randomness test and normality test. Randomness test is to determine that the residuals are randomly distributed and normality test on the Kolmogorov-Smirnov statistics is to ensure that the normality assumptions are not violated. The Runs Test value is 3.7767, since the value of $|Z| = 0.192$<asymp. Sig (2-tailed) = 0.847, therefore, H_0 is accepted and this test supported the conclusion that there is enough evidence that the residual is randomly distributed. Since the Kolmogorov-Smirnov statistics (0.192) gives the significant p-value = 0.200>0.05, therefore, H_0 is accepted. There is enough evidence at 0.05 significant levels that the standardized residual is normal. This statement is supported by the scatter plot and histogram in Fig. 2.

From here, the best regression model would therefore be represented by:

$$\hat{Y} = 0.428 + 0.273X_3 - 0.634X_6 + 0.416X_1X_2 - 0.747X_1X_3 + 1.51X_1X_6 - 0.228X_2X_5 \quad (6)$$

where, X_3 is Lithium, X_6 is Magnesium, X_1X_2 is interaction between Sodium-Sulphate, X_1X_3 is the interaction between Sodium-Lithium, X_1X_6 is the interaction between Sodium-Magnesium and X_2X_5 is the interaction between Sulphate-Chlorine. This interaction factor could be considered to reflect ion-exchange reactions between groundwater and the aquifer matrix corresponded to the positive

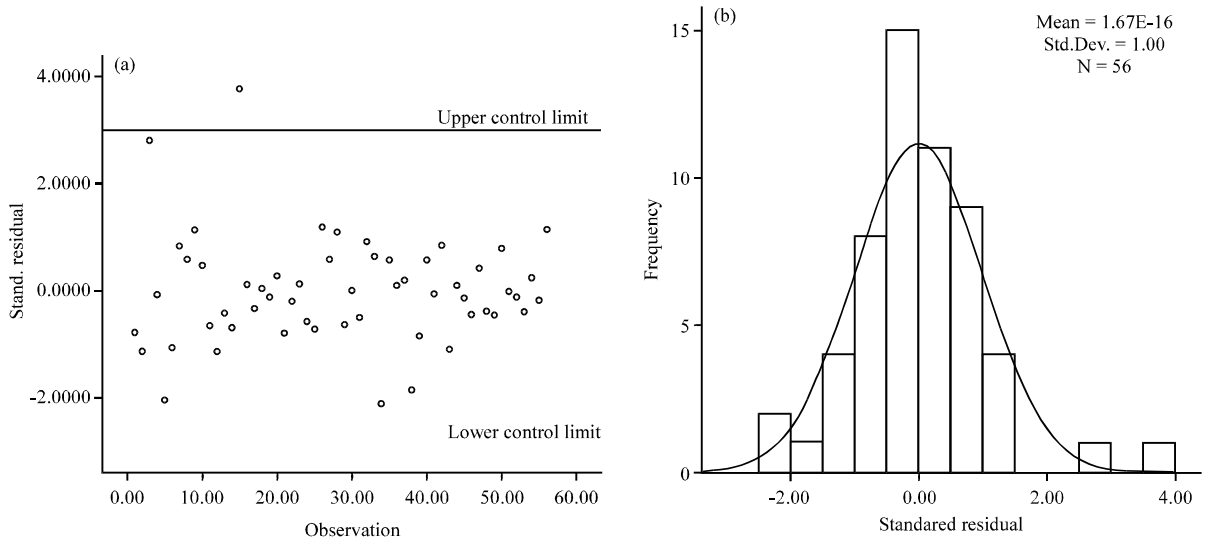

Fig. 2(a-b): (a) Standardized residual scatter plot and (b) Histogram with normal curve

loadings especially between Sodium and Sulphate. Sodium played a substantial role in controlling the behaviour of Sulphate and Magnesium in the shallow aquifer which will increase the salinity (EC level).

Model accuracy measurement: The Mean Absolute Percentage Error (MAPE) is commonly used in quantitative forecasting methods because it produces a measure of relative overall fit. The absolute values of all the percentage errors are summed up and the average is computed (Levy and Lemeshow, 1991). In this study, MAPE is used to verify the best model obtain. It usually expresses accuracy as a percentage and is defined by the formula:

$$MAPE = \frac{1}{a}\sum_{t=1}^{a}\left|\frac{A_t - F_t}{A_t}\right| \times 100 \quad (7)$$

where, A_t is the actual value and F_t is the forecast (estimated) value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value of this calculation is summed for every fitted or forecast point in time and divided again by the total number of fitted point's a. In this case, the number of a = 3, number of data reserved for this purpose. In general a MAPE of 10% is considered very well, a MAPE in the range 11-25% or even higher is quite common. The lower MAPE value the best the model can be used in forecasting or evaluating the missing values. By substituting the remaining observation that has no been included in the model building analysis, the value of MAPE obtained is 2.022%. This value indicates that this model could be best used for estimation of missing value or forecasting.

CONCLUSION

EC is widely used for monitoring the mixing of fresh and saline water. The groundwater with high EC level is not appropriate for drinking purposes attributed to its high salinity and elevated concentration of several minor elements. In this study, the model obtained clearly stated the

contribution of each parameter in determining the EC level. Na is dominant ions of seawater, high levels of Na ions in coastal groundwater may indicate a significant effect of seawater mixing. Eventually, Na is not independently significant in estimation of EC level. The interaction between Na-SO_4, Na-Li and Na-Mg has given a significant impact in the model. Two dummy variables (Tides and Borehole position) have been created to be included in the process of model building but have been eliminated during the modeling process. The dummy variables do not show any significant effect in estimation of EC level. The application of the model to such an island proved useful in demonstrating the mechanism of seawater intrusion in monitoring the water quality. The uses of variable interaction effects in the statistical model especially for environmental datasets have shown a significant impact parallel with the environmental theory. The interpretation on the environmental theory supported by the statistical modelling plays an important role in this task of problem solving and decision making. For further analysis, the remaining 72 models (Table 2) will be analysing using MRA with higher interaction. The highest interaction that can be considered for this dataset is until 5th order. With the higher interaction effects, the model is expected to give more significant.

ACKNOWLEDGMENT

The data for this study is financially supported by the Ministry of Science, Technology and Innovation, Malaysia (under Science Fund Grant No 04-01-10-SF0065. The authors thank the Mr. Lin Chin Yik and his project team members for providing the data for this research. The authors would also like to thank anonymous reviewers for their useful comments and enlightened ideas.

REFERENCES

Abdullah, N., H.J. Zainodin and A. Ahmed, 2011. Improved stem volume estimation using p-value approach in polynomial regression models. Res. J. Forest., 5: 50-65.

Abdullah, N., H.J. Zainodin and J.B.N. Jonney, 2008. Multiple regression models of the volumetric stem biomass. WSEAS Trans. Math., 7: 492-502.

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Stat. Math., 22: 203-217.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control, 19: 716-723.

Al-Naeem, A.A., 2008. Hydrochemical processes and metal composition of Ain Umm-Sabah natural spring in Al-Hassa Oasis Eastern province, Saudi Arabia. Pak. J. Biol. Sci., 11: 244-249.

Alslaibi, T.M., Y.K. Mogheir and S. Afifi, 2011. Assessment of groundwater quality due to municipal solid waste landfills leachate. J. Environ. Sci. Technol., 4: 419-436.

Baharuddin, M.F.T., R. Hashim and S. Taib, 2009. Electrical imaging resistivity study at the coastal area of Sungai Besar, Selangor, Malaysia. J. Applied Sci., 9: 2897-2906.

Golub, G.H., M. Heath and G. Wahba, 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21: 215-223.

Hair, J.F., W.C. Black and B.J. Babin, 2010. Multivariate Data Analysis: A Global Perspective. 7th Edn., Pearson Education Inc., New Jersey, USA., ISBN: 9780135153093, Pages: 800.

Hannan, E.J. and B.G. Quinn, 1979. The determination of the order of an autoregression. J. R. Stat. Soc. Ser. B, 41: 190-195.

Knighton, A.D., K. Mills and C.D. Woodroffe, 1991. Tidal-creek extension and saltwater intrusion in Northern Australia. Geology, 19: 831-834.

Kristie, W., 2007. Salinity Management Handbook. West Region Publ., South Queensland, Australia.

Levy, P. and S. Lemeshow, 1991. Sampling of Populations: Methods and Applications. John Wiley and Sons Inc., New York, USA.

Lin, C.Y., M.H. Abdullah, S.M. Praveena, H.Y. Aminatul and B. Musta, 2012. Delineation of temporal variability and governing factors influencing the spatial variability of shallow groundwater chemistry in a tropical sedimentary Island. J. Hydrol., 432-433: 26-42.

Lind, D.A., W.G. Marchal and R.D. Mason, 2005. Statistical Techniques in Business and Economics. 11th Edn., McGraw-Hill Inc., New York, USA.

Mitra, B.K., C. Sasaki, K. Enari, N. Matsuyama and S. Pongpattanasiri, 2007. Suitability assessment of shallow groundwater for irrigation in sand dune area of northwest Honshu Island, Japan. Int. J. Agric. Res., 2: 518-527.

Naeem, M., K. Khan, S. Rehman and J. Iqbal, 2007. Environmental assessment of ground water quality of Lahore Area, Punjab, Pakistan. J. Applied Sci., 7: 41-46.

Pedhazur, E.J. and L.P. Schmelkin, 1991. Measurement, Design and Analysis: An Integrated Approach. Routledge, Hillsdale, NJ., USA., ISBN-13: 9780805810639, Pages: 819.

Ramanathan, R., 2002. Introductory Econometrics with Applications. 5th Edn., South-Western/Thomson Learning, Ohio, USA., ISBN-13: 9780030341861, Pages: 688.

Rice, J., 1984. Bandwidth choice for nonparametric kernel regression. Ann. Statistics, 12: 1215-1230.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464.

Shibata, R., 1981. An optimal selection of regression variables. Biometrika, 68: 45-54.

Thirumalini, S. and K. Joseph, 2009. Correlation between electrical conductivity and total dissolved solids in natural waters. Malaysian J. Sci., 28: 55-61.

Voudouris, K., A. Panagopoulos and J. Koumantakis, 2000. Multivariate statistical analysis in the assessment of hydrochemistry of the Northern Korinthia prefecture alluvial aquifer system (Peloponnese, Greece). Nat. Resour. Res., 9: 135-143.

Zainodin, H.J., A. Noraini and S.J. Yap, 2011. An alternative multicollinearity approach in solving multiple regression problem. Trends Applied Sci. Res., 6: 1241-1255.