International Journal of Pure and Applied Mathematics
Volume 87 No. 3 2013, 443-451
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: http://dx.doi.org/10.12732/ijpam.v87i3.8

MULTIPLE LINEAR REGRESSION EQUATION FOR CHLORIDE ESTIMATION OF THE GROUNDWATER FOR CHANTHABURI, THAILAND

Jatupat Mekparyup[1], Kidakan Saithanu[2], Preeyarat Naksuwan[3], Mattika Hongboonme[4], Thaned Rojsiraphisal[5] §

[1,2,3,4]Department of Mathematics
Faculty of Science
Burapha University
Chonburi, 20131, THAILAND

[5]Department of Mathematics
Faculty of Science
Chiang Mai University
Chiang Mai, 50200, THAILAND

Abstract: The objective of this study is to estimate Chloride of the groundwater in Chanthaburi, Thailand using multiple regression analysis. The multiple linear regression equation proposed in this study related to Chloride, Total dissolved solids, Iron, Nitrate and Total hardness data which are obtained from the Department of Groundwater Resources, Thailand. Results shows that the best fitted equation to estimate Chloride (y) of the groundwater in Chanthaburi relates to Total dissolved solids (x_1) and Total hardness (x_4). It can be estimated by

$$\widehat{log(y)} = 1.22026 + 0.0050531x_1 - 0.0019421x_4$$

with adjusted coefficient of determination 0.607 and standard error of estimation 0.717.

AMS Subject Classification: 62J05
Key Words: multicollinearity, variance inflation factor, standard of ground-

Received: June 25, 2013

§Correspondence author

water quality, total dissolved solids, total hardness

1. Introduction

Due to lack of civilization in the rural area, water supply is costly and cumbersome process. Therefore, the groundwater becomes a principal source of water for drinking and other activities for many areas including the rural of Chanthaburi, Thailand. Prior to use, however, the groundwater must be determined to ensure that the quality of the groundwater is suitable for consumption [1].

Monitoring quality of groundwater is usually examined in four aspects which are (*i*) physical examination; i.e., check clearness and smell of the groundwater [2]; (*ii*) chemical examination; i.e., determine the amount of chemicals in the groundwater such as Iron [3], Manganese, Copper, Zinc, Sulphate, Chloride [4], [5], [6], [7], [8], [9], [10], Fluoride, Nitrate [3], [5], [7], [11], Total hardness [4], [7], [8], Total alkalinity [4] and Total dissolved solids [3], [4], [5], [6], [7]. (*iii*) Toxins examination; i.e., monitor any toxic contaminants in the groundwater such as Arsenic, Cyanide, Lead, Mercury, Cadmium and Selenium, etc. (*iv*) Bacteria examination; i.e., determine the bacteria and Escherichia coli (E. coli) in the groundwater [12].

Since some chemicals are easy to examine but some are not. Moreover, the monitoring of the groundwater quality requires expertise and tools for monitoring its quality [2]. For example, to examine the Chloride of the groundwater, it is tested by the Argentometric method which is only available in the chemical laboratory. In the past, there are several researches [13], [14], [15], [16] conducted on the groundwater quality for different places in Thailand but the estimation of Chloride of the groundwater in Thailand has not been conducted, while there are several studies that estimate Chloride of the groundwater for various area around the world [4], [5], [6], [7], [8], [9], [10]. This process is quite expensive and requires time to evaluate. To reduce cost of examination, therefore, the present study objectifies to estimate the Chloride of the groundwater through the multiple regression analysis which also saves time and cost for examining it.

2. Material and Methods

Data used in this study, provided by the Department of Groundwater Resources of Thailand, are Chloride, Total dissolved solids, Iron, Nitrate and Total hardness from 199 different fields collected from Chanthaburi province, Thailand.

To formulate the best fit for Chloride of the groundwater, the following processes are performed. First, simple correlation coefficients (R) are calculated to identify relationship among these data. The highly correlated variables are selected to form a multiple linear regression equation for estimating Chloride of the groundwater. Then, the multiple regression is validated with four assumptions. If any assumption fails, some modification is required with re-examine these assumptions. Lastly, we compare the observed values of Chloride of the groundwater with the estimated values obtained from the best fitted multiple regression equation.

2.1. Building Multiple Regression Equation

From total of 1,342 different fields, various chemical substances of the groundwater in Chanthaburi are collected. However, different chemical information are collected differently from fields to fields. There are only 199 different fields commonly collecting information of the Chloride (y), Total dissolved solids (x_1), Iron (x_2), Nitrate or NO_3 (x_3) and Total hardness (x_4). Thus, we use these five variables in multiple regression analysis. The multiple linear regression equation to estimate the Chloride of the groundwater is generated by regression model as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon. \tag{1}$$

The model in Eq. (1) consists of the response variable, Chloride (y); four predictor variables; Total dissolved solids (x_1), Iron (x_2), Nitrate or NO_3 (x_3) and Total hardness (x_4); $\beta_i (i = 0, 1, 2, 3, 4)$ are regression coefficients and ε is the error of regression model. Note that all substances are measured in mg/l.

2.2. Checking Assumptions

In order to use the proposed multiple regression analysis, it is necessary to verify that the proposed equation satisfies some assumptions which are commonly use in [17], [18], [19]. Four assumptions used in this study to validate the proposed multiple regression analysis are: (i) normality of the error distribution using Anderson-Darling statistic described in Eq.(2) [17]; (ii) independence of the errors using Durbin-Watson statistic described in Eq. (3) [18]; (iii) homoscedasticity (constant variance) of the errors using Breusch-Pagan statistic described in Eq. (4) [19]; (iv) multicollinearity among predictor variables using

Variance Inflation Factor (VIF) described in Eq. (5) .

$$AD = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln F(Y_i) + \ln(1 - F(Y_{n+1-i}))\right] \tag{2}$$

$$DW = \sum_{i=1}^{n} (\widehat{\varepsilon}_i - \widehat{\varepsilon}_{i-1})^2 \div \sum_{i=1}^{n} \widehat{\varepsilon}_i^{\,2} \tag{3}$$

$$BP = \frac{SSR^{\square}}{2} \div \left(\frac{SSE}{n}\right)^2 \tag{4}$$

$$\mathrm{VIF}_j = \frac{1}{1 - R^2_{j|\text{ others}}}, \tag{5}$$

where $SSR^{\square}$ is sum of squares in regression between the j^{th} residual (e_j^2) and x_{ij}; SSE is sum of squares in regression error between y_j and x_{ij}; $R^2_{j|\text{ others}}$ is multiple coefficient of determination between x_{ij} and all x_i.

2.3. Validate Estimated Values Against Observed Chloride

After the multiple linear regression equation is obtained. The values of observed Chloride (mg/l) of the groundwater from the 199 observed fields are compared with the estimated values obtained from the formulated multiple regression equation.

3. Results and Discussion

We first investigate the relationship between five variables; Chloride (y), Total dissolved solids (x_1), Iron (x_2), Nitrate or NO_3 (x_3) and Total hardness (x_4). The correlation coefficient values (R) for these variables are listed in Table 1.

Highly positive correlation between the response variable y and the predictor variables x_1 and x_4 are found to be significant with (P-*value* < 0.01). The two highest correlation coefficient value are $R = 0.684$ (between Chloride, y, and Total dissolved solids, x_1) and $R = 0.421$ (between Chloride, y, and Total hardness, x_4). These two highly correlated pair are found to be the same variables as in previous studies [4], [7]. Other correlation coefficient values range from -0.215 to 0.684.

Then, we find the best possible fitted multiple regression equation using the *best subsets method*. Results show that x_1 (total dissolved solids) and x_4

variable	y	x_1	x_2	x_3
y	1.000			
x_1	0.684*	1.000		
x_2	−0.002	−0.173**	1.000	
x_3	0.012	0.064	−0.078	1.000
x_4	0.421*	0.909*	−0.215*	0.079

*Significant at 0.01 level, **Significant at 0.05 level

Table 1: Correlation coefficients between the response variable and four predictor variables

(total hardness) are selected to form a multiple linear regression equation for estimating the Chloride of the groundwater with the *Mallow C-p* = 2.8 and $S = 30.911$. The multiple linear regression is formed as:

$$\widehat{y} = -16.842 + 0.43319x_1 - 0.45537x_4 \tag{6}$$

Further analysis on the Eq. (6) is proceeded by ANOVA (as shown in Table 2). The test statistic F is 226.56 and found to be significant with *P-value* < 0.01. Then, the regression coefficients of the equation are tested consequently (as shown in Table 3). Results show that all regression coefficients are significant (*P-value* < 0.01). Note that the best fitted multiple linear regression equation in Eq. (6) is generated with adjusted coefficient of determination (R^2_{adj}) 0.695 and the standard error of the estimation (S) 30.9109.

Source	*df*	*SS*	*MS*	*F*	*P-value*
Regression	2	432941	216471	226.56	0.000
Residual Error	196	187275	955		
Total	198	620216			

SS is sum of squares; MS is mean of squares

Table 2: ANOVA for regression

Then, the multiple regression equation as in Eq. (6) is validated with four assumptions, as in Section 2.2. Result shows that the normality test of the error distribution using Anderson-Darling statistic as in Eq. (2) is 9.844. But the result is not significant comparing with the critical value 0.752. This means that the proposed multiple regression equation as in Eq. (6) fails the assumption test. Therefore, the box-cox transformation [20] is applied to transform the

Table 3: Coefficient of regression

Predictor	Coefficient	SE coefficient	T	P-*value*
Constant	-16.842	3.650	-4.61	0.000
x_1	0.43316	0.02356	18.39	0.000
x_4	-0.45537	0.03722	-12.23	0.000

Table 4: VIF values for multicollinearity testing

Predictor	x_1	x_4
VIF	5.7	5.7

Chloride data (y) and the modified equation is formulated as:

$$\widehat{\log(y)} = 1.22026 + 0.0050531x_1 - 0.0019421x_4. \tag{7}$$

The multiple regression equation as in Eq. (7) is also tested with four assumptions as in Section 2.2. Results show as follow: (*i*) the normality test is determined and is found to be significant, $AD = 0.363$ with the critical values 0.752. (*ii*) the test of independence: Durbin-Watson statistic is calculated by Eq. (3). Result shows that the test statistic value, $DW = 1.88991$ is more than critical values $DL = 1.70$. So the errors were independent. (*iii*) the test of homoscedasticity: Breush-Pagan statistic is determined by Eq. (4). Result shows that the test statistic value $BP = 7.3887$ is less than the critical values 7.8147. So the variance of error is constant. (*iv*) Test of multicollinearity: the VIF values are calculated by Eq. (5) and results are listed in Table 4. The VIF values are more than 5; i.e. there is relationship among predictor variables (x_1 and x_4) in multiple regression equation [21].

Note that the multicollinearity is found in this multiple regression analysis. This, in general, might occur if these variables actually related in nature. However, the problem can be resolved statistically by collecting more observed data into the equation.

Lastly, we show the accuracy of estimation of Chloride of the groundwater for Chanthaburi. The observed Chloride of the groundwater from 199 different fields and the predicted values from Eq. (7) are plotted in Figure 1.

4. Conclusion

In this study, we apply the multiple linear regression analysis to estimate Chloride of the groundwater for Chanthaburi Thailand. The analysis show that the

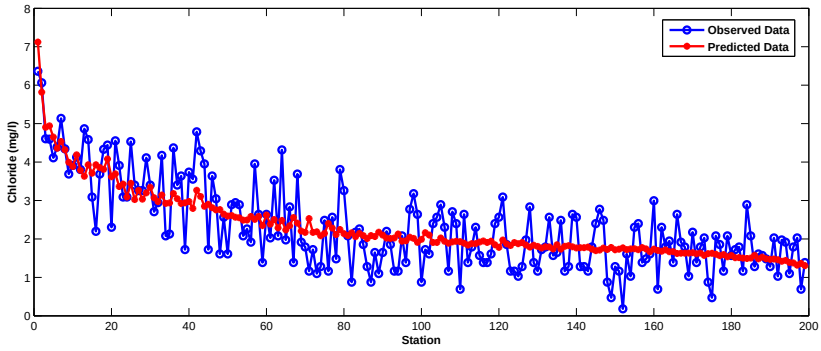

Figure 1: Comparison observed and predicted Chloride in the groundwater.

predictors used to estimate Chloride are Total dissolved solids (x_1) and Total hardness (x_4) with adjusted coefficient of determination (R^2_{adj}) of 0.607 and the standard error of the estimation (S) of 0.717.

Acknowledgments

TR thanks to Chiang Mai University for fully support of this work. We also thanks to the Department of Groundwater Resources, Thailand for kindly providing all data.

References

[1] K. Jothivenkatachalam, A. Nithya, S. C. Mohan, Correlation analysis of drinking water quality in and around Perur block of Coimbatore District, Tamil Nadu, India, *Rasayan Journal Chemistry*, **3**, No. 4 (2010), 649-654.

[2] C.O. Akinbile, M.S. Yusoff, Environmental impact of leachate pollution on groundwater supplies in Akure, Nigeria, *International Journal of Environmental Science and Development,* **2**, No. 1 (2011), 81-89.

[3] T.C. Cavaller, T.L. Lavy, J.D. Mattice, Persistence of Selected Pesticides in Ground-Water Samples, *Ground Water*, **29** No. 2 (1991), 225-231.

[4] S. G. Daraigan, A. S. Wahdain, A. S. Ba-Mosa, M. H. Obid, Linear correlation analysis study of drinking water quality data for Al-Mukalla City,

Hadhramout, Yemen, *International Journal of Environmental Sciencs*, **1**, No. 7 (2011), 1692-1701.

[5] J. Das, R.K. Sahoo, B.K. Sinha, Urban Ground Water Pollution: A Case Study in Cuttack City, India, *Ground Water Monitoring & Remediation*, **22**, No. 3 (2002), 65-103.

[6] N.J. Raju, Seasonal evaluation of hydro-geochemical parameters using correlation and regression analysis, *Research Communications*, **91**, No. 6 (2006), 820-827.

[7] M. A. Joarder, F. Raihan, J.B. Alam, S. Hasanuzzaman, Regression Analysis of Ground Water Quality Data of Sunamganj District, Bangladesh, *International Journal of Environmental Research*, **2**, No. 3 (2008), 1735-6865.

[8] T.D. Steele, A Bivariate-Regression Model for Estimating Chemical Composition of Streamflow or Groundwater, *Hydrological Sciences*, **21**, No. 1 (1976), 149-161.

[9] S.K. Golfinopoulos, G. B. Arhonditsis, Multiple regression models: A methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics, *Chemosphere*, **47** (2002), 1007-1018.

[10] R. Muñoz-Carpena, A. Ritter, Y. C. Li, Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to Everglades National Park, *Journal of Contaminant Hydrology*, **80** (2005), 49-70.

[11] B.E. Logan, D. LaPoint, Treatment of perchlorate- and nitrate-contaminated groundwater in an autotrophic, gas phase, packed-bed bioreactor, *Water Research*, **36** (2002), 3647-3653.

[12] G. Bitton, S.R. Farrah, R.H. Ruskin, J. Butner, Y.J. Chou, Survival of Pathogenic and Indicator Organisms in Ground Water, *Ground Water*, **21**, No. 4 (1983), 405-410.

[13] W.E. Sanford, Correcting for Diffusion in Carboon-14 Dating of Groundwater, *Groundwater*, **35**, No. 2 (1997), 357-361.

[14] A. Thapinta, P. F. Hudak, Use of geographic information systems for assessing groundwater pollution potential by pesticides in Central Thailand, *Environment International*, **29**, No. 1 (2003), 87-93.

[15] A.R. Lawrence, D.C. Gooddy, P. Kanatharana, W. Meesilp, V. Rammnarong, Groundwater evolution beheath Hat yai, a rapidly developing city in Thailand, *Hydrogeology Journal*, **8** (2000), 564-575.

[16] A.D. Gupta, M.H. Nachabe, Evaluation of Ground Water Monitoring Network by Principal Component Analysis, *Groundwater*, **39**, No. 2 (2001), 181-191.

[17] P. A. Lewis, Distribution of the Anderson-Darling Statistic, *The Annals of Mathematical Statistics*, **32**, No. 4 (1961), 1118-1124.

[18] J. Durbin, G.S. Watson, Testing for Serial Correlation in Least Squares Regression. II, *Biometrika*, **38**, No. 2 (1951), 159-177.

[19] T.S. Breusch and A.R. Pagan, A Simple Test for heteroscedasticity and Random Coefficient Variation, *Econometrica*, **47**, No. 5 (1979), 1287-1294.

[20] G.E. Box, D.R. Cox, An Analysis of Transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, No. 2 (1964), 211-252.

[21] M.H. Kutner, J.N. Christopher, J. Neter, *Applied liner regression models*, 4th ed, McGraw-Hill/Irwin , USA, 1996.