Cluster and regression analysis for predicting salinity in groundwater

Phiraphat Aphiphan[1,*], *Uma* Seeboonruang [2], and *Somyot* Kaitwanidvilai[3]

[1]Master degree student, Environmental and energy engineering for sustainability, Department of Civil Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand, 10520.

[2,3]Assoc.Prof.Dr, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand, 10520.

Abstract. Groundwater salinity is a major problem particularly in the northeastern region of Thailand. Saline groundwater can cause widespread saline soil problem resulting in reducing agricultural productivity as in the Lower Nam Kam River Basin. In order to better manage the salinity problem, it is important to be able to predict the groundwater salinity. The objective of this research was to create a cluster-regression model for predicting the groundwater salinity. The indicator of groundwater salinity in this study was electrical conductivity because it was simple to measure in field. Ninety-eight parameters were measured including precipitation, surface water levels, groundwater levels and electrical conductivity. In this study, the highest groundwater salinity at 3 wells was predicted using the combined cluster and multiple linear regression analysis. Cross correlation and cluster analysis were applied in order to reduce the number of parameters to effectively predict the quality. After the parameter selection, multiple linear regression was applied and the modeling results obtained were R^2 of 0.888, 0.918, and 0.692, respectively. This linear regression model technique can be applied elsewhere in the similar situation.

1 Introduction

Groundwater is an important natural resource for ecosystem, organism and human living. The groundwater salinity problem was important to study at the risk areas. Groundwater quality monitoring information was required in the interested area. However, continuous water quality measurement spent a lot of time and investment.

The salinity of groundwater may be derived from natural salts and human actions. The important issue of this research is groundwater salinity due to rapid population, rapid industrialization growth and the use of enormous chemicals in agriculture because of poor management. The flow of water in groundwater will increase the saline and the evaporation process will bring the saline to surface cause salinity problems. In the long term, salinity problem will severely affect the environment such as salty soil, water salinity and a shortage of fresh water. Electrical conductivity was a useful indicator of saline in this research because it was easy to measure in fieldwork. A few numbers of literatures were used regression equation for groundwater quality prediction data in different areas. Modeling used time series techniques was an alternative tool that can be used to determine the relationship between water quality and index variables for unknown parameters [1,6,8,9]. So predicting the fluctuation of groundwater quality is important for proper water The title is set in bold 16-point Arial, justified. The first letter management and land use [5,9]. Ground water samples from different area have been analyzed for correlation between electrical conductivity and parameters. An attempt has been made to develop linear regression for predicting the concentration of water quality constituents having significant correlation coefficients with electrical conductivity. Water quality indexes prediction was applied by multiple linear regression modeling [2,7]. Cluster analysis was used to assess the water quality and it useful to manage, control pollution and protect water quality [12].

This research is necessary to predict the salinity of groundwater for better planning or proper management about salinity area in the future because it is the one of the main causes of salinity problems. So the purpose of this study was to create a cluster-regression model for predicting the groundwater salinity and finds the relationships of variables.

2 Methodology

2.1 Data Preparation

First step in this research, the lower basin data was focused area for this study. It was sub-basin of the Mekong River basin located in Nakhon Phanom province in the northeastern of Thailand [11]. Fig. 1 shows well locations. In this area, groundwater faces salinity problem because this area may have salt pits or

* Corresponding author: jack_ozaka@hotmail.com

rock salt which dissolves into high conductivity value. The data was collected into monthly from December 2010 to October 2012. The data of the groundwater were from 27 wells, each of well has water level, pH and EC data.There were data of 12 water gates and rainfall data from 5 areas. Data obtained from previous research was incomplete. It was necessary to convert data with average data, remove the abnormal values and interpolation data (piecewise linear interpolation method) by the software, PAST (Paleontological Statistics Software Package for Education and Data Analysis). Fig. 2 shows incomplete data that we have presented by plotting graph and it shows the missing data and this is not suitable for next process. Fig. 3 shows the complete graph and this graph has no missing data because the missing data has fulfilled by interpolation and outlier to the data. This data is a sample of the conductivity data of BLK_Y, BDD_P, BDD, BDD_C and BDY_S wells. We had the descriptive statistics of incomplete and complete data, we have compared the incomplete data with the complete data to observe the variance and standard deviation of the data.

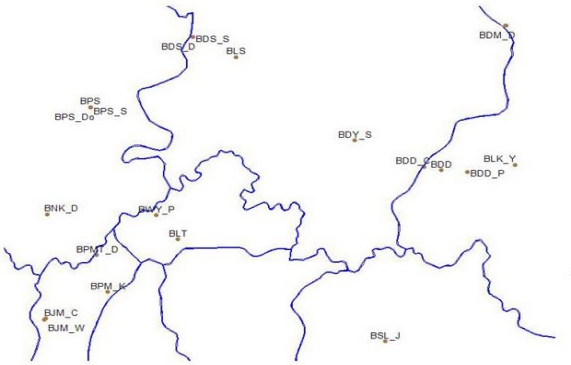

Fig. 1. Well locations

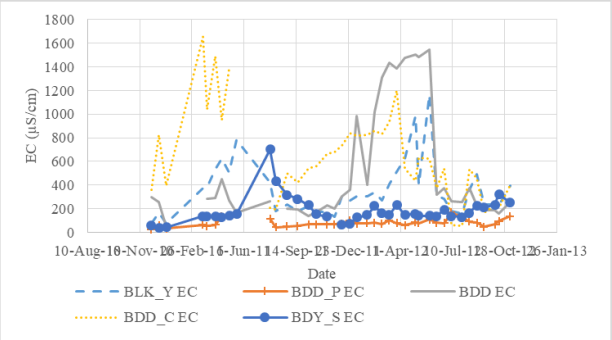

Fig. 2. Samples of incomplete data

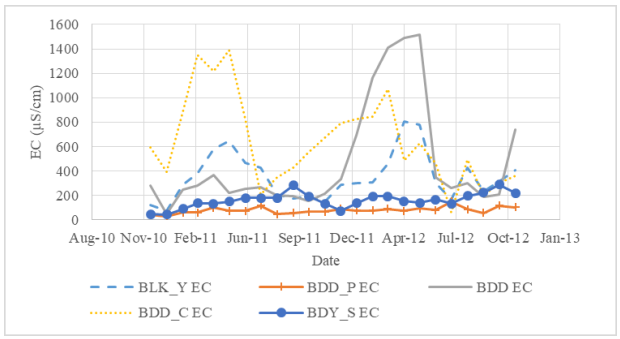

Fig. 3. Samples of complete data

2.2 Cross correlation analysis

After the data preparation step, the data has been set. The variables used in this research were 98 variables, which included conductivity, pH, groundwater level of well, precipitation and surface water level data. We would like to create 3 regression models. So we created three experiments. We used 3 highest electrical conductivity as the main variables in 3 experiments which in each experiment had 1 main variable be dependent variable and remaining 97 variables be independent variables. These dependent variables were Y_1, Y_2 and Y_3 and they were the electrical conductivity of BDS_S, BDS_D and BPS_Do wells, they had highest electrical conductivity were 27365.58 µS/cm, 134402.334 µS/cm, 95527.84 µS/cm, which their conductivity value higher than conductivity of seawater. The remaining 95 variables were defined as independent variables X_1 to X_{95}. In cross correlation step, we need to create 3 experiments by each experiment had 98 variables and we used 1 main variable (dependent variable) matching one by one with 98 independent variables. Cross correlation is the finding relationship between variable and variable of lag time. In general, Cross correlation is given by Eqn. (1)

$$r_k(x_t,y_t) = \frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(X_{t-k}-\bar{X})}{\sqrt{\sum_{t=1}^{n}(X_t-\bar{X})^2\sum_{t=1}^{n}(Y_t-\bar{Y})^2}} \quad \text{and}$$

$$r_{-k}(x_t,y_t) = \frac{\sum_{t=k+1}^{n}(Y_{t-k}-\bar{Y})(X_t-\bar{X})}{\sqrt{\sum_{t=1}^{n}(X_t-\bar{X})^2\sum_{t=1}^{n}(Y_t-\bar{Y})^2}} \quad (1)$$

When r is coefficient cross correlation, y is dependent variable, x is independent variable. Cross correlation is useful for finding the relationships of two variables and it can explain about lag time.

2.3 Cluster analysis

In this step, the data from cross correlation analysis will be classified. If any variable had no relationship from cross correlation, it will be cut off. Correlated variables were grouped by cluster analysis by the software, SPSS (Statistical Package for the Social Sciences). Representative data selection by Selected the most cross correlated value data of each group to represent the group. This step is important. This step was the process of grouping the variables and finding the representatives of the groups to simplify the regression analysis. The

variables in this research were quite numerous. It was necessary to find the representatives of them in order to make predictions and the grouping will be in the dendrogram. Example of dendrogram is shown in Fig. 4.

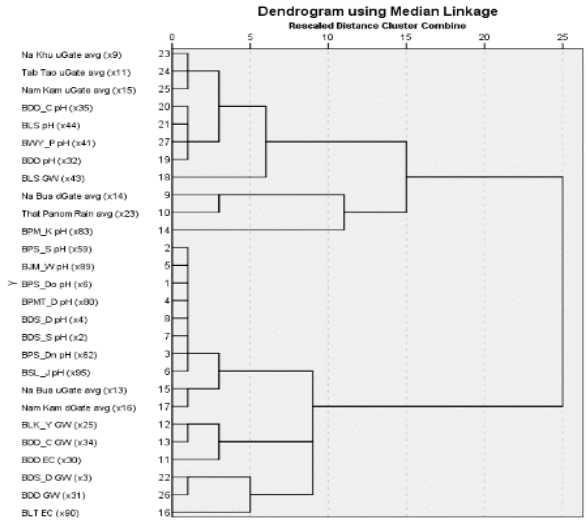

Fig. 4. Example Dendrogram of cluster Y3

2.4 Multiple linear regression

In final step, after grouping each group's representative data was used for create multiple linear regression models. A multiple linear regression model is applied to study any linear relationship between one dependent variable and several of independent variables. The multiple regression model is given by Eqn. (2)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m \quad (2)$$

When Y is the dependent variable, b_o is the intercept, b_i is the regression coefficients or slope in linear regression. Regression model is the equation for find the relationship of variables and predicting interested variable. This model is very useful in statistics and time series. The regression model can be used to predict the accuracy of forecasting.

3 Result and Discussion

3.1 Measured and interpolated data comparison

According to research, In the process of preparing the complete data. It had made the missing data range more complete and result in lower standard deviation and variance. This made the data suitable for data entry for cross correlation analysis.

3.2 Parameter elimination by cross correlation

In the cross correlation process, we had 3 main variables which in each main variable had matching with other variables and it shows in Table 1.

Table 1. The number of cross correlation between dependent variable with independent variable

Dependent variable	Number of independent variables
Y_1	69
Y_2	71
Y_3	27

3.3 Grouping variables by cluster analysis

In the clustering process, we had 3 cluster models by importing cross correlated data into grouping and in each cluster models were many groups in itself. So we used the variances and standard deviations to decide how many groups in each models and the number of appropriate groups for each models is shown in Table 2.

Table 2. Group of cluster model

Cluster Model	groups
Y_1	8
Y_2	4
Y_3	10

3.4 Multiple linear regression

For the regression process, Table 3 shows the model of Y_1 Y_2 Y_3 had R^2 of 0.888, 0.918 and 0.692 respectively. The statistical significance of models is 95%, the regression models of Y_1 Y_2 Y_3 can be described by regression equations with R^2 of 88.8%, 91.8%, 69.2%, respectively, indicated the regression models appropriated for predicting the dependent variables, which were electrical conductivity at 3 wells. Regression model of Y_1, Y_2 and Y_3 were

$Y_1 = 26489.345 + 20522.46X_{76} - 21411.258X_{77}$

$Y_2 = 88786.137 + 0.871X_{60} - 36632.942X_3$

$Y_3 = - 199546.303 + 33901.825X_{62} + 25275.749X_{35} - 15236.37X_{83}$

Fig. 5 shows an example of graph plotting between observed data (Y_{obs}) and computed data (Y_{fc}) for Y_1 model. This graph shows these lines were very similar and the mean absolute percent error of observed data (Y_{obs}) and computed data (Y_{fc}) for Y_1 is 6.01%. It can be explained that the computed data was close to the observed data. It possible to explain the computed data, it can use to be a substitute for observed data.

Table 3. Models summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Y_1	.942	.888	.877	2010.42372
Y_2	.958	.918	.910	11245.21347
Y_3	.832	.692	.646	12133.25304

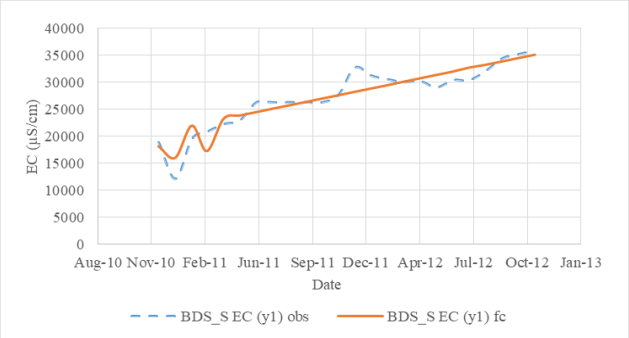

Fig. 5. Example of graph plotting between observed data (Yobs) and computed data (Yfc)

4 Conclusion

Groundwater is very important natural resource. Groundwater salinity is a huge problem for agriculture and environment. Predicting groundwater salinity is vital for appropriate planning and management. In this research, electrical conductivity at 3 observation wells was predicted by using the new cluster – regression models. Cross correlation and cluster analysis techniques were applied in order to eliminate unnecessary variables for prediction. Then, multiple regression equations were formed for the 3 locations. The models can predict the measured groundwater salinity very well. This model can be applied elsewhere to predict other variables.

References

1. Seeboonruang U., A Multiple Regression Analysis for Predicting Salinity in Shallow Groundwater. King Mongkut's Institute of Technology Ladkrabang. *IMEC 2017*.
2. Joarder M.A.M., Raihan F., Alam J.B. and Hasanuzzaman S., Regression analysis of ground water quality data of Sunamganj District, Bangladesh, *Int. J. Environ. Res. Vol* **2**(3), pp. 291-296 (2008).
3. Aflatooni M. and Mardaneh M., Time series analysis of ground water table fluctuations due to temperature and rainfall change in Shiraz plain. *International Journal of Water Resources and Environmental Engineering Vol.* **3**(9), pp. 176-188 (2011).
4. Hajigholizadeh M. and Assefa M., Melesse. Assortment and spatiotemporal analysis of surface water quality using cluster and discriminant analyses. *Catena 151* p.247–258 (2017).
5. Anderson M.P. and Woessner W.W., Applied Groundwater Modeling.Academic Press, San Diego 1992.
6. Maiti S. and Tiwari R.K., A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction. *Environmental Earth Sciences Vol* **71** pp.3147 (2014).
7. Agarwal M. and Agarwal A., Linear Regression And Correlation Analysis Of Water Quality Parameters: A Case Study Of River Kosi at District Rampur, India. *International Journal of Innovative Research in Science, Engineering and Technology Vol.* **2**, 2013.
8. Seeboonruang U., An application of time-lag regression technique for assessment of groundwater fluctuations in a regulated river basin: a case study in Northeastern Thailand. *Environmental Earth Sciences Vol* **73** pp 6511–6523 (2015).
9. Jain C.K. and Sharma M.K., Relationship among Water Quality Parameters of Groundwater of Jammu District. *Pollution Research,* **16** (4), 241-246 (1997).
10. Hayashi M., *Temperature-Electrical Conductivity relation of water for environmental monitoring and geophysical data*. University of Calgary (2003).
11. Thailand Mineral Resource Department (TMRD). *The study of salt rock in the Lower Namkam Basin Irrigation Project-Nakhon Panom Province*. Final Report, Thailand (1998).