Predictive Modelling Physico-chemical Properties Groundwater in Coastal Plain Area of Vinhlinh and Giolinh Districts of Quangtri Province, Vietnam

Nguyen Hong Giang
ThuDauMot University
Tran Dinh Hieu
ThuDauMot University
Hoang Ngo Tu Do (hoangngotudo@hueuni.edu.vn)
Hue University

Short Report

Keywords: Groundwater, Machine learning, Physico-chemical properties, Prediction

Posted Date: October 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1005920/v1

Predictive Modelling Physico-Chemical Properties Groundwater In Coastal Plain Area of Vinhlinh and Giolinh Districts of Quangtri Province, Vietnam

Nguyen Hong Giang[1] , Tran Dinh Hieu[1*], Hoang Ngo Tu Do[2*]
[1]Faculty of Architecture, ThuDauMot University, ThuDauMot 820000, Vietnam, Giangnh@tdmu.edu.vn, hieutd@tdmu.edu.vn
[2]Faculty of Geology and Geography of Sciences University, Hue University, 49118, Vietnam, hoangngotudo@@hueuni.edu.vn
*Correspondence: hieutd@tdmu.edu.vn, hoangngotudo@hueuni.edu.vn. Tel.: +8490.515.3333.

Abstract

This paper presents to study the performance of machine learning techniques consisting of Multivariate Adaptive Regression Spline(MARS), Multilayer Perceptron (MLP), and Decision Tree Regression (DTR) for estimating physico-chemical properties groundwater in coastal plain area in Vinhlinh and Giolinh districts of Quangtri province of Vietnam. To deploy the MLP and DTR, different types of transfer and kernel functions were tested, respectively. Determining the results of MARS, MLP and DTR showed that three models have suitable carrying out for forecasting water quality components. Comparison of outcomes of MARS model with MLP, DTR models indicates that this model has good performance for forecasting the elements of water quality, its level of accuracy is slightly more than other. To assess the accurate values of the models according to the measurement parameters indicated that order models were MARS, DTR, and MLP, respectively.

Keywords: Groundwater, Machine learning, Physico-chemical properties, Prediction.

Highlights

- Machine learning methods are used for spatial modeling of physico-chemical properties of groundwater.
- MARS performances suitable precision compared to the DTR and MLP models.
- The quality of water parameters (i.e. CaCo3, Ca, and Co2) of the coastal plain area was predicted.

1. Introduction

The attendance of contaminants in natural freshwater consider one of the most crucial environmental problems in many areas of developing countries, where several communities are hardly approaching a potable water supply [1]. Low-income communities, which lean on the untreated surface and groundwater supplies for domestic and agricultural purposes are the most affected by poor water quality [2]. Unfortunately, they are also do not get adequate tools to monitor quality of water regularly [3,4]. Thus, they are increasingly expected to obtain reliable assessments of quality of water, which can be used [5].

To evaluate the quality of water for drinking and agricultural irrigation that several variables are routinely monitored. This process makes big database, but it can be time-consuming for data acquisition while the accurate rendering of the multivariate data may be defying.

With regard to use machine learning for forecasting physico-chemical parameters in water, using Artificial Neural Network (ANN) estimated river water quality components [6-13]; employing MARS predict physicochemical in water [14-18]; deploying DTR forecast quality of water [19-26]. Furthermore. MARS, MLP, and DTR model also belongs to nonparametric learning, and the model is used in those areas [27-36].

This paper presents the prediction of the physico-chemical properties of ground-water using MLP, DTR, and MARS models. The input vectors used in the models are leaned on 290 samples that were collected from 290 wells of households in coastal plain area in Giolinh and Vinhlinh districts of Quangtri province. This study also highlights comparison among three models that base on the results of statistic accuracy parameters such as Mean (M), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Standard Deviation (St Dev), Pearson Correlation Coefficient (R), Skewness Coefficient (Skew), Minimum (Min), Maximum (Max), and Correlation of Determination (R^2), and Nash–Sutcliffe Efficiency (NSE). The collection of results of these three models may show the working efficiency of the models for the predictive quality of water.

The structure of the paper is organized as follows. Section 1 gives the paper introduction. Section 2 presents the MARS, MLP, and DTR models and explains them clearly for understanding use throughout this paper. Section 3 describes the study results and discussions. Finally, Section 4, and Section 5 introduce the discussions and conclusions.

2. Methodology

2.1. Multivariate Adaptive Regression Splines

MARS model is a novel approach in soft computing, and it is a nonparametric regression model, and it was introduced by Friedman [37]. MARS seems like a method for a fitted relationship between independent and dependent variables in each desired phenomenon. MARS supports techniques for modelling systems with high accuracy, which is based on a dataset [38-40]. The MARS algorithm feature is the procedure of the backward and forwards stepwise, at the same time may explain and control the complex nonlinear mapping between the inputs and output variables. MARS model is highlighting input variables that have a note-worthy effect on the output variables. The general form of MARS is described as below:

$$y = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \quad (1)$$

where, *y* is output variables, β_0 is constant value, *M* is the number of functions, $h_m(x)$ is M_{th} basis function and β_m is the corresponding coefficient of $h_m(x)$. Furthermore, $h_m(x)$ shows information about the relationship between input and output variables, and it is described as below:

$$h_m(x) = Max(0, C - x) \; or \; h_m(x) = Max(C - x, 0) \quad (2)$$

where *h* is the basis function, *x* is the input variables, and *C* is the threshold value of the independent (input) variables of *x*. It is notable that *C* is called *"knots"* or *"hinges"*.

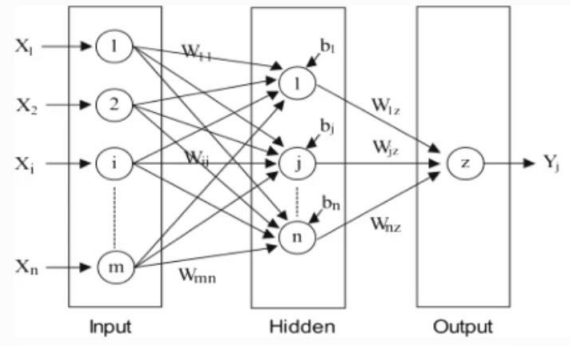
The function of backward stepwise function relates to removing basis functions one at a time until the criterion of *"lack of fit"* is a minimum. In the deletion of backward stepwise, the last important basic functions are destroyed one at a time. The lack of used fitting measurement is based on the Generalized Cross-Validation (GCV) [41]:

$$CGV = A * \sum_{j=1}^{P} (y_i - \hat{f}(x)/N \quad (3)$$

where $A = [1 - \frac{C(M)}{N}]^{-2}$ and $C(M) = 1\, trace[B(B^{-}B)^{-1}B']$ are the complexity function[37]. The GCV criterion is considered the average of residual error multiplied by a penalty to modify for the variability associated with more parameters estimation in the model [42].

2.2. Multilayer perception

MLP model is a member of neural network method [43]. It may simulate arbitrarily complex nonlinear processes for any systems in terms of inputs and outputs. A MLP structure in **Figure 1** demonstrates a three-layer neural network consisting of inputs layer, hidden layer (layers) and outputs layer [44,45]. With regards shown in **Figure 1**, w_i is the weight and b_i is the bias for each neuron. The values of weight and bias can be assigned progressively and corrected during the training process in order to compare predicted outputs with known outputs. As networks are often trained using a backpropagation algorithm [45].



Fiure 1. Diagram of MLP architecture

In this paper, the experiments have found the optimal model with the values of the core parameters that showed in **Table 1**.

Table 1. MLP basic component

Number of inputs	2	Number of outputs	1
Number of hidden layers	4	Hidden layer sizes	100/100/100/100
Learning rate init	0.001	Alpha	0.0001
Momentum	0.9	Max iter	10000
Iter no change	10	Power_t	0.5
Beta 1	0.9	Beta 2	0.999

2.3. Decision tree regression

A decision tree is a data structure that includes nodes and edges. Mentions about a member of the decision tree, tree C4.5 algorithm is supervised learning method [46,47]. The C4.5 algorithm inherently employs a single-pass pruning process of overfitting mitigation, and it can do with both discrete and continuous data. Tree C4.5 algorithm consists of four steps to generate a decision tree. Four steps consist of choosing an attribute as a root, making branch every value, putting the dataset in the wing, and repeating the second process until every class has the same value, respectively. The formula of Entropy is indicated as below:

$$Entropy\ (S) = \sum_{i=1}^{n} (-p_i)(Log_2 p_i \quad (4)$$

where *S* is entropy, and *p* is class proportion in the output. In addition, the attribute with the highest gain value is used as the root attribute. Equation 5 points out the formula of the Gain as below:

$$Gain\ (S, A) = Entropy(S) \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S) \quad (5)$$

where, *S* is a set of case; *A* is an attribute of case; /S_i/ is a number of cases to *i*; and /S/ is number of cases in the set.

2.4. Performance metrics

Forecasting results are based on the calculation and comparison of the actual values to the forecasted values. These metrics of the *a*ccuracy measurement parameters include the MSE, MAE, RMSE, R^2, R. Furthermore, the error metrics are defined as follows [48-50]:

$$MSE = \frac{\sum_{t=1}^{n}(x_t - x_t')^2}{n} \quad (6)$$

$$\mathrm{MAE} = \frac{\sum_{t=1}^{n}|x_t - x_t'|}{n} \quad (7)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(x_t - x_t')^2}{n}} \quad (8)$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(x_t - x_t')^2}{\sum_{t=1}^{n}\left(x_t - \frac{1}{n}\sum_{t=1}^{n} x_t\right)^2} \quad (9)$$

$$R = \frac{\sum_{t=1}^{n}(x_t - \bar{x})\,(x_t' - \bar{x'})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(x_t' - \bar{x'})^2}} \quad (10)$$

$$NSE = 1 - \frac{\sum_{t=1}^{n}(x_t - x_t')^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \quad (11)$$

where x_t, x_t' are the estimated value and observed value in the period time *t,* and *n* is the number of the observed values in the testing data. $\bar{x}$, $\bar{x'}$ are mean of the observed value. The R^2 and R should be approaching *1* to indicate strong model performance, and the MSE, MAE, and RMSE should be as close to *zero* as possible.

The methodology of this study is described by the diagram in **Figure 1**.

2.5. Study area

The place of study is about 450 km^2 and covers the Giolinh and Vinhlinh coastal plain of Quangtri province of Vietnam. It is surrounded by Quangbinh province in the north, Thachhan river in the south, 50–150 m high hills in the west, and the East Sea (see **Figure. 2**). The coastal plain is relatively flat with an elevation between 0 and 5m except for coastal sand-dunes with 11–22m high, which provide a natural embankment system for seawater prevent [51]. During the dry season (from June to August), the saltwater pervades (i.e. where total mineralization of water M=1g/l) is often inspected at about 30

and 35km from the main tributaries of Benhai, Hieu, and Thachhan rivers from the estuary [52]. Thus, groundwater from dug wells and shallow wells scattered in the coastal is the main water source for drinking and domestic use for residents.

3. Results

3.1. Data collection

Study data includes 290 groundwater samples obtained at wells of households in coastal plain area of Vinhlinh and Giolinh districts. The predominant chemical compositions in these samples consisted of three main ingredients as Calcium Carbonate (CaCo3, mg/l) Calcium (Ca, mg/l), and Carbon Dioxide (Co2, mg/l). In addition, there were some other physico-chemical components (as Ammonium, Magie, and Iron Oxide), but their contents were not significant in these samples. Three input variables include two independent variables such as Ca and Co2, and one dependent variable as CaCo3, which was collected from five provinces. The statistical characteristic results are also pointed out in **Table 2**. The range of the following characteristics was computed from the observed monthly precipitation time series: the Mean, Min, and Max values, St Dev, Skew. The input data patterns of 290 items were randomly selected with two parts. The first part was used for the training phase, which contained about 70% of the entire data. The second part was used for the test phase, which contained about the remaining 30%.

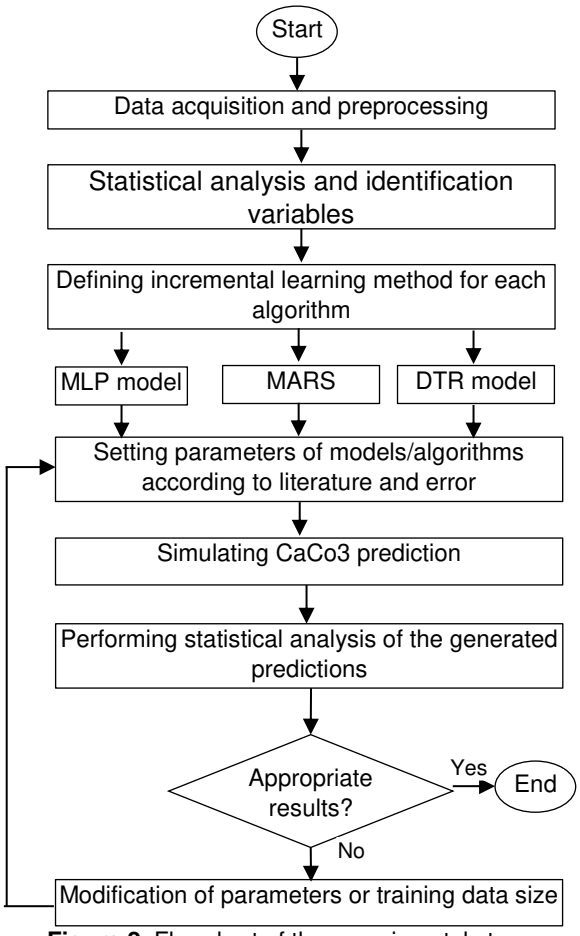


Figure 2. Flowchart of the experimental steps conducted in this study

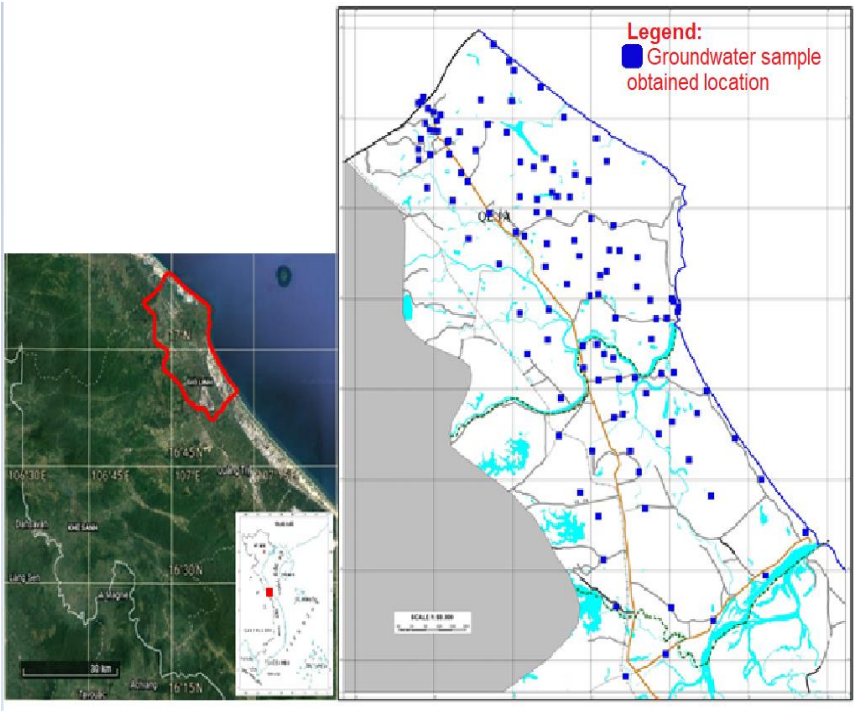


Figure 3. The loaction for samples collection.

Table 1. Statistical characteristics of physico-chemical components data

Item	St Dev	Mean (mg/l)	Min (mg/l)	Max (mg/l)	Skewness
CACO3	4.23	1.3	0	25.8	5.06
Ca	16.1	6.05	0	87.55	3.57
Co2	2.42	0.79	0	12	3.56

3.2. Analysis of Simulation Results

MARS, MLP, and DTR models were generated in the Python platform. The output function of MARS is presented as below:

MARS = $7.907 - 0.249F_1 - 0.129F_2$, where $F_1 = \max(0, Ca-55.79)$, $F_2 = \max(0, 55.79-Ca)$. F_i is the basis function. F_1 may be explained as the maximum value of 0 and Ca-55.79. The minus sign ahead of the maximum value is equivalent to a minimum value. In addition, the MARS analysis indicates that the most important is Ca. Furthermore, the output function for MLP and DTR do not occur.

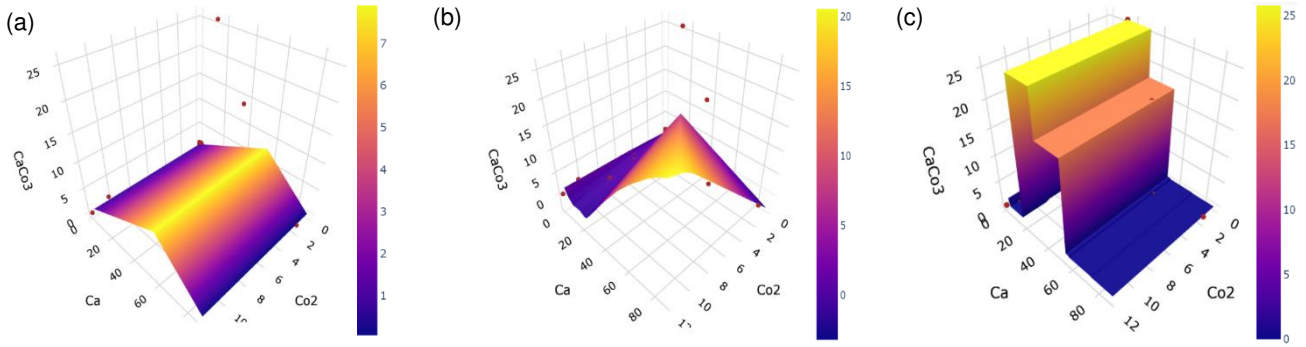


Figure 4. Physico-chemical properties prediction with (a) MARS model, (b) MLP model, and (c) DTR model

The data in **Figure 4(a, b, c)** shows the relationship between the three variables. The content of Co2 and Ca increase leads to the content of CaCo3 increase. The MLP model makes a forecasting form that resembles a cone shape. In the meantime, the DTR and MARS charts look like the image of papers with some folds. Through these three images, it is hard to judge which model gives the best estimating. Hence, the values of performance metrics of the three models are presented in **Table 2**, **Figure 5 – 6.** The R^2 and NSE of the three models for the training and testing phases are from 0.889 to 0.947 and are closer to 1. In addition, The MAE, RMSE, and MSE values are also

Table 1. Statistical characteristics of physico-chemical components data

Item	St Dev	Mean (mg/l)	Min (mg/l)	Max (mg/l)	Skewness
CACO3	4.23	1.3	0	25.8	5.06
Ca	16.1	6.05	0	87.55	3.57
Co2	2.42	0.79	0	12	3.56

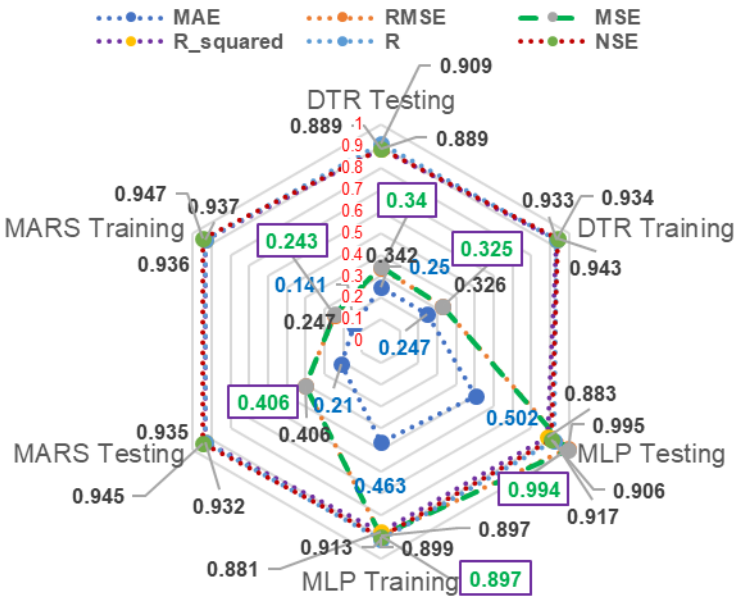


Figure 5. The radar chart of accuracy parameters

from 0.21 to 0.995 and are closer to 0. These show that the forecast results are very consistent compared with the actual data. As for the experimental results for each specific model, it indicates that MARS model for training phase with R^2, NSE MAE, RMSE, and MSE values are 0.937, 0.947, 0.141, 0.243, and 0.247 respectively, and these properties are better than DTR and MLP models. Regarding the testing phase results, the highest accuracy for forecast is MARS model with R^2 = 0.935 and NSE = 0.945, the second-highest is MLP model with R^2 = 0.917 and NSE = 0906, and the lowest is DTR model with R^2 = 0.889 and NSE = 0.889.

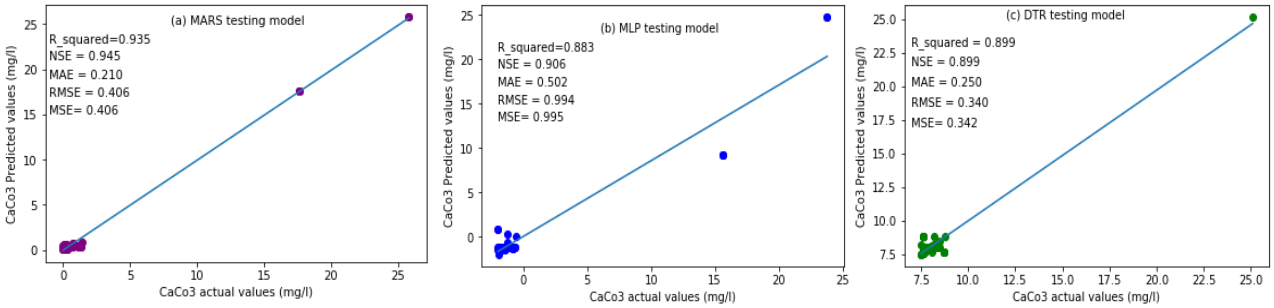


Figure 6. The best performance indicators for CaCo3 prediction **(a)** MARS testing model, **(b)** MLP testing model, **(c)** DRT testing model

Table 2. Accuracy parameters for physico-chemical components prediction

Parameter	DTR		MLP		*MARS*	
	Testing	***Training***	***Testing***	***Training***	***Testing***	***Training***
MAE	0.250	0.247	0.502	0.463	0.210	0.141
RMSE	0.340	0.325	0.994	0.897	0.406	0.243
MSE	0.342	0.326	0.995	0.897	0.406	0.247
R^2	0.889	0.933	0.883	0.881	0.935	0.937
R	0.909	0.934	0.917	0.913	0.932	0.936
NSE	0.889	0.943	0.906	0.899	0.945	0.947
GCV					0.136	0.136

The Taylor charts check the performance of estimated and actual values based on the standard deviation and CC, which are contained simultaneously in assessing the respective models [53,54]. The standard deviation and CC between the actual and predicted datasets for the models are present in the Taylor diagram, and it also can be seen overall consistency between observed and estimated values when the CC value is approaching up to 1, as pointed in **Figure 7**. This can be considered for the MARS model with $CC_{\text{training phase}}$ = 0.936, $CC_{\text{testing phase}}$ = 0.932, MLP model with $CC_{\text{training phase}}$ = 0.913, $CC_{\text{testing phase}}$ = 0.917, and DTR model with $CC_{\text{training phase}}$ = 0.934, $CC_{\text{testing phase}}$ = 0.909. The large number of correlation coefficients indicate that there is a strong relationship. The Taylor plot also shows that these models are optimal with the highest accuracy [53]. In other words, if the standard deviation of the

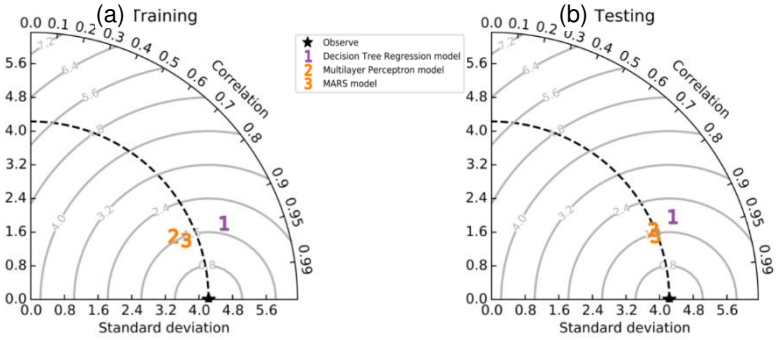


Figure 7. Taylor diagram representing the best performance of MARS, MLP, DTR models at **(a)** Training phase, **(b)** Testing phase.

predicted value of the higher standard deviation of the observed value, it will lead to an over estimation and vice versa [55]. In addition, Furthermore, GCV indicator of MARS brings about equilibrium between flexibility and generalization ability of the function of MARS model [56].

4. Discussions

This paper described a comparative study and analysis of MARS, MLP, and DTR models in estimating physico-chemical properties of groundwater. The different circumstances, influential factors, and indicators have been observed for the experimentation. The following key findings are as the predictive errors in the case of the models decreased if the testing set decreased; MARS was the highlight in comparison to other models. The detection of CaCo3, Ca, and Co2 attendance in drinking water is a cause of public health. Therefore, it also indicated that the overfitting existence of any of these forms of chemical component renders the water unfit for drinking. Hence the factors need to be any additional studies in order to interpret the widespread contamination of the two districts' groundwater by the chemical items. Due to the limitation of groundwater analysis equipment, the study did not detect any more chemicals in the water that affect public health in these two areas. However, the results of this study also contribute to supporting local authorities to have appropriate solutions to help households use clean water.

5. Conclusions

In this study, using the MARS, DTR, and MLP predicted physico-chemical properties groundwater in coast plain of Vinhlinh and Giolinh, which is located in the north middle of Vietnam. For phases of training and testing carried out in the models, the observed data consisting of Co2 and Ca was used as inputs, while CaCo3 was used as output. The stimulated results pointed out that the three models have a high suitable presentation for forecasting water quality components. The best performance was related to the MARS. The results of DTR and MLP also showed that their accuracy is a suitable presentation for practical purposes. Furthermore, the carrying out of a comparison of three models showed that the outcomes of MARS and DTR models were slightly more reliable in comparing with MLP.

Author Contributions

Conceptualization, Discussion, Writing, Material, Methods, and Funding, Acquisition, Review, and Editing: Nguyen Hong Giang, Tran Dinh Hieu; Data collection, Writing: Hoang Ngo Tu Do. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Thu Dau Mot University, VietNam. The APC was funded by Thu Dau Mot University.

Acknowledgments

We would like to thank Quangtri province of Departments of Natural Resources and Environment, and Science and Technology. We also thank the Vinhlinh and Giolinh residence.

Data availability

All data generated or analysed during this study are included in this published article.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

[1] Organisation mondiale de la santé, Światowa Organizacja Zdrowia, World Health Organization, & World Health Organisation Staff. (2004). *Guidelines for drinking-water quality* (Vol. 1). World health organization.
Google Scholar

[2] Ayoko, G. A., Singh, K., Balerea, S., & Kokot, S. (2007). Exploratory multivariate modeling and prediction of the physico-chemical properties of surface water and groundwater. *Journal of Hydrology*, *336*(1-2), 115-124.
Google Scholar

[3] Resh, V. H. (2008). Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. *Environmental Monitoring and Assessment*, *138*(1), 131-138.
Google Scholar

[4] Omarova, A., Tussupova, K., Berndtsson, R., Kalishev, M., & Sharapatova, K. (2018). Protozoan parasites in drinking water: A system approach for improved water, sanitation and hygiene in developing countries. *International journal of environmental research and public health*, *15*(3), 495.
Google Scholar

[5] Bonansea, M., Rodriguez, M. C., Pinotti, L., & Ferrero, S. (2015). Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). *Remote Sensing of Environment*, *158*, 28-41.
Google Scholar

[6] Zamani, M. A. T., & Saybani, M. ARTIFICIAL NEURAL NETWORK MODEL FOR PREDICTION OF ENVIRONMENTAL STATUS OF URBAN CATCHMENT OF PENCHALA RIVER, KUALA LUMPUR, MALAYSIA.
Google Scholar

[7] Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, *53*(1), 3-13.
Google Scholar

[8] İlhan, N., Yetiş, A. D., Yeşilnacar, M. İ., & Atasoy, A. D. S. (2021). Predictive modelling and seasonal analysis of water quality indicators: three different basins of Şanlıurfa, Turkey. *Environment, Development and Sustainability*, 1-35.

Google Scholar

[9] Khan, F. M., Gupta, R., & Sekhri, S. (2021). Superposition learning-based model for prediction of E. coli in groundwater using physico-chemical water quality parameters. *Groundwater for Sustainable Development*, *13*, 100580.
Google Scholar

[10] Najah, A., El-Shafie, A., Karim, O. A., & El-Shafie, A. H. (2014). Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. *Environmental Science and Pollution Research*, *21*(3), 1658-1670.
Google Scholar

[11] Raheli, B., Aalami, M. T., El-Shafie, A., Ghorbani, M. A., & Deo, R. C. (2017). Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. *Environmental Earth Sciences*, *76*(14), 1-16.
Google Scholar

[12] Niroobakhsh, M., Musavi-Jahromi, S. H., Manshouri, M., & Sedghi, H. (2012). Prediction of water quality parameter in Jajrood River basin: application of multi layer perceptron (MLP) perceptron and radial basis function networks of artificial neural networks (ANNs). *African Journal of Agricultural Research*, *7*(29), 4131-4139.
Google Scholar

[13] Emamgholizadeh, S., Kashi, H., Marofpoor, I., & Zalaghi, E. (2014). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology*, *11*(3), 645-656.
Google Scholar

[14] Haghiabi, A. H. (2016). Prediction of longitudinal dispersion coefficient using multivariate adaptive regression splines. *Journal of Earth System Science*, *125*(5), 985-995.
Google Scholar

[15] Esmaeilbeiki, F., Nikpour, M. R., Singh, V. K., Kisi, O., Sihag, P., & Sanikhani, H. (2020). Exploring the application of soft computing techniques for spatial evaluation of groundwater quality variables. *Journal of Cleaner Production*, *276*, 124206.
Google Scholar

[16] Ahmadi, M. H., Mohseni-Gharyehsafa, B., Farzaneh-Gord, M., Jilte, R. D., Kumar, R., & Chau, K. W. (2019). Applicability of connectionist methods to predict dynamic viscosity of silver/water nanofluid by using ANN-MLP, MARS and MPR algorithms. *Engineering Applications of Computational Fluid Mechanics*, *13*(1), 220-228.
Google Scholar

[17] Najafzadeh, M., & Ghaemi, A. (2019). Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environmental monitoring and assessment*, *191*(6), 1-21.
Google Scholar

[18] Bhatt, A. H., Karanjekar, R. V., Altouqi, S., Sattler, M. L., Hossain, M. S., & Chen, V. P. (2017). Estimating landfill leachate BOD and COD based on rainfall, ambient temperature, and waste composition: Exploration of a MARS statistical approach. *Environmental Technology & Innovation*, *8*, 1-16.
Google Scholar

[19] Jalal, D., & Ezzedine, T. (2020, June). Decision Tree and Support Vector Machine for Anomaly Detection in Water Distribution Networks. In *2020 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 1320-1323). IEEE.
Google Scholar

[20] Lu, H., & Ma, X. (2020). Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere*, *249*, 126169.
Google Scholar

[21] Saghebian, S. M., Sattari, M. T., Mirabbasi, R., & Pal, M. (2014). Ground water quality classification by decision tree method in Ardebil region, Iran. *Arabian journal of geosciences*, *7*(11), 4767-4777.
Google Scholar

[22] Liao, H., & Sun, W. (2010). Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method. *Procedia Environmental Sciences*, *2*, 970-979.
Google Scholar

[23] Gakii, C., & Jepkoech, J. (2019). A classification model for water quality analysis using decision tree.
Google Scholar

[24] Jaloree, Shailesh, Anil Rajput, and Sanjeev Gour. "Decision tree approach to build a model for water quality." *Binary Journal of Data Mining & Networking* 4 (2014): 25-28.
Google Scholar

[25] Chandanapalli, S. B., Reddy, E. S., & Lakshmi, D. R. (2018). DFTDT: distributed functional tangent decision tree for aqua status prediction in wireless sensor networks. *International Journal of Machine Learning and Cybernetics*, *9*(9), 1419-1434.
Google Scholar

[26] He, Q., Dong, Z., Zhuang, F., Shang, T., & Shi, Z. (2012, July). Parallel decision tree with application to water quality data analysis. In *International symposium on neural networks* (pp. 628-637). Springer, Berlin, Heidelberg.

Google Scholar

[27] Genuer, R., Poggi, J. M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research*, *9*, 28-46.
Google Scholar

[28] Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. M. (2020). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, 107312.
Google Scholar

[29] Bengio, Y., Delalleau, O., & Simard, C. (2010). Decision trees do not generalize to new variations. *Computational Intelligence*, *26*(4), 449-467.
Google Scholar

[30] Al Iqbal, M. R., Rahman, S., Nabil, S. I., & Chowdhury, I. U. A. (2012, December). Knowledge based decision tree construction with feature importance domain knowledge. In *2012 7th international conference on electrical and computer engineering* (pp. 659-662). IEEE.
Google Scholar

[31] Khan, Y. A., Shan, Q. S., Liu, Q., & Abbas, S. Z. (2020). A nonparametric copula-based decision tree for two random variables using MIC as a classification index. *Soft Computing*, 1-16.
Google Scholar

[32] Zheng, X., Dan, C., Aragam, B., Ravikumar, P., & Xing, E. (2020, June). Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics* (pp. 3414-3425). PMLR.
Google Scholar

[33] Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., & Khazaeni, Y. (2019, May). Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning* (pp. 7252-7261). PMLR.
Google Scholar

[34] Khaldi, R., El Afia, A., & Chiheb, R. (2019). Forecasting of BTC volatility: comparative study between parametric and nonparametric models. *Progress in Artificial Intelligence*, *8*(4), 511-523.
Google Scholar

[35] Devianto, D., Permathasari, P., Yollanda, M., & Ahmad, A. W. (2020, May). The Model of Artificial Neural Network and Nonparametric MARS Regression for Indonesian Composite Index. In *IOP Conference Series: Materials Science and Engineering* (Vol. 846, No. 1, p. 012007). IOP Publishing.
Google Scholar

[36] Kohler, M., KRZY ZAK, A., & Langer, S. (2019). Deep learning and MARS: a connection. *stat*, *1050*, 8.
Google Scholar

[37] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.
Google Scholar

[38] Sekulic, S., & Kowalski, B. R. (1992). MARS: a tutorial. *Journal of Chemometrics*, *6*(4), 199-216.
Google Scholar

[39] Steinberg, D. (2001). An alternative to neural nets: multivariate adaptive regression splines [57]. *PC AI*, *15*(1), 38-41.
Google Scholar

[40] Gutiérrez, Á. G., Schnabel, S., & Contador, J. F. L. (2009). Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecological Modelling*, *220*(24), 3630-3637.
Google Scholar

[41] Attoh-Okine, N. O., Cooger, K., & Mensah, S. (2009). Multivariate adaptive regression [57] and hinged hyperplanes (HHP) for doweled pavement performance modeling. *Construction and Building Materials*, *23*(9), 3020-3023.
Google Scholar

[42] LeBlanc, M., & Tibshirani, R. (1994). Adaptive principal surfaces. *Journal of the American Statistical Association*, *89*(425), 53-64.
Google Scholar

[43] Khoshhal, J., & Mokarram, M. (2012). Model for prediction of evapotranspiration using MLP neural network. *International Journal of Environmental Sciences*, *3*(3), 1000-1009.
Google Scholar

[44] Ramchoun, H., Idrissi, M. A. J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer Perceptron: Architecture Optimization and Training. *Int. J. Interact. Multim. Artif. Intell.*, *4*(1), 26-30.
Google Scholar

[45] Ashiquzzaman, A., & Tushar, A. K. (2017, February). Handwritten Arabic numeral recognition using deep learning neural networks. In *2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 1-4). IEEE.
Google Scholar

[46] Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, *23*(6), 529-535.
Google Scholar

[47] Kamber, M., & pei, J. (2006). *Data Mining.* Morgan kaufmann.
Google Scholar

[48] Kardani, N., Zhou, A., Nazem, M., & Shen, S. L. (2020). Estimation of bearing capacity of piles in cohesionless soil using optimised machine learning approaches. *Geotechnical and Geological Engineering*, *38*(2), 2271-2291.
Google Scholar

[49] Yang, J. H., & Yang, M. S. (2005). A control chart pattern recognition system using a statistical correlation coefficient method. *Computers & Industrial Engineering*, *48*(2), 205-221.
Google Scholar

[50] Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533-1543.
Google Scholar

[51] Krutwagen, M. (2007). *Impact of shrimp pond wastewater on the estuaries and the issue of salinity intrusion in the Quang Tri Province* (Bachelor's thesis, University of Twente).
Google Scholar

[52] Tam, V. T., Batelaan, O., Le, T. T., & Nhan, P. Q. (2014). Three-dimensional hydrostratigraphical modelling to support evaluation of recharge and saltwater intrusion in a coastal groundwater system in Vietnam. *Hydrogeology Journal*, *22*(8), 1749-1762.
Google Scholar

[53] Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183-7192.
Google Scholar

[54] Ghorbani, M. A., Deo, R. C., Yaseen, Z. M., Kashani, M. H., & Mohammadi, B. (2018). Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. *Theoretical and applied climatology*, *133*(3), 1119-1131.
Google Scholar

[55] Abba, S. I., Pham, Q. B., Usman, A. G., Linh, N. T. T., Aliyu, D. S., Nguyen, Q., & Bach, Q. V. (2020). Emerging evolutionary algorithm integrated with kernel principal component analysis for modeling the performance of a water treatment plant. *Journal of Water Process Engineering*, *33*, 101081.
Google Scholar

[56] Deo, R. C., Samui, P., & Kim, D. (2016). Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. *Stochastic Environmental Research and Risk Assessment*, *30*(6), 1769-1784.
Google Scholar

214