proceedings

MDPI

Proceedings

Estimating Chlorophyll-a and Dissolved Oxygen Based on Landsat 8 Bands Using Support Vector Machine and Recursive Partitioning Tree Regressions †

Nimisha Wagle [1], Tri Dev Acharya [2,3,4] and Dong Ha Lee [3,*]

[1] Department of Survey, Minbhawan, Kathmandu 44600, Nepal; wagle1996@gmail.com
[2] Institute of Industrial Technology, Kangwon National University, Chuncheon 24341, Korea; tridevacharya@kangwon.ac.kr
[3] Department of Civil Engineering, Kangwon National University, Chuncheon 24341, Korea
[4] School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing 102616, China
* Correspondence: geodesy@kangwon.ac.kr; Tel.: +82-33-250-6232
† Presented at the 6th International Electronic Conference on Sensors and Applications, 15–30 November 2019. Available online: https://ecsa-6.sciforum.net/.

Published: 14 November 2019

Abstract: In general, water quality mapping is done by interpolation of in situ measurement samples. Often, these parameters change with time. Due to geographic variability and the lack of budget in Nepal, such measurements are done less often. Remote sensors that collect spectral information continually can be very useful in the regular monitoring of water quality parameters. Landsat Operational Land Imager (OLI) bands have been used to estimate water quality parameters. In this work, we model two water quality parameters: chlorophyll-a (Chl-a) and dissolved oxygen (DO) using sequential minimal optimization regression (SMOreg), which implements a support vector machine (SVM) algorithm and recursive partitioning tree (REPTree) regressions. A total of 19 measurements were taken from Phewa Lake, Nepal and various secondary bands were derived from using Landsat 8 Operational Land Imager (OLI) bands. These bands underwent feature selection, and regression models were created based on selected bands and sample data. The results showed satisfactory modelling of water quality parameters using Landsat 8 OLI bands in Phewa Lake. Due to a limited number of data, cross-validation was done with 10 folds. The SVM showed a better result than the REPTree regression. For future studies, the performance can be further evaluated in large lakes with larger sample numbers and other water quality parameters.

Keywords: classification; SMOreg; REPTree; surface water; Landsat; Phewa Lake

1. Introduction

Water is one of the significant environments for living animals to endure. Living organisms inside water resources are facing a great threat from a wide range of physical processes, including land use/land cover change, pollution, and global climate change as well as human interventions [1]. Lakes and their supplies store assets and fulfil both human necessities, ranging from drinking water to diversion, and natural prerequisites to help significant levels of biodiversity [2]. Due to the expanding populace developments and the fast pace of modernization and urbanization areas, as well as climate change, water quality is being deteriorated. These phenomena will continue to increase even more in the future, and many types of research have recognized declining water quality as one of the most crucial threats to society [3]. This has led to a growing need for the monitoring of water quality parameters in lakes and reservoirs. Water quality is measured based on various

physical, chemical and biological parameters. Chlorophyll-a (Chla-a) and dissolved oxygen (DO) are very important parameters for determining water quality.

Chl-a is the major indicator of a trophic state because it acts as a link between nutrient concentration, particularly phosphorus, and algal production. A eutrophication phenomenon is often related to Chl-a concentrations [4]. Eutrophication, determined by the algal bloom, is an enrichment of water by nutrient salts that causes structural changes to the ecosystem, which causes degradation in water quality and depletion of fish species [5]. Similarly, DO refers to the level of free, non-compound oxygen present in water or other liquids. It is an important parameter in assessing water quality because it influences the organisms living within a body of water. In limnology (the study of lakes), DO is an essential factor second only to water itself [6]. A DO level that is too high or too low can harm aquatic life and affect water quality. So, Chl-a and DO are major parameters to determine water quality.

The major purpose of this study is to estimate water quality parameters using sequential minimal optimization regression (SMOreg) and recursive portioning (REPTree) regression techniques based on Landsat 8 bands. For this, secondary bands were derived from the fundamental Operational Land Imager (OLI) bands, and these bands underwent feature selection with the sample data obtained from the in situ measurements. Finally, models were created using two different machine learning regression techniques. The machine learning techniques were used for regression analysis because these techniques reduce human error, and the modelling can be done using multiple variables which gives better accuracy.

2. Materials and Methods

2.1. Case Study

The Landsat imagery of June 2017 was used for the regression purpose. The in situ measurements from the same month was also taken for the preparation of the training dataset.

The case study area is Phewa Lake (Figure 1), which is located in the Kaski district of Nepal. This lake is the second largest lake in Nepal, with an area of 5.2 sq.km [7]. In the southern and western parts of the lake, there are hills and trees, and there is a less human settlement in those areas, whereas in the eastern and northern parts of the lake, there is a huge human settlement that affects the quality of water.

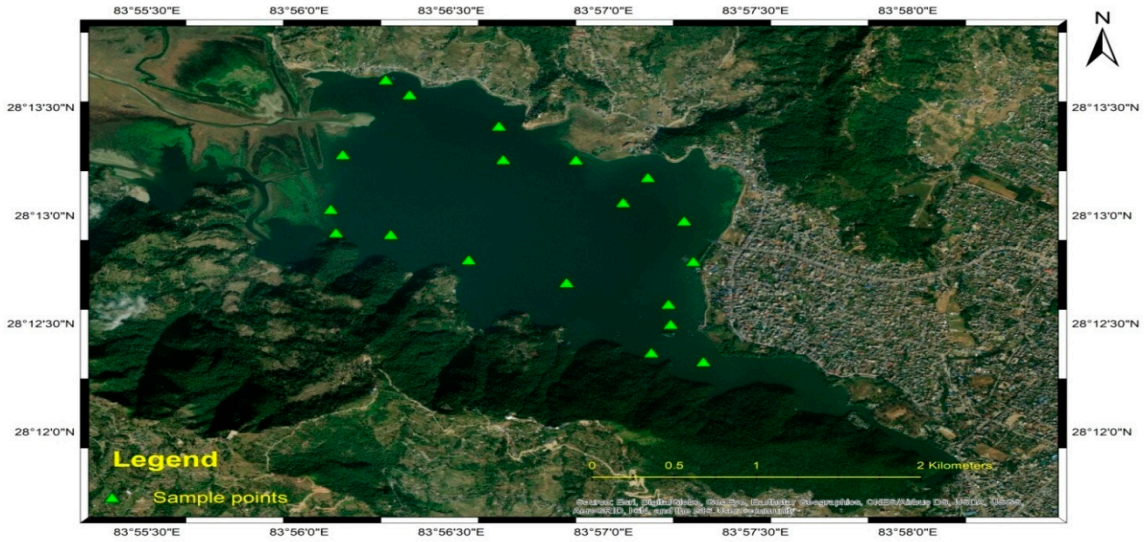

Figure 1. Location of sample points on Phewa Lake.

2.3. Method

After preprocessing the at-satellite reflectance, the images were first used to extract all seven OLI bands' values, from which different secondary bands were derived. The derived secondary bands were obtained from the difference of the various bands, a ratio of the various OLI bands, and the normalized difference of the bands and logarithmic multiplication of the various bands. A total of 44 secondary bands were derived. All the derived bands were not necessarily useful, so band selection was done by ranking the bands according to the correlation coefficient values with Chl-a and DO, and only those bands which had a higher correlation with Chl-a and DO were chosen for the training dataset. The selection of the variables was done in the Waikato Environment for Knowledge Analysis (WEKA) software. After forming the training dataset, two regression methods were applied: REPTree and SMOreg.

Recursive partitioning (REPTree) is a kind of binary tree utilized for grouping or regression assignments. It plays out a hunt over every conceivable split by expanding a data proportion of the node polluting influence and then choosing the covariate demonstrating the best split [8]. Recursive partitioning creates a decision tree that correctly classifies members by splitting them into sub-populations based on several dichotomous independent variables. It is easy to understand and attempt to limit the utilization of all given datasets.

A support vector machine (SVM) can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin) [9]. SMOreg uses the same principles as the SVM for classification, with only a few minor contrasts. However, the main idea is always the same; that is, to minimize the error by individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated. It tries to fit the error within a certain threshold.

The modelling was done on the WEKA software. For SMOreg, 19 instances were used as the training data set, and cross-validation with 10 folds was used for the validation, as the data were in limited numbers. The SMOreg function was used for the modelling purpose. The correlation coefficient obtained from the modelling was high enough for further prediction. The same obtained model was used for the prediction of the Chl-a and DO values of the different sample points. Later, these sample values were interpolated and maps were prepared in a GIS platform. Similarly, for the REPTree tree-based regression, WEKA was used. The obtained model was used for the predictions of Chl-a and DO, as was done in SMOreg.

3. Results and Discussion

Out of 44 derived bands, only 2 bands had a high correlation with Chl-a. For the SMOreg method in estimating Chl-a, two derived bands were used for the modelling purpose: Green-Blue and Red-Blue. The equation obtained from the SMOreg was:

$$\text{Chl-a} = 0.7981*(\text{Normalized})\text{Green-Blue}-0.0447*(\text{Normalized})\ (\text{Red-Blue}) \times 0.0583 \quad (1)$$

For DO in the SMOreg method, out of 44 bands, only 6 derived bands were used for the modeling purpose as they had a better correlation co-efficient. The bands were: TCT_B, TCT_G, Blue*log(Red), Red*log(Blue), Blue*Blue*Log(Red), Blue+Red. The equation obtained from the SMOreg was:

$$\begin{aligned}\text{DO} = {} & 0.1667 * (\text{Normalized})\ \text{DVI} + 0.2554 * (\text{Normalized})\ \text{NIR} + \text{Red} + \text{Green} + 0.3506 * (\text{Normalized}) \\ & \text{TCT\_B} + 0.0326 * (\text{Normalized})\ \text{TCT\_G} - 0.1676 * (\text{Normalized})\ \text{Blue*Log(Red)} + 0.0531 * (\text{Normalized}) \\ & \text{Red*Log(Blue)} - 0.4457 * (\text{Normalized})\ \text{Blue*Blue*Log(Red)} + 0.182 * (\text{Normalized})\ \text{Blue+Red} + 0.2102.\end{aligned} \quad (2)$$

However, for the REPTree, only one band (Green-Blue) was used as it had a high correlation value for the Chl-a. The formed tree is shown in Figure 2a. Similarly, for DO in the REPTree, the same variables were used as in the SMOreg. The tree formed is shown in Figure 2b.

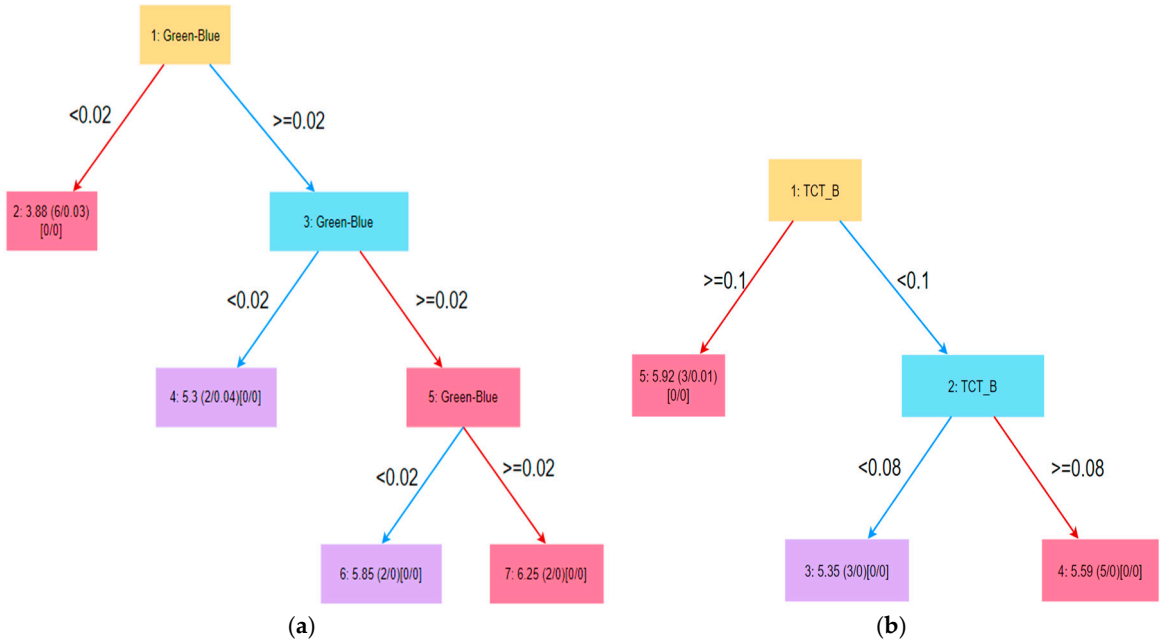

Figure 2. Tree formed by recursive partitioning tree (REPTree) for water quality estimation from Landsat 8: (**a**) chlorophyll-a; (**b**) dissolved oxygen.

In this work, a 10-fold cross-validation technique was used because there were limited numbers of instances. It was used to fix the problem of overfitting. Results from the cross-validation are given in Table 1.

Table 1. Results of 10-fold cross-validation for both methods.

Measures	Chlorophyll-a		DO	
	SMOreg	**REPTree**	**SMOreg**	**REPTree**
Correlation coefficient	0.8102	0.7846	0.8879	0.8059
Mean absolute error	0.4199	0.4433	0.0718	0.1133
Root mean squared error	0.6043	0.6525	0.1042	0.1342
Relative absolute error	40.5519%	42.5538%	36.6678%	57.89%
Root relative squared error	53.5106%	59.7607%	42.2431%	54.42%

The maps obtained from the data interpolation are show in in Figure 3.

From the analysis of the DO maps, we found that the area of the lake which is affected by human intervention has a low amount of dissolved oxygen. On the other hand, the area near the forest has a high amount of dissolved oxygen. Similarly, from the analysis of the chlorophyll maps, the shore of the lake has a high amount of Chl-a. The algal substances swept from the middle and collected inshore can be the reason for this. Since Chl-a is the primary indicator of phytoplankton, the Chl-a maps indicate that phytoplankton is abundant near the shore and less phytoplankton in the middle of the lake. The regression model formed using the SMOreg is a multivariate model. Multivariate models are capable of assessing a large number of variables and interrelations are, therefore, more successful in defining and predicting the values [5], and this multivariate model's accuracy is better than the single variable model obtained from the REPTree.

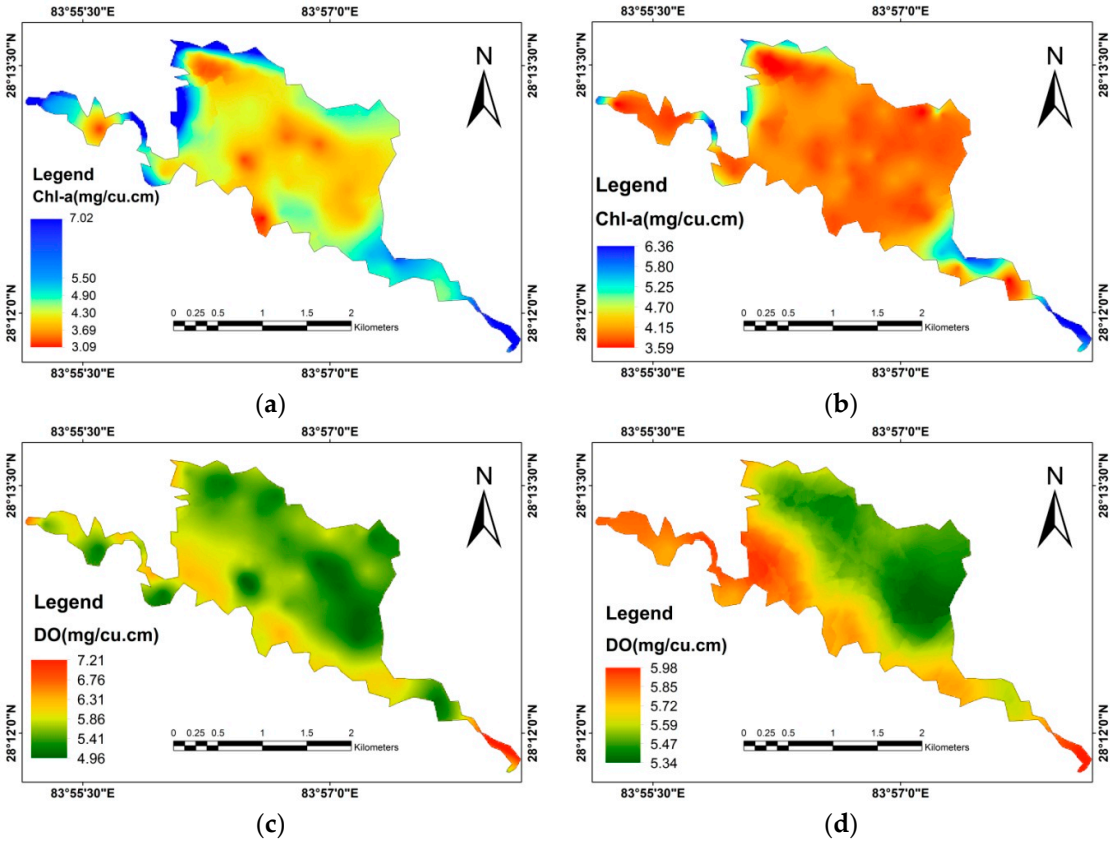

Figure 3. (**a**) Chl-a map prepared by using SMOreg of Phewa Lake; (**b**) Chl-a map prepared by using REPTree of Phewa Lake; (**c**) DO map prepared by using SMOreg of Phewa Lake; (**d**) DO map prepared by using REPTree of Phewa Lake.

4. Conclusions

In this study, Landsat OLI bands were utilized for finding water quality parameters. A machine learning approach was used for the modelling purpose. Two regression methods, namely SMOreg and REPTree, were used for the modelling. Various secondary bands were derived from the primary OLI bands. Thus, these bands along with the data from the in situ measurements were utilized to create a regression model for predicting the values of the water quality parameters, i.e., Chl-a, of the lake. From the present study, it is concluded that the SMOreg creates a better model of regression analysis than the REPTree. These machine learning techniques are better than the other regression analysis techniques as they create a model using multivariable, a model which is the best fit for the instances. The present study also concludes the efficacy of Landsat imagery for establishing a cost-effective method for determining the value of the water quality parameters of the lake. Routine observation of lake water quality using remote sensing may be considered by different organizations as an alternative method to field survey for recording and processing water quality information for various works including fisheries. For future studies, the performance can be further evaluated in large lakes with larger sample numbers and other water quality parameters.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mushtaq, F.; Pandey, A.C. Assessment of land use/land cover dynamics vis-à-vis hydrometeorological variability in Wular Lake environs Kashmir Valley, India using multitemporal satellite data. *Arab. J. Geosci.* **2014**, *7*, 4707–4715.

2. Ismail, K.; Boudhar, A.; Abdelkrim, A.; Mohammed, H.; Mouatassime, S.; Kamal, A.; Driss, E.; Idrissi, E.; Nouaim, W. Evaluating the potential of Sentinel-2 satellite images for water quality characterization of artificial reservoirs: The Bin El Ouidane Reservoir case study (Morocco). *Meteorol. Hydrol. Water Manag.* **2019**, *7*, 31–39.
3. Torbick, N.; Hession, S.; Hagen, S.; Wiangwang, N.; Becker, B.; Qi, J. Mapping inland lake water quality across the Lower Peninsula of Michigan using Landsat TM imagery. *Int. J. Remote Sens.* **2013**, *34*, 7607–7624.
4. Han, L.; Jordan, K.J. Estimating and mapping chlorophyll-a concentration in Pensacola Bay, Florida using Landsat ETM + data. *Int. J. Remote Sens.* **2005**, *26*, 5245–5254.
5. Liu, X.; Fei, D.; He, G.; Liu, J. Use of PCA-RBF model for prediction of chlorophyll-a in Yuqiao Reservoir in the Haihe River Basin, China. *Water Sci. Technol. Water Supply* **2014**, *14*, 73–80.
6. Kramer, D.L. Dissolved oxygen and fish behavior. *Environ. Biol. Fishes* **1987**, *18*, 81–92.
7. Rai, A.K. Evaluation of natural food for planktivorous fish in Lakes Phewa, Begnas, and Rupa in Pokhara Valley, Nepal. *Limnology* **2000**, *1*, 81–89.
8. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; CRC Press: New York, NY, USA, 2017; ISBN 9781351460491.
9. Drucker, H.; Surges, C.J.C.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. In *Proceedings of the Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1997; Volume 1, pp. 155–161.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).