MODELLING THE CHLOROPHYLL-A CONCENTRATION OF LAGUNA LAKE USING HIMAWARI-8 SATELLITE IMAGERY AND MACHINE LEARNING ALGORITHMS FOR NEAR REAL TIME MONITORING

E.R.G. Martinez [a], R.J.L. Argamosa [b], R.B. Torres [c], A.C. Blanco [b, c]

[a] Philippine Science High School - Main Campus, Diliman, Quezon City, 1101 - b2022ermartinez@pshs.edu.ph
[b] Training Center for Applied Geodesy and Photogrammetry, University of the Philippines Diliman, Quezon City, 1101
[c] Department of Geodetic Engineering, University of the Philippines Diliman, Quezon City, 1101

Commission 4, WG 7

KEY WORDS: Laguna Lake, Chlorophyll-a Concentration, Himawari-8, Machine Learning, C2RCC.

ABSTRACT

Recent studies have investigated the use of satellite imaging combined with machine learning for modelling the Chlorophyll-a (Chl-a) concentration of bodies of water. However, most of these studies use satellite data that lack the temporal resolution needed to monitor dynamic changes in Chl-a in productive lakes like Laguna Lake. Thus, the aim of this paper is to present the methodology for modelling the Chl-a concentration of Laguna Lake in the Philippines using satellite imaging and machine learning algorithms. The methodology uses images from the Himawari-8 satellite, which have a spatial resolution of 0.5-2 km and are taken every 10 minutes. These are converted into a GeoTIFF format, where differences in spatial resolution are resolved. Additionally, radiometric correction, resampling, and filtering of the Himawari-8 bands to exclude cloud-contaminated pixels are performed. Subsequently, various regression and gradient boosting machine learning algorithms are applied onto the train dataset and evaluated, namely: Simple Linear Regression, Ridge Regression, Lasso Regression, and Light Gradient Boosting Model (LightGBM). The results of this study show that it is indeed possible to integrate algorithms in Machine Learning in modelling the near real-time variations in Chl-a content in a body of water, specifically in the case of Laguna Lake, to an acceptable margin of error. Specifically, the regression models performed similarly with a train RMSE of 1.44 and test RMSE of 2.51 for Simple Linear Regression and 2.48 for Ridge and Lasso Regression. The linear regression models exhibited a larger degree of overfitting than the LightGBM model, which had a 2.18 train RMSE.

1. INTRODUCTION

Several studies have linked Chlorophyll-a (Chl-a) concentration with the trophic status of an aquatic ecosystem, which can be classified as oligotrophic, mesotrophic, or eutrophic (Sakamoto, 1966), (National Research Council (US) Committee on Water Quality Criteria, 1974), (Dobson et al., 1974), (Jones et al., 1979). Thus, analysis of the current conditions and predictive modelling of the Chl-a concentration of a body of water may support decisions for its proper management. Recent studies have modeled Chl-a concentrations using imaging and machine learning. In 2014, Kim et al. (2014) proposed machine learning approaches to coastal water quality monitoring using GOCI satellite data. This study highlighted that hourly available GOCI images were useful in discussing spatiotemporal distributions of the water quality parameters with tidal phases in the west coast of Korea. More recently, the study of Li et al. (2021) discussed the remote estimation of Chl-a based on artificial intelligence that can provide an effective and robust method to monitor the lake eutrophication on a macro-scale, and offer a better approach to elucidate the response of lake ecosystems to global change. Furthermore, Saberioon et al. (2020) proposed that Sentinel-2A, when coupled with machine learning algorithms, could be employed as a reliable, inexpensive, and accurate instrument for monitoring the biophysical status of small inland waters like fishponds and sandpit lakes. Several studies have also investigated the case of Chl-a concentration in Laguna Lake. Specifically, Blanco et al. (2020) used regression analysis of Sentinel-3 OLCI images to produce algal classification maps, Syariz et al. (2019) used neural networks trained on the same Sentinel-3 data to produce models that outperform existing 3-band and 2-band models in terms of accuracy, and Jalbuena et al. (2019) inputted Landsat-8 data in the Bio-Optical Model Based tool for Estimating water quality and bottom properties from Remote sensing images (BOMBER) tool and processed this data through the Water Color Simulator (WASI) tool to produce Chl-a maps.

Most of these studies, however, have relied on freely available satellite images with temporal resolution ranging from a few hours to a few weeks. This is not enough to capture the short time period variations of Chl-a, especially in highly productive lakes like Laguna Lake. As a result, there is the need for real time monitoring of Chl-a to serve as an early warning on potential fish kills resulting from hypoxia.

In line with this, the objective of this paper is to present the methodology for modelling the Chl-a concentration Laguna Lake in the Philippines using satellite imaging and machine learning algorithms. Additionally, this paper will present initial observations on the present accuracy of the models developed. Due to the dynamic nature of water quality, the Himawari-8 AHI was used to monitor Chl-a with a high temporal resolution (10 minutes) to detect changes within a single day; given that the satellite can capture the image every 10 minutes, these multiple images can be input to the model close to real time. The purpose of this study is to show the methods for modelling Chl-a using the Himawari-8 AHI spectral bands as features and Sentinel-3 OLCI C2RCC Chl-a as inputs for various regression algorithms that will be studied.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVI-4-W3-2021-211-2022

2. STUDY AREA AND METHODOLOGY

2.1 Study Area: Laguna Lake

The study area is Laguna Lake. It is the largest lake in the Philippines and is the third largest freshwater lake in Southeast Asia, with an area of 949 km^2 (Laguna Lake Development Authority, 2016). It is located in the island of Luzon and cuts across cities and municipalities including parts of the National Capital Region. This lake provides life support services, particularly to Rizal and Laguna, as a source of water for irrigation, industrial cooling, and as a domestic water supply. It also serves as a reservoir, waste sink, power generator, and transport route. Furthermore, it is a place for recreation. With these, the ecological health and integrity of Laguna Lake must be well maintained. Therefore, modelling the water quality parameters of Laguna Lake may aid in successful maintenance.

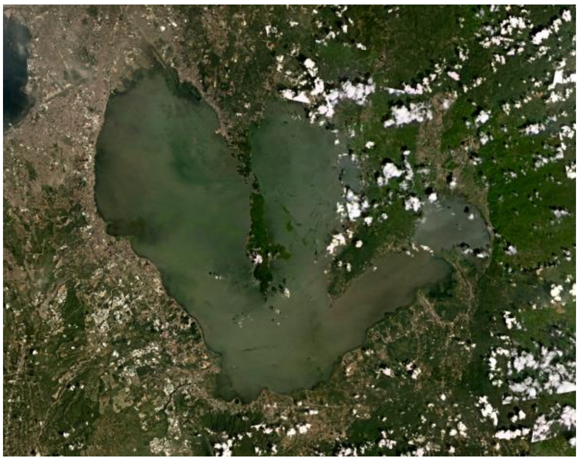

Figure 1. Satellite image of Laguna Lake and the surrounding area from Planet Explorer taken in the month of October 2019.

Figure 2. Map of the Philippines with Laguna Lake highlighted and boxed in red.

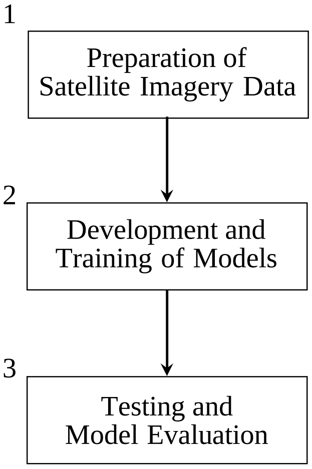
1
Preparation of Satellite Imagery Data
2
Development and Training of Models
3
Testing and Model Evaluation

Figure 3. Flowchart for the major steps in the Methodology.

2.2 Methodology

Shown in Figure 3 is the major steps in the methodology of this paper. The methodology starts with the preparation of satellite imagery, followed by the development and training of models, and finally ended with testing and model evaluation.

Preparation of Satellite Imagery Data The methodology begins with the processing of Himawari-8 image data provided by the Japan Aerospace Exploration Agency dating back to October 2019. Specifically, 4 Himawari-8 raster images and 1 Sentinel-3 raster image were obtained from 09:10 to 10:10 on 9 October 2019, when cloud cover was low. The raster images were first converted into a GeoTIFF format using the geo2grid plugin. This plugin is a set of command line tools for mapping satellite instrument files to uniform grids for further processing (Space Science and Engineering Center University of Wisconsin - Madison, n.d.). Chl-a concentration map was generated using Sentinel-3 OLCI Level-1 data from the Copernicus Open Access Hub as input to the Case 2 Regional Coast Colour (C2RCC) processor. This processor is a set of neural networks trained on simulated top-of-atmosphere reflectance which produces inherent optical properties (IOPs), Chl-a, total suspended matter (TSM), and yellow substance maps (NASA SeaDAS, n.d.). Through Kriging interpolation, the different bands were interpolated to a spatial resolution of 0.5 km to resolve differences in spatial resolution. This is followed by the radiometric correction, resampling, and filtering of the Himawari-8 bands to exclude cloud contaminated pixels. Finally, the bands are stacked into one raster image file.

Development and Training of Models The final raster image files were then processed in Python. These image files were first imported and transformed into a dataframe using the rs_learn Python library. Rows with empty or invalid values and outliers were then removed, and the various pairwise band ratios among the 4 visible and near-infrared bands were calculated. The dataframe was then split into two dataframes to be used as train and test data, using 15% of the original dataframe as test data. Three of the four models, namely Simple Linear Regression, Ridge Regression, and Lasso Regression, were then applied onto the train dataset. The last two regression models aim to address the problem of variable selection within Simple Linear Regression through regularization, with Ridge Regression minimizing the squared sum of the coefficients (L2 regularization) and Lasso Regression minimizing the absolute sum of the coefficients (L1 regularization) (Muthukrishnan and Rohini, 2016). Grid search with cross validation was used to evaluate each regression model's performance when the fold is equal to the number of rows in the training data (leave-one-out or LOO method). The fourth model, Light Gradient Boosting Model (LightGBM), a gradient boosting model which makes use of tree-based learning algorithms (LightGBM, n.d.), was tuned and evaluated by an automatic hyperparameter search module from the Optuna framework, which is a modular hyperparameter optimization software for machine learning models (Optuna, n.d.).

Testing and Model Evaluation The models were then assessed for their root mean squared errors, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (1)$$

where n = number of samples
$\hat{y}_i$ = predicted value
y_i = actual value

Moreover, coefficients for linear models and feature importance were examined wherein the coefficients determine the direction (increase or decrease) of Chl-a. Feature importance can be calculated by permuting a column of features and then creating a model from it through gradient boosting models such as LightGBM. The feature with the highest value of feature importance contributes the most to the model.

3. RESULTS AND DISCUSSION

Figure 4 presents the Chl-a concentration map of Laguna Lake. Most Chl-a is concentrated in the center of the lake, with a mode concentration of roughly 15 µg/L. Table 1 highlights the root mean squared error values for the four models mentioned previously. For the train data, the linear models produce the least amount of error, but exhibit a higher degree of overfitting as compared to LightGBM, as the train RMSE is much lower than the test RMSE as shown in Table 1. The coefficients/effect of each feature for the linear models is presented in Table 2. It must be noted that the coefficients for the Ridge Regression and Lasso Regression models are scaled to a standard normal scaling based on the train data.

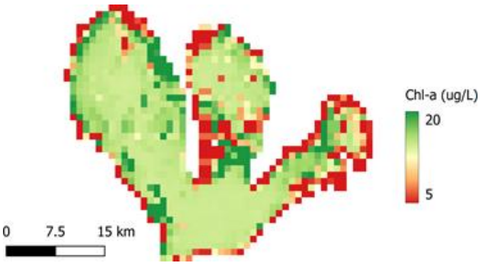

Figure 4. Chl-a concentration map of Laguna Lake produced via C2RCC. Raster image taken at 09:30 on 9 October 2019. Most Chl-a is concentrated in the center of the lake, with a mode concentration of roughly 15 µg/L

Features	Coefficients	
	Simple	Ridge/Lasso
Blue	6.1484	-0.3411
Green	-25.0678	-2.2091
Red	17.0432	3.4449
NIR	5.3999	1.0366
Blue/Green	-2.5681	-0.9750
Blue/Red	0.4069	1.1700
Blue/NIR	13.2410	0.1199
Green/Blue	209.1132	0.1101
Green/Red	-0.9421	-0.6370
Green/NIR	-19.9818	0.3851
Red/Blue	-134.3557	-3.2749
Red/Green	37.5423	-0.3465
Red/NIR	-1.0870	0.4168
NIR/Blue	-59.4948	0.2034
NIR/Green	26.5272	0.3000
NIR/Red	0.1069	-0.7075

Table 2. Coefficients of the Linear Models (Ridge and Lasso have the same coefficients). The coefficients of Simple Linear Regression can be larger due to overfitting caused by a lack of penalties.

Certain bands and band ratios have a large effect on the predicted Chl-a concentration. For example, an increase in the ratio of Green to Blue results in a significant increase in the predicted Chl-a. This is shown by the large positive values of the coefficients in all regression models. On the other hand, an increase in the ratio of Red to Blue results in a significant decrease in the predicted Chl-a, which is shown by the large negative value of the coefficients in all models. Still further, Red/NIR and NIR/Red seem to have a smaller effect on predicted Chl-a. In general, it can be observed that the simple linear regression model has numerous large coefficients when compared to Ridge or Lasso Regression. This indicates possible overfitting, which the Ridge and Lasso address through the introduction of penalties. Figure 5 summarizes the significance of features of the LightGBM model. The Blue band is found to be the feature that contributes most to the model's accuracy.

Model	Hyper-parameters	Train RMSE ((µg/cm3)	Test RMSE (µg/cm3)
Simple Linear	N/A	1.44	2.51
Ridge and Lasso	α = 1.05, solver = saga	1.44	2.48
LightGBM	objective = regression l1 (lasso penalty), learning rate = 0.068005652	2.18	2.51

Table 1. Results of the optimized models for predicting Chl-a. All models had similar Test RMSE, with the linear regression models overfitting to the train data causing a lower Train RMSE.

.

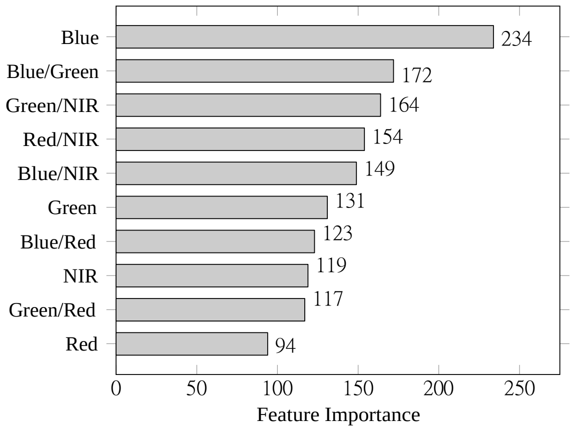

Figure 5. Important Features in the LightGBM Model. The Blue band has the largest effect on Chl-a concentration, followed by the band ratios Blue/Green and Green/NIR.

4. CONCLUSION

The results of this study have confirmed that it is possible to integrate algorithms in Machine Learning in modelling the variations in Chl-a content in a body of water, specifically in the case of Laguna Lake in the Philippines, to an acceptable margin of error. Moreover, with the high temporal resolution provided by the Himawari-8 satellite, and with the methodology presented in this paper, data collection may be conducted for several short time periods to observe the behavior of Chl-a concentration in the Lake. The models shown may, in turn, be applied to produce almost real-time monitoring software to monitor the dynamic behavior of the lake for a period of time.

5. ACKNOWLEDGEMENTS

This research was done as part of the Development of Integrated Mapping, Monitoring, and Analytical Network System for Manila Bay and Linked Environments (MapABLE) Project under the Integrated Mapping For Manila Bay (IM4ManilaBay) Program. The Project was implemented by the University of the Philippines Training Center for Applied Geodesy and Photogrammetry (TCAGP), through the support of the Department of Science and Technology (DOST) of the Republic of the Philippines and the Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD). 10-minute Himawari data for October 2019 was provided by the P-Tree System, Japan Aerospace Exploration Agency (JAXA), and the NICT Science Cloud. This paper also acknowledges the support of the Philippine Science High School – Main Campus Science Immersion Program.

REFERENCES

Blanco, A. C., Manuel, A., Jalbuena, R., Ticman, K., Medina, J. M., Gubatanga, E., Santos, A., Ana, R. S., Herrera, E., Nadaoka, K., 2020. Estimation of Chl-a concentration in Laguna Lake using SENTINEL-3 OLCI images. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 17–21.

Dobson, H. F. H., Gilbertson, M., Sly, P. G., 1974. A Summary and Comparison of Nutrients and Related Water Quality in Lakes Erie, Ontario, Huron, and Superior. *Journal of the Fisheries Research Board of Canada*, 31(5), 731-738. https://doi.org/10.1139/f74-099.

Jalbuena, R., Blanco, A., Manuel, A., Santos, J. et al., 2019. Bio optical modelling of Laguna Lake using BOMBER tool and WASI-derived inverted parameters. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, XLII-4/W16, 277–282.

Jones, R. A., Rast, W., Lee, G. F., 1979. Relationship between summer mean and maximum chlorophyll a concentrations in lakes. *Environmental Science & Technology*, 13(7), 869–870.

Kim, Y. H., Im, J., Ha, H. K., Choi, J.-K., Ha, S., 2014. Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing*, 51(2), 158–174.

Laguna Lake Development Authority, 2016. Laguna de Bay. https://llda.gov.ph/laguna-de-bay/.

Li, S., Song, K., Wang, S., Liu, G., Wen, Z., Shang, Y., Lyu, L., Chen, F., Xu, S., Tao, H. et al., 2021. Quantification of chlorophyll-a in typical lakes across China using Sentinel-2 MSI imagery with machine learning algorithm. *Science of The Total Environment*, 778, 146271.

LightGBM, n.d. LightGBM documentation. https://lightgbm.readthedocs.io/en/latest/index.html.

Muthukrishnan, R., Rohini, R., 2016. Lasso: A feature selection technique in predictive modeling for machine learning. *2016 IEEE international conference on advances in computer applications (ICACA)*, IEEE, 18–20.

NASA SeaDAS, n.d. C2RCC processor overview. https://seadas.gsfc.nasa.gov/help-8.0.0/c2rcc/C2RCCTool.html.

National Research Council (US) Committee on Water Quality Criteria, 1974. *Water Quality Criteria, 1972: A Report of the Committee on Water Quality Criteria, Environmental Studies Board, National Academy of Sciences, National Academy of Engineering, Washington, DC, 1972*. Environmental Protection Agency.

Optuna, n.d. Optuna: A hyperparameter optimization framework. https://optuna.readthedocs.io/en/stable/.

Saberioon, M., Brom, J., Nedbal, V., Souček, P., Císař, P., 2020. Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters. *Ecological Indicators*, 113, 106236.

Sakamoto, M., 1966. Primary production by phytoplankton community in some Japanese lakes and its dependence on lake depth. *Arch Hydrobiol*, 62, 1–28.

Space Science and Engineering Center University of Wisconsin - Madison, n.d. Geo2grid. https://www.ssec.wisc.edu/software/geo2grid/index.html.

Syariz, M. A., Lin, C.-H., Blanco, A. C., 2019. Chlorophyll-a concentration retrieval using convolutional neural networks in Laguna Lake, Philippines. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, XLII-4/W19, 401–405.