Machine learning approaches to estimate chlorophyll-a concentration using GOCI satellite data

Hyun-Young Choi (1), Young-Jun Kim (1), Jungho-Im (1,2)

[1] Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Ulsan 44919, Republic of Korea

[2] Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, New York, USA

Email: hyong56@unist.ac.kr; kimyj@unist.ac.kr; ersgis@unist.ac.kr

Abstract: Water quality has been an important issue in Korea since most industrial facilities and residential areas located in the coastal region and monitoring the coastal water quality has been considered important. In this study, we attempted to estimate the concentration of chlorophyll-*a* (chl-*a*) over the southern coast of the Korean peninsula using Geostationary Ocean Color Imager (GOCI) satellite data. Recently, the ocean color remote sensing technologies have been widely used in the field of water quality monitoring due to its continuous spatial distribution through a wide area. Although Korea Ocean Satellite Center (KOSC) provide some algorithms to retrieve water quality indicators such as chl-*a*, total suspended solids (TSS), and colored dissolved organic matter (CDOM), the products show low consistency with *in situ* data. To improve the accuracy of estimating chl-*a* concentration, the machine learning method of applied in this study. We used GOCI remote sensing reflectance (R_{rs}) data processed by the GOCI Data Processing System (GDPS v2.0.0) of 8 spectral bands and their ratio as the input variables of the machine learning algorithm. A total of 36 variables were initially used, and we applied the Boruta algorithm as the feature selection method to decrease the dimension of the input variables. The variables which confirmed through the feature selection were used as the final variables. *In situ* chl-*a* data was collected from the FerryBox program, which is the automatic water quality monitoring systems on ships provided by Korea Marine Environment Management Corporation (KOEM). The estimated chl-*a* concentration from GOCI data was compared with the *in situ* data from 2013 to 2016. Four machine learning approaches including Random Forest (RF), Extreme Gradient Boost (XGB), Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN) were attempted for chl-*a* estimation and the results show that RF outperformed the other three models. The coefficient of determination (R^2) and root-mean-square-error (RMSE) between the estimated and *in situ* chl-*a* was about 0.93 and 0.4572 μg/L for train dataset, 0.47 and 0.9119 μg/L for test dataset, respectively. It seems to be a quite meaningful result for estimating chl-*a* concentration compare to the performance (R^2 = 0.23) of the OC3G algorithm provided from KOSC for the same test dataset.

Keywords: Water quality, Chlorophyll-*a*, GOCI, Remote sensing, Machine learning