FEATURE SELECTION OF OPTICAL SATELLITE IMAGES FOR CHLOROPHYLL-A CONCENTRATION ESTIMATION

M. V. Nguyen [1,2], H. J. Chu [1], C. H. Lin [1], M. J. Lalu [3]

[1]Department of Geomatics, National Cheng Kung University, Taiwan – (p68087019, honejay, linhung)@mail.ncku.edu.tw
[2]Institute of Geography, Vietnam Academy of Science and Technology, Vietnam – manh.ig239@gmail.com
[3]Department of Geomatics Engineering, Institut Teknologi Sepuluh Nopember- lmjaelani@geodesy.its.ac.id

KEY WORDS: Feature Selection, Water Quality, Chlorophyll-a, Remote Sensing, Optical Satellite Images, Turbid Inland Water

ABSTRACT:

Healthy inland freshwater sources, such as lakes, reservoirs, rivers, and streams, play crucial roles in providing numerous benefits to surrounding societies. However, these inland water bodies have been severely polluted by human activities. Therefore, long-term monitoring and real-time measurements of water quality are essential to identify the changes of water quality for unexpected environmental incidents avoidance. The success of satellite-based water quality studies relies on three key components: precise atmospheric correction method, optimization algorithm, and regression model. Previous studies integrated various algorithms and regression models, including (semi-) empirical or (semi-) analytical algorithms, and (non-) linear regression models, to obtain satisfactory results. Nevertheless, the selection of appropriate algorithm is complex and challenging because of the fact that the changes in chemical and physical properties of water can lead to different method determination. To alleviate the aforementioned difficulties, this study proposed a potential integration which comprises an optimization method for efficient water-quality model selection, ordinary least squares regression, and an accurately atmospheric corrected dataset. Prime focus of this study is water-quality model selection which optimizes an objective function that aims to maximize prediction accuracy of regression models. According to the experiments, the performance of the selected water-quality model using proposed procedures, dominated that of the existing algorithms in terms of root-mean-square error (RMSE), the Pearson correlation coefficient (r), and slope of the regressed line (m) between measured and predicted chlorophyll-a.

1. INTRODUCTION

Healthy inland freshwater sources, including lakes, reservoirs, rivers and streams, play a crucial role in providing and preserving benefits for the biodiversity and habitat of every living species in surrounding societies. In past few decades, driven by increasing population, global warming, urbanization, modernization and other factors, the water quality of lakes have fluctuated frequently and jeopardized our health (Wu et al., 2014). Long-term monitoring and real-time measurements of water quality, which are necessarily demands to provide immediate and proper treatments, aiming to identify significant changes of water sources for the avoidance of the unexpected environmental incidents (United States Environmental Protection Agency [US-EPA], 2016).

There are various domain experts devoting their efforts to aquatic environmental studies by collecting, analyzing water samples in the professional laboratory. The limitation of conventional approach is that the water quality could be represented only at sampling sites, but for extensive water bodies, the spatial-temporal synoptic view, which is important for long-term monitoring and managing, is impossibly obtained.

Over past decades, numerous previous studies have integrated various algorithms and regression models to obtain satisfactory results. Nevertheless, the selection of appropriate algorithm is complex and challenging because of the fact that the changes in chemical and physical properties of water can lead to different method determination

To alleviate the aforementioned difficulties, this study proposed a potential integration which comprises an optimization method for efficient water-quality model selection, ordinary least squares regression, and an accurately atmospheric corrected dataset.

2. STUDY AREA

Locating at approximately 60 km northeast of Tokyo, Lake Kasumigaura is the second largest lake with surface area of 160 km^2 in Japan. There is seasonal or year-to-year fluctuation in water level. The average and maximum depths are about 4 meters and 7 meters, respectively. In Japan, CGER (2010) indicated that the average Chlorophyll-a (Chla) in Kasumigaura Lake is declined from 87 to 61 mg m^{-3}.

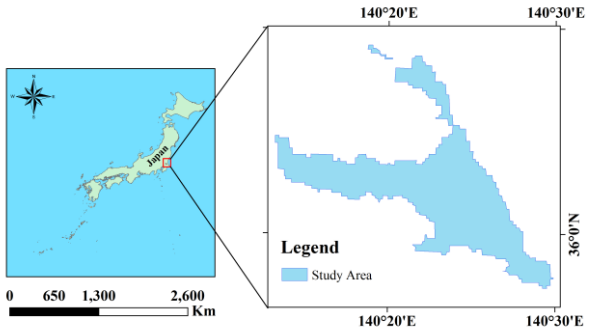

Figure 1: Study area

3. MATERIALS

The research material consisted of a datasets including in-situ measurements of Chla concentration as well as satellite imageries. Chla samples were collected during a field dedicated campaigns by University of Tsukuba and the National Institute for Environmental Studies (NIES) (20 samples on 18 May 2010). Beside the ground truth data on Chla, the European Space Agency (ESA) supplied the ENVISAT-MERIS satellite imageries, which were captured around water sampling time. During field campaigns and image acquisition periods, large amount of precipitation lead to higher water-level at Lake Kasumigaura.

3.1. In-situ Chorophyll-a Concentration

To integrate in-situ Chla measurements and remote sensing reflectance, denoted by $R_{rs}(\lambda)$, during feature selection as well as model calibration and validation stage, a collection of 20 water samples is collected to coincide with the acquisition time of satellite imagery. The sample locations are averagely distributed along the Lake Kasumigaura during field campaign on 18 May 2010.

The majority of in-situ Chla measurements are reported at very high concentrations, exceed 35 mg m^{-3}. The possible reason in the high Chla retrievals due to the water environments are getting warmer in the summer that lead to increase algal growth rates. Also, a significant amount of precipitation brings more nutrients runoff into the lake, feeding algal bloom . We also observed from the descriptive statistics that the Chla ranges are widely variation, 46.10 mg m^{-3} in 2010.

3.2. Remotely Sensed Data

In this study, the MERIS imagery with 300-meters spatial resolution were utilized to deliver the temporal-spatial data over Lake Kasumigaura. Despite ENVISAT's MERIS lost communication in 2012, the potential capability of collected dataset for extensive water quality has still experimented in recent studies as its unique bands, for instance, 510 nm, 665 nm, 681.25 nm, 708.75 nm and 753.75 nm bands which are not available in operating sensors (Zhang et al., 2019; Attila et al., 2018; Smith et al., 2018). By continuing to use the aforementioned bands, since 2016, a new sensor, which is developed based on MERIS, named Ocean and Land Color Instrument (OLCI), has been successfully placed in orbit, transmitting the 21 reflectance bands of the ocean and land, in which 15 bands are exactly what have been instrumented in ENVISAT-MERIS.

In the field of satellite-based water quality monitoring, an accurate atmospheric correction is an essential prerequisite for potentially improving the prediction of some key parameters concentrations in water bodies. Implication for future use as well as the limitations of atmospheric correction methods on Sentinel-3 datasets, we made use of MERIS bands which were precisely corrected using New- the standard Gordon and Wang algorithm with an iterative process and a bio-optical model (N-GWI) in this study (Jealani et al., 2014).

Jealani et al. (2014) proved that N-GWI outperformed the other four existing AC algorithms for turbid waters, including GWI (Stumpf et al., 2003; Bailey et al., 2010), Management Unit of the North Sea Mathematical Models (MUMM) (Li, 2003), Case-2 Water Processor (C2WP); (Doerffer and Schiller, 2007), and Self-Contained Atmospheric Parameters Estimation (SCAPE-M) (Guanter et al., 2007; Guanter et al., 2010).

4. METHODOLOGY

This part concentrates on application of an optimization method, named Neighborhood Component Feature Selection (NCFS) (Yang et al., 2012), for alleviating the difficulties in model selection. Feature selection is known as an efficient technique for selecting a small subset of significant features from a given high-dimensional features dataset. On the concept of the study, a feature is expressed as an algorithm, which is possibly created based on the optimization architecture of the existing algorithm. The following subsections have been prepared for further discussions about definition of feature, the principle of NCFS and the accuracy assessment.

4.1. Generation of Feature Space

In this study, feature space refers to the ***m***-dimensional dataset, of which a dimension corresponds to a variable that has been created as following steps. Firstly, considering the optimization architecture of the existing (semi-)empirical algorithms, included the original three-band models as Equation (1) (cite). Then, integrating multiple bands into architecture of the algorithm below to create a new variable. Due to the applied method named: "Neighborhood Component Feature Selection", hence we consider all variable as features.

$$[R_{rs}(\lambda_1)^{-1} - R_{rs}(\lambda_2)^{-1}].\, R_{rs}(\lambda_3) \qquad (1)$$

4.2. Standardization

Since the features vector x_i corresponding to the response y_i is created in different scales, the standardization plays an important pre-processing step to make the feature weights meaningful and comparable after feature selection, and also to make the optimization solver coverage faster than that without the standardization process. In this study, the feature space vector x_i is standardized to have zero mean and unit standard deviation.

4.2. Neighborhood Component Feature Selection (NCFS)

Given a training dataset $T=\{(x_1,y_1),\cdots(x_i,y_i),\cdots(x_n,y_n)\}$, where x_i represents a **m**-dimension features vector, and $y_i \in R$ is the corresponding response of x_i; and **n** denotes the number of training samples.

The NCFS method consists of three main steps. The first step begins with the standardization of the feature vectors x_i, which are created in different scales. The second step adopts leave-one-out (LOO) cross-validation to estimate the goodness of a defined loss function for sample *i* in the training dataset **T**. Afterwards, the

third step minimizes the summarized loss function to determine the optimal weights for every features.

Only directions along which the parameters contribute significantly to reducing the objective function are preserved relatively intact. In directions that do not contribute to reducing the objective function, components of the weight vector corresponding to such unimportant directions are decayed away through the use of the regularization throughout training.

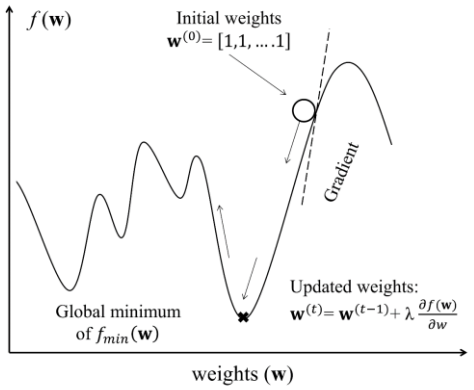

Figure 2: Minimizing the regularized loss function

4.3. Accuracy Assessment

There are three main criteria which were used in accuracy assessment, included root-mean-square-error (RMSE), Pearson's correlation coefficient (r) between predicted and measured Chla, slope (m) of the linear regression line between predicted and measured Chla also be calculated to visualize how close to 1:1 these regressed lines were.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\text{Chla}_{\text{pred},i}-\text{Chla}_{\text{meas},i})^2}{N}},$$

$$\text{r} = \frac{\sum_{i=1}^{N}(\text{Chla}_{\text{pred},i}-\overline{Chla}_{pred})(\text{Chla}_{\text{meas},i}-\overline{Chla}_{meas})}{\sqrt{\left(\sum_{i=1}^{N}(\text{Chla}_{\text{pred},i}-\overline{Chla}_{pred})^2\right)\left(\sum_{i=1}^{N}(\text{Chla}_{\text{meas},i}-\overline{Chla}_{meas})^2\right)}},$$

$$\text{m} = \frac{\sum_{i=1}^{N}(\text{Chla}_{\text{pred},i}-\overline{Chla}_{pred})(\text{Chla}_{\text{meas},i}-\overline{Chla}_{meas})}{\sum_{i=1}^{N}(\text{Chla}_{\text{pred},i}-\overline{Chla}_{pred})^2},$$

where $\text{Chla}_{\text{pred,i}}$, $\text{Chla}_{\text{meas,i}}$ were predicted and measured Chla of sample i; $\overline{\text{Chla}}_{\text{pred}}$, $\overline{\text{Chla}}_{\text{meas}}$ were average predicted Chla and average measured Chla of all validation samples, respectively; N indicated number of validation samples.

5. RESULTS AND DISCUSSIONS

5.1. NCFS Features Selection

The experimental flowchart has been implemented with consideration of the first 10 bands in MERIS images, following the steps: 1) a set of five-band groups is determined, each group consists of four spectral bands, of which have frequently dominated the previous studies, specifying at wavelengths of B_7 (665), B_8 (681), B_9 (709) and B_{10} (754), then, in sequential order, the fifth band from the remaining six bands is added to a group of four bands above; 2) a dataset of non-correlated 30 dimensions is created, which are based on the architecture of the formulas Eq. (1). After processing, the results pointed out the entire algorithms with its corresponding weights, as well as the errors of the regularized loss function during iterative runs.

Band	Wavelength centre (nm)	Band notation
1	412.5	$B_1(412)$
2	442.5	$B_2(442)$
3	490.0	$B_3(490)$
4	510.0	$B_4(510)$
5	560.0	$B_5(560)$
6	620.0	$B_6(620)$
7	665.0	$B_7(665)$
8	681.25	$B_8(681)$
9	708.75	$B_9(709)$
10	753.75	$B_{10}(754)$

Table 1. Band notation of the first 10 MERIS spectral bands

5.2. Algorithm accuracy assessment

In order to generate the predictive models from those aforementioned algorithms, the statistical calibration procedure, named the Ordinary least square regression (OLS) which is utilized. The coefficients of calibrated models along with its corresponding predictors, such as intercept a_0; slope a_1, a_2; the coefficient of determination (R^2) are reported in Table 3.

Model name	Predictors	
	X1	X2
Mdl1	$[R_{rs}^{-1}(665)$ - $R_{rs}^{-1}(709)].R_{rs}(681)$	$[R_{rs}^{-1}(665)$ - $R_{rs}^{-1}(709)].R_{rs}(490)$
Mdl2	$[R_{rs}^{-1}(665)$ - $R_{rs}^{-1}(709)].R_{rs}(681)$	$[R_{rs}^{-1}(665)$ - $R_{rs}^{-1}(709)].R_{rs}(510)$
Mdl3	$[R_{rs}^{-1}(709)$ - $R_{rs}^{-1}(560)].R_{rs}(681)$	**
Mdl4	$[R_{rs}^{-1}(709)$ - $R_{rs}^{-1}(620)].R_{rs}(681)$	**
3BG08 (Gitelson et al., 2008)	$[R_{rs}^{-1}(665)$ - $R_{rs}^{-1}(709)].R_{rs}(754)$	**

** Not available

Table 2. The predictor(s) of models.

Model name	a_0	a_1	a_2	R^2
Mdl1	-9.5	65	174	0.91
Mdl2	-8.9	1.7	198	0.91
Mdl3	108	-295	**	0.94
Mdl4	**48.6**	**-247.8**	**	**0.96**
3BG08	24.91	115.14	**	0.44

** Not available

Table 3. Coefficients derived from calibration using OLS method.

Models	RMSE (mg m^{-3})	*r*	m
Mdl1	6.3	0.91	1.015
Mdl2	6.4	0.91	1.021
Mdl3	6.4	0.91	1.031
Mdl4	**4.6**	**0.94**	**1.014**
3BG08	6.2	0.90	0.623

Table 4: Comparison of RMSE, slope (*m*) and correlation coeficient (*r*) and with other studies

The predictive capability of all calibrated models above were evaluated on an independent validation dataset consisting of remaining 10 samples in 2010. In order to identify the best fitted model, the comparison was concentrated on four criteria, such as root-mean-square-error (RMSE), the Pearson's correlation coefficient r and the slope m of the linear correlation line between measured and predicted Chla.

According to the Table 4, we clearly obtained that the algorithms which has selected from NCFS methods mostly provided higher accuracy to than the existing widely applied (semi-) empirical algorithm from previous studies. The performance of Mdl4 has reached the RMSE of 4.6. Moreover, the remaining criterias, such as Pearson' correlation coefficient (r) and slope (m) of the regressed line between predicted and the measured Chla, which also proved the outstanding predictive capability of Mdl4. The regressed line is very close to the line 1:1 when the slope is of 1.014, the coefficient (r) is of 0.94.

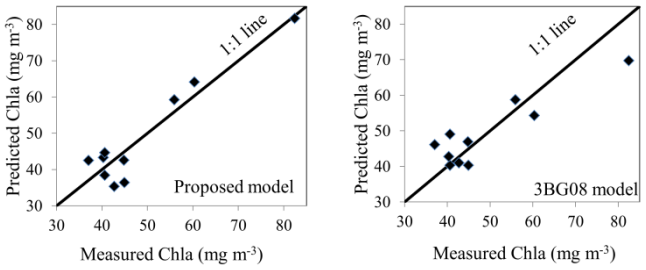

Figure 3. The comparison between measured and predicted Chla in proposed model (left) and reference model – 3BG08 (right)

6. CONCLUSIONS

From previous studies, NCFS has been proved as an effective and efficient method to select the significant features in many classification applications. Throughout this study, NCFS has been firstly experimented in regression, applied in the field of satellite-based water quality monitoring, it has showed the capability of selecting important features, which is express as algorithms to make accurate predictive models for estimation of Chla concentration.In case the widely applied algorithms from previous studie, such as 3BG08 that could not well perform in the present study area, NCFS could be used as an alternative approach to create new algorithms, aims to better fit with the study area. The selected algorithm, the predictive model in this study need to be further validated in different seasons, in this study, it has successful predicted the Chla concentration on unseen data of 2010. The selected algorithm has been found out using ENVISAT-MERIS data, which has stopped since 2012. However, Sentinel 3-MSI has instrumented the similar sensor to MERIS, operating in orbit, the selected algorithm could be further applied using Sentinel-3 remotely sensed data.

REFERENCES

Wu, J., Liu, W., Zeng, H., Ma, L. and Bai, R. (2014), "Water Quantity and Quality of Six Lakes in the Arid Xinjiang Region, NW China", Environmental Processes, Vol. 1 No. 2, pp. 115–125.
United States Environmental Protection Agency [US-EPA]. (2016), "Online source water quality monitoring for water quality surveillance and response systems", U.S.Environmental Protection Agency, p. 114.

Jaelani, L.M. (2014), Development of a New Atmospheric Correction Algorithm for Turbid Inland Waters, University of Tsukuba.

Zhang, F., Li, J., Shen, Q., Zhang, B., Tian, L., Ye, H., Wang, S. and Lu, Z. (2019), "A soft-classification-based chlorophyll-a estimation method using MERIS data in the highly turbid and eutrophic Taihu Lake", International Journal of Applied Earth Observation and Geoinformation, Vol. 74, pp. 138–149.

Attila, J., Kauppila, P., Kallio, K.Y., Alasalmi, H., Keto, V., Bruun, E. and Koponen, S. (2018), "Applicability of Earth Observation chlorophyll-a data in assessment of water status via MERIS — With implications for the use of OLCI sensors", Remote Sensing of Environment, Vol. 212 , pp. 273–287.

Smith, M.E., Robertson Lain, L. and Bernard, S. (2018), "An optimized Chlorophyll a switching algorithm for MERIS and OLCI in phytoplankton-dominated waters", Remote Sensing of Environment, Vol. 215, pp. 217–227.

Yang, W., Wang, K., and Zuo, W. (2012), Neighborhood Component Feature Selection for High-Dimensional Data. Journal of Computers, 7(1), 161-168.

Stumpf, R.P., Arnone, R.A., Gould, R.W., Martinolich, P.M. and Ransibrahmanakul, V. (2003), "A partially coupled ocean-atmosphere model for retrieval of water-leaving radiance from SeaWiFS in coastal waters", NASA Tech. Memo, SeaWiFS postlaunch technical report series, Vol. 206892, pp. 51–59.

Bailey, S.W., Franz, B.A. and Werdell, P.J. (2010), "Estimation of near-infrared water-leaving reflectance for satellite ocean color data processing", Optics Express, Vol. 18 No. 7, p. 7521.

Li, Y. (2003), "Atmospheric correction of SeaWiFS imagery for turbid coastal and inland waters: comment", Applied Optics, Vol. 42 No. 6, p. 893.

Doerffer, R. and Schiller, H. (2007), "The MERIS case 2 water algorithm", International Journal of Remote Sensing, Vol. 28 No. 3–4, pp. 517–535.

Guanter, L., González-Sanpedro, M.D.C. and Moreno, J. (2007), "A method for the atmospheric correction of ENVISAT/MERIS data over land targets", International Journal of Remote Sensing, Vol. 28 No. 3–4, pp. 709–728.

Guanter, L., Ruiz-Verdú, A., Odermatt, D., Giardino, C., Simis, S., Estellés, V., Heege, T., Domínguez-Gómez, J.A. and Moreno, J. (2010), "Atmospheric correction of ENVISAT/MERIS data over inland waters: Validation for European lakes", Remote Sensing of Environment, Vol. 114 No. 3, pp. 467–480.