Article

Retrieval of Chlorophyll-a and Total Suspended Solids Using Iterative Stepwise Elimination Partial Least Squares (ISE-PLS) Regression Based on Field Hyperspectral Measurements in Irrigation Ponds in Higashihiroshima, Japan

Zuomin Wang 1, Kensuke Kawamura 2, Yuji Sakuno 1,*, Xinyan Fan 3, Zhe Gong 3 and Jihyun Lim 3

1 Graduate School of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashihiroshima, Hiroshima 739-8527, Japan; wangzuomin123@gmail.com
2 Social Sciences Division, Japan International Research Center for Agricultural Sciences (JIRCAS), 1-1 Ohwashi, Tsukuba, Ibaraki 305-8686, Japan; kamuken@affrc.go.jp
3 Graduate School for International Development and Cooperation (IDEC), Hiroshima University, 1-5-1 Kagamiyama, Higashihiroshima, Hiroshima 739-8529, Japan; xinyanfan5160@gmail.com (X.F.); gongzhe79@gmail.com (Z.G.); limjihyun7@gmail.com (J.L.)

* Correspondence: sakuno@hiroshima-u.ac.jp; Tel.: +81-82-424-7773

Academic Editors: Yunlin Zhang, Claudia Giardino, Linhai Li, Deepak R. Mishra and Prasad S. Thenkabail

Received: 16 December 2016; Accepted: 9 March 2017; Published: 13 March 2017

Abstract: Concentrations of chlorophyll-a (Chl-a) and total suspended solids (TSS) are significant parameters used to assess water quality. The objective of this study is to establish a quantitative model for estimating the Chl-a and the TSS concentrations in irrigation ponds in Higashihiroshima, Japan, using field hyperspectral measurements and statistical analysis. Field experiments were conducted in six ponds and spectral readings for Chl-a and TSS were obtained from six field observations in 2014. For statistical approaches, we used two spectral indices, the ratio spectral index (RSI) and the normalized difference spectral index (NDSI), and a partial least squares (PLS) regression. The predictive abilities were compared using the coefficient of determination ($R^2$), the root mean squared error of cross validation (RMSECV) and the residual predictive deviation (RPD). Overall, iterative stepwise elimination based on PLS (ISE–PLS), using the first derivative reflectance (FDR), showed the best predictive accuracy, for both Chl-a ($R^2 = 0.98$, RMSECV = 6.15, RPD = 7.44) and TSS ($R^2 = 0.97$, RMSECV = 1.91, RPD = 6.64). The important wavebands for estimating Chl-a (16.97% of all wavebands) and TSS (8.38% of all wavebands) were selected by ISE–PLS from all 501 wavebands over the 400–900 nm range. These findings suggest that ISE–PLS based on field hyperspectral measurements can be used to estimate water Chl-a and TSS concentrations in irrigation ponds.

Keywords: chlorophyll-a; hyperspectral; irrigation ponds; partial least squares regression; total suspended solids

1. Introduction

Agriculture is by far the greatest water consumer in the world, and consequently, a major cause of water pollution. The primary pollutants from agriculture are excess nutrients and pesticides [1]. In agricultural activity, non-point source pollution, such as irrigation water and surface runoff water containing fertilizer from farmland, contributes to excessive nutrient concentrations [2]. Meanwhile, excess nutrients that cause eutrophication, hypoxia and algal blooms in surface water bodies and
coastal areas contribute to the primary global water quality problem [1]. Eutrophication has become a widespread matter of concern during the past 50 years, especially in coastal and inland waters [3].

The chlorophyll-\(a\) (Chl-\(a\)) concentration in water is the most widely applied parameter to assess the water quality status of lakes, particularly with respect to their trophic quality [4]. Since Chl-\(a\) is the primary photosynthetic pigment of all plant life [5], the concentration of Chl-\(a\) indicates phytoplankton biomass and eutrophication in lakes [6]. The concentration of total suspended solids (TSS) is another commonly used indicator for water quality assessment [7]. TSS consists of organic and inorganic materials suspended in the water [8]. Increased TSS decrease light transmission through the water [9], and therefore affect light availability to phytoplankton, thus resulting in a decrease of phytoplankton primary production [10].

However, traditional water quality monitoring requires in situ measurements and sampling, then returning the samples to the laboratory to measure water quality indicators (e.g., Chl-\(a\) and TSS), which is costly and time consuming [11]. Remote sensing makes it possible to monitor the state of the globe routinely, and is cost effective and useful, with the benefits of its passive nature and wide spatial coverage [12]. Earlier studies have demonstrated several algorithms developed for satellite sensors to estimate ocean and coastal water quality parameters, such as the Chl-\(a\) algorithm OC3, created for the moderate resolution imaging spectroradiometer (MODIS) data, and OC4, created for sea-viewing wide field-of-view sensor (SeaWiFS) data [13]. The geostationary ocean color imager (GOCI) also shows good performance, using the linear combination index (LCI) method to monitor Chl-\(a\) [14]. Further, a three-band semi-analytical reflectance model, originally developed by Gitelson et al. (2003) [15], and a normalized difference chlorophyll index (NDCI) [16], both performed well for assessing Chl-\(a\) in turbid productive water [16–18]. For estimating TSS concentrations, an algorithm with a single wavelength created for MODIS and medium spectral resolution imaging spectrometer (MERIS) data has been proved to be satisfactory [19].

Unlike ocean and coastal water, inland water usually has a smaller surface area and more complicated spectral features, especially irrigation ponds, which are often impacted by human use such as agriculture activities. Consequently, inland water quality monitoring presents higher requirements for both temporal and spatial resolution of satellite sensor data; hence currently used satellite sensors often have limited practical applicability in assessing relatively smaller inland water bodies. Since there are a limited number of wavebands for Landsat and other multispectral sensors, finding more informative wavebands to improve the performance of water quality estimation is necessary. With respect to in situ measurements, a two-band ratio approach, for example the ratio spectral index (RSI), has performed well for estimating Chl-\(a\) concentrations in inland waters [18,20,21], especially using the ratio of near-infrared (NIR) regions to red wavebands, such as the reflectance ratio of 705 nm to 670 nm performed by Han et al. (1997) [22]. Normalized difference spectral indices (NDSI) are another type of spectral indices frequently used to select the optimum bands for spectral analysis. As similar studies that have been done before mainly focused on vegetation parameters retrieval [23–25], optimum bands have been calculated from combinations of all available bands in the hyperspectral spectrum, a considerable range for hyperspectral analysis. Water quality parameters retrieval requires a similarly broad approach.

Partial least squares (PLS) regression, which was developed by Wold (1966) [26], is widely used to extract valuable information for spectroscopic analysis. PLS regression uses all available wavebands without multi-collinearity issues. The eigenvectors of the explanatory variables are manipulated such that the corresponding scores (latent variables) not only explain the variance of the explanatory variables (wavebands) themselves, but also are highly correlated with the response variables (Chl-\(a\) and TSS) [27]. However, PLS is considered limited because it treats each wavelength as independent, which incorporate noise created by non-informative wavelengths [28]. There is increasing evidence to indicate that wavelength selection can affect the performance of PLS analysis [29], since wavelength selection for PLS models is performed to eliminate uninformative variables and choose the variables that contribute the most to the predictive ability of the calibration model [30]. Iterative
stepwise elimination PLS (ISE–PLS), developed by Boggia et al. (1997) [31], combines PLS regression and the most useful information from hundreds of wavebands into the first several factors [32]. This method was developed to eliminate useless wavebands in PLS analysis.

The objective of this study is to develop models to estimate Chl-a and TSS using in situ spectral reflectance data and statistical approaches. We used several regression analyses including (a) a simple linear regression at each waveband of reflectance and the first derivative reflectance (FDR) to explore informative wavelength regions for Chl-a and TSS estimation; (b) all available two-band combination spectral indices (RSI and NDSI); and (c) a PLS regression using original reflectance and FDR datasets. In the PLS analyses, the predictive ability of ISE–PLS was compared with that of a standard full spectrum PLS (FS–PLS) and the spectral indices (RSI and NDSI).

2. Study Area

The study area is located in Higashihiroshima, Japan, as shown in Figure 1. Higashihiroshima is a core city in the central region of Hiroshima Prefecture, with a total area of 635.32 km² covering nearly 7.5% of the prefecture’s total area. Paddy fields, totalling 36.8 km², cover 14.9% of the Hiroshima Prefecture. Consequently, Higashihiroshima has the largest rice production of the 86 cities, towns and villages in Hiroshima Prefecture [33]. The city has an estimated population of 183,834 people, and its population density was 289.36 people per km² in 2011. The number of irrigation ponds in Hiroshima Prefecture approaches approximately 21,000. This qualifies as the second largest number in Japan; a quarter of the total irrigation ponds in Japan are in Higashihiroshima, the average beneficiary area is 3.36 ha, and the average number of beneficiary farmhouses is approximately 9 [34]. The monthly mean temperature ranges from 2.2 °C in January to 25.8 °C in August, and the monthly precipitation ranges from 43.3 mm in December to 232.1 mm in July, referring to the minimum and maximum values, respectively. To assess changes in water quality status and environments, six ponds, including both eutrophic ponds and non-eutrophic ponds, were selected for this study. Descriptions of the six ponds are listed in Table 1.

![Figure 1. Locations of Higashihiroshima and the six irrigation ponds used in this study.](image)

<table>
<thead>
<tr>
<th>No.</th>
<th>Name of Pond</th>
<th>Alt. (m)</th>
<th>Depth (m)</th>
<th>Area (ha)</th>
<th>Coordinate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nanatsu-ike</td>
<td>245</td>
<td>2.3</td>
<td>8.1</td>
<td>34°26′06.46″N 132°41′39.69″E</td>
</tr>
<tr>
<td>2</td>
<td>Shitami-Oike</td>
<td>221</td>
<td>1.5</td>
<td>2.5</td>
<td>34°24′28.56″N 132°42′22.09″E</td>
</tr>
<tr>
<td>3</td>
<td>Okuda-Oike</td>
<td>228</td>
<td>3.3</td>
<td>2.9</td>
<td>34°24′25.24″N 132°43′43.16″E</td>
</tr>
<tr>
<td>4</td>
<td>Yamanaka-ike</td>
<td>231</td>
<td>2.6</td>
<td>1.2</td>
<td>34°24′14.15″N 132°43′12.21″E</td>
</tr>
<tr>
<td>5</td>
<td>Yamanakaike-kamiike</td>
<td>231</td>
<td>1.1</td>
<td>0.1</td>
<td>34°24′15.29″N 132°43′14.45″E</td>
</tr>
<tr>
<td>6</td>
<td>Budou-ike</td>
<td>210</td>
<td>1.6</td>
<td>1</td>
<td>34°24′02.78″N 132°42′45.89″E</td>
</tr>
</tbody>
</table>
3. Materials and Methods

3.1. Measurement of Water Surface Reflectance

Measurements of water surface reflectance were performed using an ASD FieldSpec HandHeld-2 spectrometer (ASD Inc., Boulder, CO, USA) with a spectral range of 350–1050 nm and a probe field angle of 10°. Spectral readings were taken approximately 1 m above the water surface between 10:30 and 13:00 on a day with clear skies. Surveys were conducted six times between 3 January 2014, and 28 June 2014. From these data, a total of 36 datasets were obtained.

With respect to the spectral data, the ranges 325–399 nm and 901–1075 nm from each spectrum were identified as noise and removed. Subsequently, spectral data were smoothed using a moving and normalized Gaussian filter with a sigma (standard deviation) of 2.5. The FDR was also computed and compared with the original reflectance.

3.2. Water Sampling and Chemical Analysis

The water sampling sites were consistent with the spectral reflectance measurements. Immediately after measurement of spectral reflectance, water samples were collected into two 1 L containers. The samples were maintained at constant temperature and protected from light until they were received at the laboratory for analysis.

Chl-a and TSS concentrations were determined at the laboratory of the Graduate School for International Development and Cooperation (IDEC), Hiroshima University, Japan. Chl-a was extracted using 90% acetone, the absorption of Chl-a was measured by a spectrophotometer (UVmini-1240, SHIMADZU Co., Kyoto, Japan) and pigment concentration was calculated using the equations from UNESCO. To measure the TSS, the water sample was filtered using 47 mm diameter GF/F filters. The filters were weighed before and after drying with an oven drier (SANYO Electric Co., Moriguchi, Osaka, Japan) at 105 °C for two hours. The TSS contents were quantified by the difference in the weight of the filter paper before and after filtration.

3.3. Ratio Spectral Index and Normalized Difference Spectral Indices

A combination of spectral indices between all wavebands is performed to select the optimum two-band combination. The aim of spectral indices is to construct a mathematical combination of spectral wavebands to enhance information content with respect to the parameter under study [35]. Moreover, normalization in the NDSI is effective at cancelling atmospheric disturbance or other sources of error, while enhancing and standardizing the spectral response to the observed targets [23].

For this study, two of the most commonly used spectral indices (RSI and NDSI) were calculated using the reflectance dataset. The forms to express them are as follows:

\[
\text{RSI}(i, j) = \frac{R_i}{R_j} \quad (1)
\]

\[
\text{NDSI}(i, j) = \frac{R_i - R_j}{R_i + R_j} \quad (2)
\]

where \( R_i \) and \( R_j \) are the intensity values for bands \( i \) and \( j \), respectively. Using both indices, the optimum wavebands from all combinations of two separate wavelengths were obtained.

3.4. Full Spectrum Partial Least Squares Regression

We performed PLS regression to estimate Chl-a and TSS concentrations using the reflectance and FDR datasets \((n = 36)\). The standard FS–PLS regression equation is as follows:

\[
y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \epsilon \quad (3)
\]
where the response variable \( y \) is a vector of the water quality parameters (Chl-\( a \) and TSS), the predictor variables \( x_1 \) to \( x_i \) are surface reflectance or FDR values for spectral bands 1 to \( i \) (400, 401, \ldots , 900 nm), respectively, \( \beta_1 \) to \( \beta_i \) are the estimated weighted regression coefficients, and \( \varepsilon \) is the error vector. The latent variables were introduced to simplify the relationship between response variables and predictor variables. To determine the optimal number of latent variables (NLV), leave-one-out (LOO) cross validation was performed to avoid overfitting of the model, which was based on the minimum value of the root mean squared error (RMSECV). The RMSECV is calculated as follows:

\[
\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_p)^2}{n}}
\]

where \( y_i \) and \( y_p \) represent the measured and predicted water quality parameters (Chl-\( a \) and TSS) for sample \( i \), and \( n \) is the number of samples in the dataset (\( n = 36 \)).

3.5. Iterative Stepwise Elimination Partial Least Squares Regression

The ISE–PLS is a model-wise technique [31], which is based on the wavelengths selection function of the ISE method. To improve the performance of the PLS model, the optimum wavelengths with good predictive ability are selected for model calibration. The wavelengths elimination process depends on the importance of the predictors \( (z_i) \), described as follows:

\[
z_i = \frac{|\beta_i|s_i}{\sum_{i=1}^{I} |\beta_i|s_i}
\]

where \( \beta_i \) is the regression coefficient and \( s_i \) is the standard deviation of predictor, both corresponding to the predictor variable of the waveband \( i \).

Initially, all available wavebands (501 bands, 400–900 nm) are used to develop the PLS regression model. Then variables are ranked from most contributed to least contributed according to the predictor \( z_i \); in other words, the predictor \( z_i \) represents the weight of each variable. The least contributed variable is eliminated and the PLS model is recalibrated with the remaining predictor variables [36]. The model building procedure is repeated, and in each cycle the predictor variable with the minimum importance (i.e., the less informative wavelength) is eliminated, until the final variable is eliminated. To determine the optimum number of wavelengths to include in the final model, LOO cross validation is conducted after each calibration. The final model with the maximum predictive ability is calibrated by the minimum value of RMSECV [37].

3.6. Evaluation of Predictive Ability

The coefficient of determination (\( R^2 \)) and RMSECV were selected as indices to evaluate the FS–PLS and ISE–PLS calibration models’ accuracy by using LOO cross validation. High results for \( R^2 \) and low RMSECV indicate the best model to predict Chl-\( a \) and TSS concentrations. In addition, the residual predictive deviation (RPD) was used to evaluate the predictive ability of the models, which was defined as the ratio of standard deviation (SD) of reference data in prediction to RMSECV [38]. For determining the performance ability of the calibration models, the goal RPD was at least 3 for agriculture applications; RPD values between 2 and 3 indicate a model with good prediction ability, 1.5 < RPD < 2 is an intermediate model needing some improvement, and an RPD < 1.5 indicates that the model has poor prediction ability [39].

All data handling and linear regression analyses were performed using Matlab software ver. 8.6 (MathWorks, Sherborn, MA, USA).
4. Results

4.1. Chl-a and TSS Concentrations in Irrigation Ponds

Descriptive statistics are shown in Table 2, including the sampling data, the number of samples, the minimum (Min), the maximum (Max), the mean, the standard deviation (SD) and the coefficient of variation (CV). In total, 36 samples were collected from six irrigation ponds in six sets of field measurements (3 January, 19 January, 24 March, 9 April, 24 May, and 28 June in 2014). Field samples ($n = 36$) provided a wide range of both Chl-a (SD = 46.1 µg/L, CV = 2.0) and TSS (SD = 12.8 mg/L, CV = 1.65). In the datasets, Chl-a ranged from 0 to 169.5 µg/L, and TSS ranged from 0.1 to 53 mg/L, which indicates that this study involves various water quality conditions from different ponds.

Table 2. Descriptive statistics for the Chl-a and TSS concentrations.

<table>
<thead>
<tr>
<th>Date</th>
<th>n</th>
<th>Chl-a (µg/L)</th>
<th>TSS (mg/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>3 January 2014</td>
<td>6</td>
<td>0.1</td>
<td>98.7</td>
</tr>
<tr>
<td>19 January 2014</td>
<td>6</td>
<td>0.1</td>
<td>169.5</td>
</tr>
<tr>
<td>24 March 2014</td>
<td>6</td>
<td>0</td>
<td>169.1</td>
</tr>
<tr>
<td>9 April 2014</td>
<td>6</td>
<td>0.5</td>
<td>48.5</td>
</tr>
<tr>
<td>24 May 2014</td>
<td>6</td>
<td>0.9</td>
<td>37.7</td>
</tr>
<tr>
<td>28 June 2014</td>
<td>6</td>
<td>1.6</td>
<td>133.9</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>0</td>
<td>169.5</td>
</tr>
</tbody>
</table>

SD = standard deviation; CV = coefficient of variation; $n$ = number of samples.

4.2. Comparison of Simple Linear Regression Models

In this study, several simple linear regression models were constructed, and the accuracy was compared with that of the PLS method. As shown in Table 3, distinct bands were selected as the optimal bands with respect to accuracy for all models. In the model that used the gaussian smoothed water surface reflectance and FDR, the 730 nm and 705 nm wavebands were selected, based on the linear correlation coefficient shown in Figure 2, to estimate Chl-a concentration ($R^2 = 0.14$ and 0.54); 722 nm and 704 nm were selected to estimate TSS ($R^2 = 0.05$ and 0.46). Figure 2 shows the correlation coefficient $r$ between reflectance/FDR and Chl-a/TSS with regard to each waveband. It is clear that FDR obviously improved correlation with Chl-a and TSS; moreover, spectra reflectance and absorption features were also enhanced (Figure 2b). An NIR/red algorithm developed by Han et al. (1997) [22] was introduced for comparison of the RSI selected wavebands and accuracy. The NIR/red model showed a higher $R^2$ and lower RMSE than the single waveband models. However, based on the regression between the reflectance of each waveband and Chl-a and TSS, the RSI model selected the $R_{719}/R_{662}$ ratio as the best band combination, which enhanced the performance of ratio model, giving the highest $R^2$ value of 0.72 for Chl-a. The $R_{717}/R_{630}$ ratio was the best band combination for TSS, with an $R^2$ of 0.52 (Figure 3a,b). A three-band semi-analytical algorithm for estimating Chl-a concentration was conducted, as a previous study suggested [40], and the optimal wavebands of model were tuned according to the optical properties of the water bodies. Bands 660, 703, and 740 nm were final selected for the three-band model with an $R^2$ of 0.71 and RMSE of 29.32. For another algorithm introduced in a previous study, the NDCI was evaluated using remote sensing reflectance $R_{rs}$ at an absorption peak of 665 nm ($R_{rs665}$), which is closely related to absorption by Chl-a pigments and a reflectance peak of 708 nm ($R_{rs708}$), which was sensitive to variations in Chl-a concentration in water, with a result of an $R^2$ of 0.60 and an RMSE of 28.82. For the NDSI model, bands 719 and 663 nm were the best combination for estimating Chl-a ($R^2 = 0.64$), and bands 704 and 698 nm were the best combination for TSS ($R^2 = 0.59$) (Figure 3c,d). The results showed the lowest RMSECV in the RSI model for Chl-a (24.14) and in the NDSI model for TSS (8.48). Among the models, the RSI or the NDSI showed
higher $R^2$ values and lower RMSECV values than those of the two types of single-band models in the estimation of both Chl-$a$ and TSS.

Table 3. Regression models used to estimate Chl-$a$ and TSS concentrations with two spectral data types (reflectance and FDR) and two spectral indices (RSI and NDSI).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Spectral Index</th>
<th>Model</th>
<th>$R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chl-$a$</td>
<td>Reflectance</td>
<td>Chl-$a$ = 0.0004 × $R_{720}$ + 0.0396</td>
<td>0.14</td>
<td>51.00</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>Chl-$a$ = $1 \times 10^{-5} \times R_{735}$ - 0.0004</td>
<td>0.54</td>
<td>51.01</td>
</tr>
<tr>
<td></td>
<td>NIR/red (Han et al. (1997) [22])</td>
<td>Chl-$a$ = 94.748 × $R_{719}/R_{670}$ - 88.897</td>
<td>0.60</td>
<td>28.78</td>
</tr>
<tr>
<td></td>
<td>Three-band (Gitelson et al. (2003) [15])</td>
<td>Chl-$a$ = 0.00066 × ($R_{705} - 40$) × $R_{740}$ - 0.0665</td>
<td>0.71</td>
<td>29.22</td>
</tr>
<tr>
<td></td>
<td>NDCI (Mishra et al. (2012) [16])</td>
<td>Chl-$a$ = 253.16 × ($R_{708} - R_{665}$)/($R_{708} + R_{665}$) + 36.353</td>
<td>0.60</td>
<td>28.82</td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>Chl-$a$ = 119.27 × $R_{719}/R_{662}$ - 88.052</td>
<td>0.72</td>
<td>24.14</td>
</tr>
<tr>
<td></td>
<td>NDSI</td>
<td>Chl-$a$ = 253.16 × ($R_{719} - R_{663}$)/($R_{719} + R_{663}$) + 36.353</td>
<td>0.64</td>
<td>27.19</td>
</tr>
<tr>
<td>TSS</td>
<td>Reflectance</td>
<td>TSS = 0.0009 × $R_{722}$ + 0.0501</td>
<td>0.05</td>
<td>14.81</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>TSS = $5 \times 10^{-5} \times R_{734}$ - 0.0003</td>
<td>0.46</td>
<td>14.83</td>
</tr>
<tr>
<td></td>
<td>RSI</td>
<td>TSS = $31.419 \times R_{717}/R_{663} - 17.913$</td>
<td>0.52</td>
<td>8.73</td>
</tr>
<tr>
<td></td>
<td>NDSI</td>
<td>TSS = 300.45 × ($R_{704} - R_{668} + R_{668} + 6.3668$)</td>
<td>0.55</td>
<td>8.48</td>
</tr>
</tbody>
</table>

Figure 2. Correlation coefficients ($r$) between water quality parameters (Chl-$a$ and TSS) at each wavelength: (a) reflectance; (b) FDR.

Figure 3. Distributions of $R^2$ between two wavebands using RSI (a) Chl-$a$; (b) TSS and NDSI (c) Chl-$a$; (d) TSS.
4.3. FS–PLS and ISE–PLS Models

Calibration and cross validation results between reflectance/FDR spectra and Chl-a/TSS using FS–PLS and ISE–PLS are shown in Table 4. The results showed that the optimum NLV ranged between 4 and 8 in FS–PLS and between 5 and 11 in ISE–PLS, which was determined by the LOO cross validation based on the lowest RMSEC. The RPD ranged between 1.22 and 1.32 (low accuracy) in FS–PLS and between 1.45 and 7.44 (excellent accuracy) in ISE–PLS. In particular, the selected number of wavebands and the percentage to full spectrum (that is, selected wavebands number/all \((n = 501) \times 100\%\)) were calculated to evaluate the informative wavebands for ISE–PLS. Results showed the selected wavebands number ranged between 9 and 85, and the percent ratio ranged between 1.80 and 16.97. Overall, for Chl-a, ISE–PLS using FDR showed the highest \(R^2\), highest RPD, and lowest RMSEC \((R^2 = 0.98, \text{RMSEC} = 6.15, \text{RPD} = 7.44)\); NLV = 11, and 85 wavebands were selected. Similarly, with respect to TSS, ISE–PLS using FDR showed the highest \(R^2\), highest RPD and lowest RMSEC \((R^2 = 0.97, \text{RMSEC} = 1.91, \text{RPD} = 6.64)\); NLV = 11, and 42 wavebands were selected.

The relations between observed and predicted Chl-a and TSS are shown in Figure 4. The data in this figure were used to evaluate goodness of fit in the FS–PLS and ISE–PLS models. Comparisons between the FS–PLS and ISE–PLS models were presented in combination with the \(R^2\) and RMSE from the cross validation listed in Table 4. For Chl-a, the ISE–PLS using FDR showed a higher \(R^2\) and lower RMSEC. The scatter distribution also showed a better linear relation, which can be judged by the red dots clustered along the 1:1 line in Figure 4b. Similarly, with respect to TSS, the ISE–PLS model using FDR showed better results than the others (Figure 4d, red dot). However, both red and green dots clustered vertically, particularly in Figure 4a,c, showing a large variation in the predicted values and nearly no variation in the observed values, indicating that plenty of observed Chl-a and TSS samples had low concentrations. This vertical clustering also indicates the FS–PLS and ISE–PLS using reflectance had lower predictive abilities than using FDR.

The selected wavebands in ISE–PLS using the reflectance and FDR datasets are shown in Figure 5. In the reflectance dataset, the selected wavebands were primarily in the red wavelengths (650–680 nm) for Chl-a. For TSS, the selected wavebands were in green wavelengths (560 nm), red wavelengths (620–630 nm) and red-edge wavelengths (720 nm). In the FDR datasets, a cluster of wavebands focus on the red region (670–680 nm, 690–710 nm) for Chl-a, and wavebands were also selected from other regions: blue (around 410), green (around 490 nm, 510 nm), red (around 603, 615), and the NIR region between 820 nm and 900 nm. Similarly, more wavebands were selected for TSS using FDR than using reflectance, especially in the red (around 620 nm, 680 nm, and 700 nm) and NIR (around 730 nm) regions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Spectral Data Type</th>
<th>Regression</th>
<th>Calibration</th>
<th>Cross Validation</th>
<th>Selected Wavebands Number</th>
<th>Selected Wavebands (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>NLV</td>
<td>(R^2)</td>
<td>RMSEC</td>
<td>(R^2)</td>
</tr>
<tr>
<td>Chl-a</td>
<td>Reflectance</td>
<td>FS–PLS</td>
<td>4</td>
<td>0.59</td>
<td>29.26</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>Reflectance</td>
<td>ISE–PLS</td>
<td>6</td>
<td>0.70</td>
<td>25.01</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>FS–PLS</td>
<td>8</td>
<td>0.99</td>
<td>3.25</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>ISE–PLS</td>
<td>11</td>
<td>1.14</td>
<td>1.07</td>
<td>0.98</td>
</tr>
<tr>
<td>TSS</td>
<td>Reflectance</td>
<td>FS–PLS</td>
<td>6</td>
<td>0.61</td>
<td>7.87</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>Reflectance</td>
<td>ISE–PLS</td>
<td>5</td>
<td>0.62</td>
<td>7.76</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>FS–PLS</td>
<td>5</td>
<td>0.93</td>
<td>3.39</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>FDR</td>
<td>ISE–PLS</td>
<td>11</td>
<td>1</td>
<td>0.84</td>
<td>0.97</td>
</tr>
</tbody>
</table>

FDR = first derivative reflectance; NLV = number of latent variables; RMSEC = root mean square error from calibration; RMSECV = root mean square error from cross validation; RPD = the residual predictive deviation.
Figure 4. Relations between measured and cross-validated prediction values of Chl-α (a) Reflectance; (b) FDR and TSS; (c) Reflectance; (d) FDR using FS–PLS and ISE–PLS.

Figure 5. Selected wavebands in ISE–PLS using reflectance or FDR datasets (n = 36) to estimate: (a) and (c) Chl-α; (b) and (d) TSS. Green bars = Chl-α; red bars = TSS.
5. Discussion

5.1. Evaluation of the Predictive Abilities of Simple Linear Regression Models

In the present study, models established by single waveband and two waveband combinations were compared using PLS. For single waveband models, FDR showed a better $R^2$ than smoothed reflectance both for Chl-$a$ and TSS, indicating that the accuracy can be improved by enhancing the features of absorption and reflectance from the smoothed reflectance. However, all single waveband models showed poor accuracy for estimating both Chl-$a$ and TSS concentrations. According to previous research, single band focus on 670–750 nm is better at determining TSS concentrations [19], especially in turbid water. Single band focus showed no predictive ability in that research, and simple linear regression using two wavebands combinations showed poor accuracy for TSS, which may indicate that TSS is difficult to detect using single-band or two-band combinations in relatively clear water; as shown in our results, most observed TSS values were low. The three-band model was successfully used for Chl-$a$ retrieval in turbid water bodies [18,40]. As for this research, the optimal spectral bands selected from the iterative band tuning are in accord with the previous research [17]; however, even the result shows a considerable $R^2$, but the relatively high RMSE may indicate a low accuracy model, which may be attributed to different compositions of optically active constituents (Chl-$a$, tripton, CDOM) [40]. The NDCI is a special case of the NDSI: two bands of NDCI are determined by the reflectance peak and spectral absorption peak, and the normalizing of two bands reflectance can eliminate uncertainties in the estimation of $R_{rs}$ [16]. As a comparison, the result of the NDSI has a slight improvement with an $R^2$ of 0.64 and an RMSE of 27.19 than the NDCI with an $R^2$ of 0.60 and an RMSE of 28.82, which may indicate that a combination of wavebands at 719 and 663 nm in the NDSI can better reflect the Chl-$a$ variations in this research area. Among all tested combinations of the RSI and the NDSI, the best $R^2$ values were obtained using the NIR waveband (719 nm) and the red region (662 nm for the RSI, 663 nm for the NDSI) to estimate Chl-$a$ concentrations, which agrees with the findings of other research. In most available research on the measurement of chlorophyll content in water, the absorption trough is located at near 670 nm, caused by absorption of Chl-$a$ [22,41] and the reflectance peak near 710 nm, caused by the fluorescence of Chl-$a$ [42–44]. On account of these characteristics, the two waveband models, particularly the NIR/red ratio, have been widely used for Chl-$a$ retrieval, and a variety of algorithms have been based mainly on the ratio of reflectance peak (about 710 nm) to reflectance trough (about 670 nm) [22,45]. Similarly, in the present study two wavebands, from the NIR and red regions respectively, were selected by the NDSI, confirming the water body reflection characteristics.

5.2. Evaluation of the Predictive Abilities of FS–PLS and ISE–PLS

As we expected, the PLS models exhibited better predictive abilities than models that use single wavebands or the index-based (RSI and NDSI) approaches, which shows the PLS method is potentially useful in retrieval of inland water quality parameters [46,47]. In our PLS analyses, results using ISE–PLS models with the FDR dataset showed higher $R^2$ and lower RMSECV values than those of the reflectance dataset. These results are consistent with the research of Han and Rundquist (1997) [22], who noted that FDR was better correlated with chlorophyll concentration than raw reflectance, and that random noise and the effects of suspended matter could be reduced by FDR [46]. After eliminating outliers and useless predictors, ISE–PLS calibrated more potential models than FS–PLS, both for Chl-$a$ and TSS, with the wavelengths relevant to water quality. As a consequence, predictive ability was further enhanced, which is reflected in the results of evaluation indices. PLS-based waveband selection greatly improved predictions for both Chl-$a$ ($R^2$ from 0.43 to 0.98, RMSECV from 35.15 to 6.15, RPD from 1.32 to 7.44) and TSS ($R^2$ from 0.40 to 0.97, RMSECV from 9.98 to 1.91, RPD from 1.27 to 6.64). The PLS models in combination with wavelength selection had an improved performance also supported by other previous research [29,36,48]. However, the $R^2$ for Chl-$a$ using ISE–PLS reached 0.98, a result that does not rule out the possibility of overfitting; therefore, the solution method for this condition should be the subject of additional research and validation.
5.3. Importance of Selected Wavebands in ISE–PLS

Our results showed 16.97% of all available wavelengths that were selected for predicting Chl-a and 8.38% were also selected for predicting TSS by ISE–PLS, which indicates that less than 20% of the waveband information from field hyperspectral data contributes to the prediction for water quality parameters (Chl-a and TSS) and over 80% were redundant. In the reflectance dataset, wavebands primarily in the red wavelengths were selected: between 630 and 710 nm for Chl-a; for TSS, 560 nm, 620–630 nm, and 720 nm. In the FDR dataset, the selected wavebands for estimating both Chl-a and TSS involved more regions than the reflectance dataset. Nevertheless, similar wavelengths in the visible and NIR regions were selected; blue (410 nm), green (approximately 490 nm, 510 nm), and red (approximately 603 nm, 615 nm) for Chl-a; and blue (approximately 420 nm), green (approximately 490 nm, 510 nm), red (approximately 620 nm, 680 nm and 700 nm), and NIR (approximately 730 nm) for TSS. Intensive absorption by Chl-a resulted in reflectance troughs around 440 and 670nm (Figure 5a) [49]. Low absorption of algal pigments or the scattering of phytoplankton cells and inorganic suspended materials might cause the reflectance peak near 570 nm [41]. The reflectance spectrum peak near 700 nm had a strong correlation with Chl-a concentration [42,50,51]. Several previous studies of inland water quality also proved these wavelengths have the potential to predict Chl-a and TSS concentrations [52–54]. This study brings obvious evidence that the ISE-PLS model may be considered as a unified approach for remote quantification of constituent concentrations in water quality assessment. Using this method, more informative wavebands can be selected from hundreds of hyperspectral wavebands, which indicates the accuracy and efficiency can be enhanced by ISE-PLS when it comes to using hyperspectral sensors in satellites with a high temporal and spatial resolution to monitor relatively small area inland water quality in the future.

6. Conclusions

The present study develops models for estimating Chl-a and TSS concentrations in irrigation ponds using water surface reflectance spectral data. Our results show that PLS regression analysis has high potential for predicting Chl-a and TSS based on field hyperspectral measurements, and that ISE wavebands selection in combination with PLS regression analysis can enhance predictive ability. Chl-a and TSS concentrations were estimated with high accuracy by using ISE-PLS, which explains 98% of the variance for Chl-a and 97% of the variance for TSS. The important wavebands for estimating Chl-a and TSS using ISE–PLS represented 16.97% and 8.38%, respectively, of all 501 wavebands over the 400–900 nm range. The selected wavebands approximately match the absorption peaks published by previous researchers. Compared to the estimation of water quality parameters by satellite sensors such as MODIS, ISE–PLS selected more informative wavebands, especially the wavelength at approximately 700 nm. These results provide useful insights for future analyses on the assessment of water quality in irrigation ponds, especially when using satellite imagery.

Acknowledgments: This study was supported by the Environmental Research and Technology Development Fund (S9) of the Ministry of the Environment, Japan, and JSPS KAKENHI (24560623, 15K14041, 16H05631).

Author Contributions: Yuji Sakuno and Kensuke Kawamura designed this study and the fieldwork; Xinyan Fan, Zhe Gong, Jihyun Lim, and Zuomin Wang performed the fieldwork; Zuomin Wang carried out the laboratory analysis, analyzed the data and wrote the manuscript; Zuomin Wang, Kensuke Kawamura and Yuji Sakuno revised the paper.

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

References


42. Gitelson, A.A. The peak near 700 nm on radiance spectra of algae and water: Relationships of its magnitude and position with chlorophyll concentration. Int. J. Remote Sens. 1992, 13, 3367–3373. [CrossRef]


© 2017 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/).