ANALYSIS AND MODEL DEVELOPMENT OF DIRECT HYPERSPECTRAL CHLOROPHYLL-A ESTIMATION FOR REMOTE SENSING SATELLITES

Sivert Bakken[1], Geir Johnsen[2], Tor A. Johansen [1]

Center for Autonomous Marine Operations and Systems
Department of Engineering Cybernetics(1) &
Department of Biology(2)
Norwegian University of Science and Technology (NTNU)
Trondheim, Norway

ABSTRACT

The advancement of instruments makes processing of hyperspectral data for monitoring phytoplankton dynamics more viable. Methods used operationally for retrieval of chl-a, an important indicator of phytoplankton, are developed for multispectral systems and optically deep waters. Coastal waters are important for the aquaculture industry, marine science, and environmental monitoring. Data rate limitations make the use of sensors with high spectral resolution difficult. Here, estimation of chl-a from top of the atmosphere reflectance using "Partial Least Squares"- and "Least Absolute Shrinkage and Selection Operator"-regression is compared with the internal consistency of the OC4 algorithm by NASA Ocean Biology Processing Group.

The models perform better in terms of NRMSE and R^2 when validated with a subset of the total data and with a separate scene. This is demonstrated by using experimental hyperspectral scenes from the Hyperspectral Imager for the Coastal Ocean (HICO) mission, processed through SeaDAS.

***Index Terms*—** Hyperpsectral Imaging, Remote Sensing, Chlorophyll-a Concentration, Ocean Color

1. INTRODUCTION

The study of Ocean Color has many potential societal benefits [1]. Chl-a concentration monitoring can provide the aquaculture industry and the government information regarding water quality, biogeochemical cycles, and fisheries management. For research, the chl-a concentration provides a bio-marker for the state of the marine ecosystems, as well as an aid to modeling the ocean state and monitoring climate change.

With traditional band-ratio algorithms for chl-a estimation, like OC4, it has been shown that additional spectral information improves the results [2]. Successful application of band-ratio algorithms in optically complex waters is often challenging due to overlapping of spectral signals from phytoplankton (chl-a), Colored Dissolved Organic Matter (CDOM), and Total Suspended Matter (TSM) that confound the model [3].

Nonlinear machine learning methods, e.g. Neural Networks (NN) or kernel-based regression models such as Gaussian Process Regression (GPR) and Support Vector Machines (SVM), can give good chl-a estimations, but will also have a higher level of complexity [4]. The relative relevance of the input features is less transparent, it is more challenging to foresee the model behavior by theoretical analysis, and the models themselves will be computationally more demanding to use for estimation when compared to linear prediction models [3, 4, 5]. However, for any linear model, nonlinear patterns can be compensated for by the appropriate preprocessing or by kernel functions [6].

Partial Least Squares Regression (PLSR) has been demonstrated to perform with greater accuracy than optimized band ratio algorithms when predicting chl-a concentrations with field-retrieved hyperspectral water-leaving reflectance [3]. Hyperspectral water-leaving reflectance values are dependent on an accurate atmospheric correction[2, 7]. This can be difficult to achieve in coastal regions, and hyperspectral atmospheric correction for ocean color is an active area of research [7, 8, 2].

In this paper, different approaches to perform regression analysis and model development using top-of-the-atmosphere reflectance values from the hyperspectral image data from the HICO mission are used to estimate the chl-a concentration. The chl-a concentration is computed by SeaDAS. The results from the regression are then compared to the OC4 band-ratio algorithm.With the demonstrated approach the PLSR models developed are intuitively interpretive, less computationally demanding, and generate promising results. Shown with different designs for validation.

2. HICO DATA AND SEADAS

The Hyperspectral Imager for the Coastal Ocean (HICO) mission was a hyperspectral instrument onboard the International

Space Station capable of capturing scenes with 128 different wavelengths in a range from 350 to 1080 nm at a 5 nm resolution [9].

Here, hyperspectral scenes from the HICO mission, processed and quality controlled through NASA OBPG software SeaDAS, are used as the desired values for chl-a. The reflectance data used is derived from the standard atmospheric correction provided by SeaDAS for HICO.

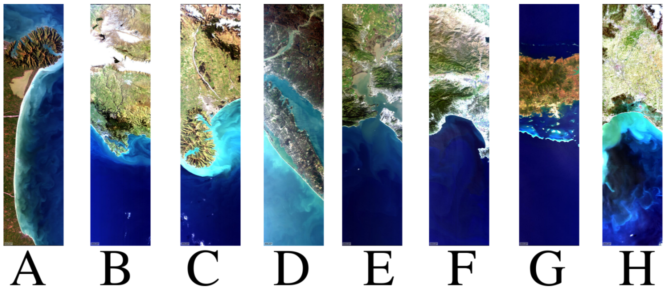

Fig. 1: HICO Sample Image Gallery Scenes from different locations around the world used for training and validation.

The scenes were selected due to the low adverse effects from atmospheric interference and good overall imaging quality from the HICO Sample Image Gallery [10]. The scenes can be seen in figure 1 and their locations are, from the left, southeast coast of New Zealand (A-C), US west coast (D-F), New Caledonia, and Italy.

The chl-a concentrations in mg/m^{-3} for the training and the verification data sets are displayed in Figure 2. The HICO chl-a data used as ground truth are derived through SeaDAS which, as standard, used the OC4 algorithm with MERIS coefficients and wavelengths [11]. In SeaDAS, the hyperspectral data is subsampled to the MERIS bandwidths when deriving chl-a. From each scene, 700 sample spectra are taken for training, and 300 different spectra are taken for verification in the initial validation procedure. In the secondary validation procedure, the fourth scene from the left, scene D, in Figure 1 is kept out and only used as a validation data set, as it has a good dynamic range. See section 4 for details about the validation procedures. The chl-a concentration for each sample can be seen in Figure 2 for the first validation procedure.

3. METHODS

A conceptual description of the different algorithms and their advantages and shortcomings are presented. A complete derivation of the algorithms is beyond the scope of this paper, but references are provided.

The preprocessing of the input variables in this paper incorporates known physical relationships [5]. Only the radiance signals from 400 to 900 nm are used, which leaves 87 spectral bands. The wavelengths outside this range have been reported as noisy [9]. No preprocessing in terms of dimensionality reduction has been performed.

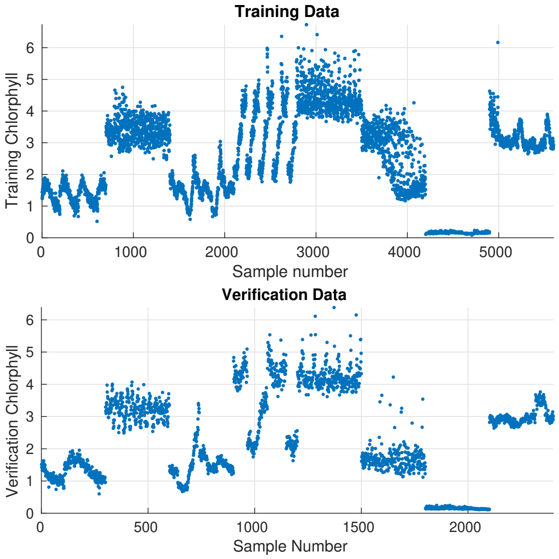

Fig. 2: Chl-a concentrations for the sampled data

Top-of-atmosphere reflectance $\rho(\lambda)$, can be defined at a given wavelength λ, to be related to the radiance $L(\lambda)$, the extraterrestrial solar irradiance $F_0(\lambda)$, and the solar-zenith angle θ_0, as given in equation (1) [7].

$$\rho(\lambda) = \pi L(\lambda)/(F_0(\lambda)\cos(\theta_0)) \tag{1}$$

First, with the assumption that the variations in $F_0(\lambda)$ are small compared to the sun angle effects, the reflectance values are approximated as $\hat{\rho}(\lambda)$ by dividing the radiance data with the solar zenith angle.

Secondly, the effect of attenuation of light is accounted for by computing the log values of the approximated reflectances as $\tilde{\rho}(\lambda) = \log_{10}(\hat{\rho}(\lambda))$[12].

Finally, the input variables have been centered and scaled to adhere to different absolute values and variation for a given wavelength [5]. This is given in equation (2), where $\hat{x}$ is the new variable, $\tilde{\rho}(\lambda)$ is the old, $\bar{\rho}(\lambda)$ is the mean, and σ_ρ is the standard deviation.

$$\hat{x} = \frac{\tilde{\rho}(\lambda) - \bar{\rho}(\lambda)}{\sigma_\rho} \tag{2}$$

3.1. OC4 by NASA OBPG

The OC4 algorithm developed by NASA OBPG [11], presented in equation (3), returns the near-surface concentration of chl-a in mg/m^{-3}. The algorithm uses an empirical relationship derived from in situ measurements of chl-a concentration and corresponding above-water remote sensing reflectances R_{rs}, with 4 spectral bands [2]. In this paper, the implementation of the OC4 algorithm uses the spectral bands closest to the ones used by the SeaWiFS multispectral imager[13, 2]. $R_{rs}(\lambda_{green})$ is the band closest to 555 nm, and $R_{rs}(\lambda_{blue})$ is the maximum of the bands closest to 443, 490, or 510 nm. The a_i coefficients used in the implementation of OC4 presented here were found using ordinary least squares on the training data presented in Figure 2 [5].

$$log_{10}(chl\_a) = \sum_{i=0}^{4} a_i \left(log_{10} \left(\frac{R_{rs}(\lambda_{blue})}{R_{rs}(\lambda_{green})} \right) \right)^i \qquad (3)$$

3.2. Partial Least Squares

Partial least-squares regression (PLSR) iteratively relates data matrices using linear multivariate models that reduce collinearity and noise within a given dataset. It is a two-step algorithm that first finds uncorrelated components in the variables of a given data set and then performs the least squares regression on these components. A more in-depth description of the algorithm can be found in [14].

This generates models with high levels of interpretability, but in their simplest form cannot accommodate for strong nonlinear effects[5].

3.3. Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) regression [15] was performed using the approach found in equation (4), where y_i is chl-a values, β_i is the regression coefficients, $\mathbf{X}$ is a matrix with all the pre-processed data, and t is a threshold that was iterated over with 10-fold cross-validation of the training data. The threshold that gave the lowest mean square error was selected.

This also generates models with high levels of interpretability, but in their simplest form cannot accommodate strong nonlinear effects. This method does not seek to find any covariation between variables. A vector of all ones is represented as $\mathbf{1}_N$.

$$\min_{\beta_0,\beta} \left\{ \|(\mathbf{y}_i - \beta_0 \mathbf{1}_N - \mathbf{X}\beta\|_2^2 \right\} \qquad s.t. \quad \|\beta\|_1 \le t \qquad (4)$$

4. RESULTS & DISCUSSION

Presented in this section is a discussion comparing the results in terms of regression coefficients, Normalized Root Mean Square (NRMSE) and R-squared (R^2) values for OC4, PLSR and LASSO.

All models were tested with the verification data set shown in Figure 2, as well as separating out the fourth scene from the left in Figure 1, scene D as the scene inhibits a good dynamic range in terms of chl-a.

4.1. OC4 Algorithm

The a_i coefficients used in equation (3) for the OC4 is computed by taking the least-squares fit of the training data set. This should give OC4 the best possible starting point.

As can be seen in Figure 4 the OC4 algorithm performs similar to previously reported tests [2, 11]. With validation using a subset of the total data i.e. the verification data set no clear trend in the residuals can be found. The algorithm struggles to determine chl-a $> 4mg/m^3$, with the coefficients found.

For the validation with a separated scene, scene D in Figure 1, the OC4 algorithm can determine a qualitative chl-a concentration within a given scene from SeaDAS, but is not able to quantify it accurately.

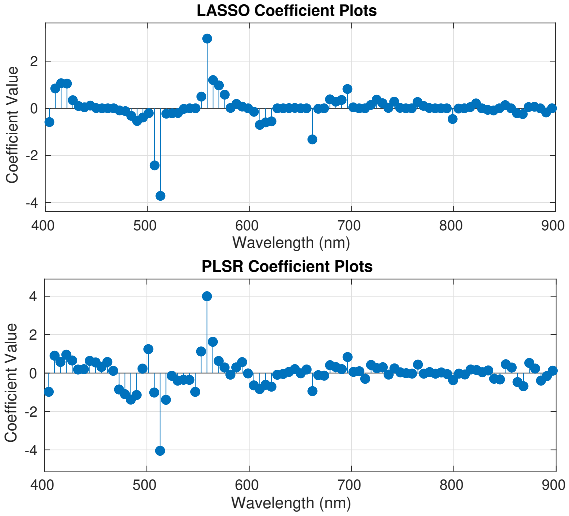

Fig. 3: Regression coefficients from PLSR and LASSO

4.2. Regression Models

LASSO and PLSR are both linear regression models that provide interpretable coefficients, as can be seen in Figure 3. The models developed deployed 10-fold cross-validation with the training data set when determining the weights of the regression. For the PLSR the average mean square error from the 10-fold CV was used to select the number of components, i.e. 20 components, to be used in the regression.

As can be seen in Figure 4, both the LASSO and PLSR models perform similarly on the given data set. From the results, PLSR has a potentially negligible higher performance in terms of NRMSE and R-squared when compared with LASSO. It should be noted that LASSO here uses 67 of the original 87 variables, which can be valuable from an operational point of view in terms of execution time. The regression models also struggle to determine higher chl-a. Possibly, due to lack of data or non-linear effects.

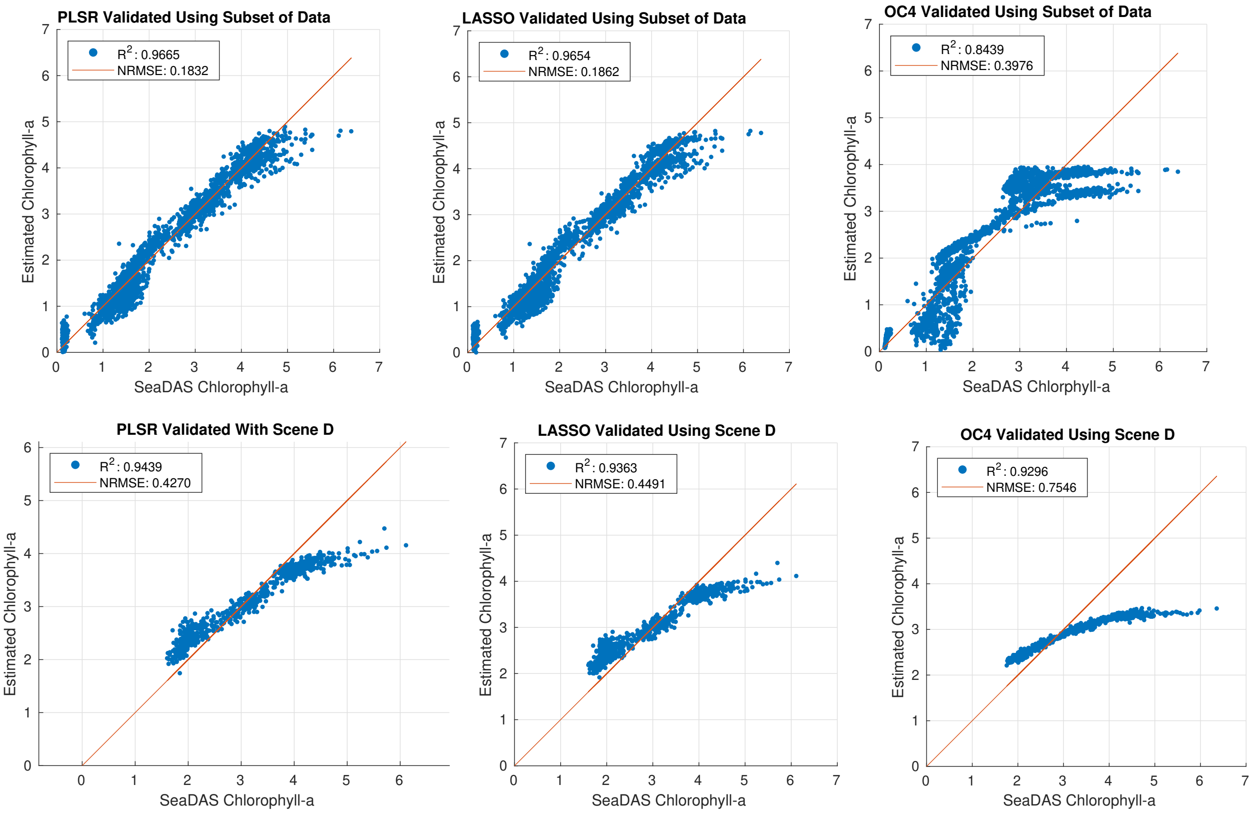

Fig. 4: Results from scenes showing Normalized Root Mean Square and R squared.

For the validation with a separated scene, scene D in Figure 1, the regression models are also able to determine a relative and more accurate quantification of the chl-a. The regression models have a better performance to the OC4 algorithm with the validation scheme using data from all scenes.

4.3. Comparison

It should be noted that this is a test of the internal consistency of OC4 within the SeaDAS software, i.e. how different bands used as a basis for OC4 will perform. A more proper data set with ground-truth samples measured by other means would be a better study, and could even yield an even higher performance increase with the approach given in this paper. The approach used here should benefit the OC4 algorithm.

With the preprocessing described in section 3 the variables are an atmospheric correction and a 4th order polynomial kernel [6] from having the same form as equation (3). The data representation chosen for regression incorporates a well-characterized non-linear physical relationship, e.g. transformation from radiance to reflectance and light attenuation. This ensures that the machine learning algorithms, i.e. different forms for regression, do not put a lot of emphasis on estimating non-linear relationships.

As can be seen in Figure 4, both the LASSO and PLSR models perform better than the OC4 algorithm in terms of NRMSE and R-squared for the used validation schemes. The two regression models have a similar performance in terms of the chosen metrics, but the execution time of the LASSO regression was on average 1.8 times faster.

From the coefficient illustrated in Figure 3, it is clear that the LASSO and PLSR approach puts emphasis on similar parts of the electromagnetic spectrum. It is also clear that some of the coefficients, $>$ 555 nm and $<$ 443 nm, have high expressive power in terms of determining the total chl-a concentration. The wavelengths 555 nm and 443nm indicate the maximum and minimum of the OC4 algorithm. That additional spectral information improves chlorophyll determination, and this corresponds well with other findings investigating band-ratio algorithms [2].

5. CONCLUSIONS & FUTURE WORK

The presented machine learning models seem may provide absolute measurements of chl-a concentration from only using the measured top-of-the-atmosphere radiance, the attitude and solar angle information related to the hyperspectral sensor.

Multivariate methods such as PLSR seems to be suitable for deriving some geophysical variables of interest such as chlorophyll-a concentration. At the same time, these linear methods can provide an interpretable derivation of results in the form of coefficients. This makes it easier to understand why the models derive the values that they do, which in return can provide reassurance to the end-user.

The LASSO model, when compared to PLSR, provided a reduction to 54% in the average computational time per pixel, encourages its use in computationally constrained systems. These results are implementation and hardware dependent.

When doing machine learning there is a considerable advantage of having large data sets with verified ground truth, but this is not widely available for hyperspectral ocean color remote sensing data, thus HICO data quality assured through SeaDAS was used. When more ground truth data become available with future space missions and systems, better models could be developed using this approach. Also, the data used in this paper does only represent a subset of the full range of naturally occurring chlorophyll-a concentrations. With more data, better models could be developed.

Preprocessing with targeted binning of spectral regions of interest for chlorophyll-a concentration, found through analysis of the regression coefficients, could improve the signal to noise ratio of the spectra and in return improve the estimation performance, and should thus be further investigated.

6. ACKNOWLEDGEMENTS

This work was supported by the Norwegian Research Council through the Centre of Autonomous Marine Operations and Systems (NTNU AMOS) (grant no. 223254), the MASSIVE project (grant no. 270959).

7. REFERENCES

[1] Trevor Platt, *Why ocean colour?: The societal benefits of ocean-colour technology*, Number 7 in IOCCG reports. International Ocean-Colour Coordinating Group, 2008.

[2] John E O'Reilly and P Jeremy Werdell, "Chlorophyll algorithms for ocean color sensors-oc4, oc5 & oc6," *Remote sensing of environment*, vol. 229, pp. 32–47, 2019.

[3] Kimberly Ryan and Khalid Ali, "Application of a partial least-squares regression model to retrieve chlorophyll-a concentrations in coastal waters using hyper-spectral data," *Ocean Science Journal*, vol. 51, no. 2, pp. 209–221, 2016.

[4] Katalin Blix and Torbjørn Eltoft, "Machine learning automatic model selection algorithm for oceanic chlorophyll-a content retrieval," *Remote Sensing*, vol. 10, no. 5, pp. 775, 2018.

[5] Kristin Tøndel and Harald Martens, "Analyzing complex mathematical model behavior by partial least squares regression-based multivariate metamodeling," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 440–475, 2014.

[6] Bernhard Scholkopf and Alexander J Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2001.

[7] IOCCG, *Atmospheric Correction for Remotely-Sensed Ocean-Colour Products*, vol. No. 10 of *Reports of the International Ocean Colour Coordinating Group*, IOCCG, Dartmouth, Canada, 2010.

[8] Amir Ibrahim, Bryan Franz, Ziauddin Ahmad, Richard Healy, Kirk Knobelspiesse, Bo-Cai Gao, Chris Proctor, and Peng-Wang Zhai, "Atmospheric correction for hyperspectral ocean color retrieval with application to the hyperspectral imager for the coastal ocean (hico)," *Remote Sensing of Environment*, vol. 204, pp. 60–75, 2018.

[9] Robert L Lucke, Michael Corson, Norman R McGlothlin, Steve D Butcher, Daniel L Wood, Daniel R Korwan, Rong R Li, Willliam A Snyder, Curt O Davis, and Davidson T Chen, "Hyperspectral imager for the coastal ocean: instrument description and first images," *Applied optics*, vol. 50, no. 11, pp. 1501–1516, 2011.

[10] "Hico - sample image gallery," `http://hico.coas.oregonstate.edu/gallery/gallery-scenes.php`, (Accessed on 04-02-2020).

[11] "Nasa ocean color," `https://oceancolor.gsfc.nasa.gov/atbd/chlor_a/`, (Accessed on 03-03-2020).

[12] Howard R Gordon, "Can the lambert-beer law be applied to the diffuse attenuation coefficient of ocean water?," *Limnology and Oceanography*, vol. 34, no. 8, pp. 1389–1409, 1989.

[13] John E O'Reilly, Stephane Maritorena, B Greg Mitchell, David A Siegel, Kendall L Carder, Sara A Garver, Mati Kahru, and Charles McClain, "Ocean color chlorophyll algorithms for seawifs," *Journal of Geophysical Research: Oceans*, vol. 103, no. C11, pp. 24937–24953, 1998.

[14] Svante Wold, Michael Sjöström, and Lennart Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and intelligent laboratory systems*, vol. 58, no. 2, pp. 109–130, 2001.

[15] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.