Estimation of Chlorophyll a, Diatoms and Green Algae Based on Hyperspectral Data with Machine Learning Approaches

PHILIPP M. MAIER[1], STEFAN HINZ[1] & SINA KELLER[1]

Abstract: Monitoring of inland waters is a major topic in terms of water quality and environmental issues. We suggest that hyperspectral remote sensing can provide a valuable data source to monitor important water parameters. In this study, we sampled hyperspectral data on The River Elbe from a research ship. The data, which we collected, contains 1383 datapoints and demonstrates similar spectral behavior to the data in the literature. A Random Forest regression shows that concentration of chlorophyll a, green algae, and diatoms is predictable based on hyperspectral data with a R^2 of around 80%. In general, the results reveal the potential of estimating chlorophyll a concentration and concentration of different algae taxa with machine learning approaches. For prospective studies, we intend to develop a generic model approach which is able to process various types of input data, measured at different inland waters.

1 Introduction

During the last decades, monitoring of inland waters has become a major research topic in terms of water quality and environmental issues. The monitoring of area-wide water bodies, is highly data intensive. Most of the currently available datasets to evaluate the quality of inland waters consist of sampled point data. To derive information on the entire water body, this data can be unreliable. Recent attempts towards such an area-wide coverage of water quality monitoring have included the application of hyperspectral sensors to gather image-based, remote sensing data. Chlorophyll a (Chl a) and turbidity function as indicators of algae existence which in turn characterize water quality and nutrition supply.

The first approaches to monitor Chl a concentrations with hyperspectral measurements were undertaken by NEVILLE & GOWER (1977). They discovered the correlation between the absorption peak as a minimum of the reflectance at a wavelength of 685 nm and the Chl a concentration in freshwater. GITELSON & KEYDON (1990) and GITELSON (1992) described hyperspectral charts of different types of inland waters. They sampled hyperspectral spectrometer-based data in different climate regions during various seasons and at several trophic states. According to their research, three significant extrema in the charts exist.

One maximum corresponds to the Chl a peak (minimum of the reflectance) discovered by NEVILLE & GOWER (1977), although it lays in a range between 670 nm to 680 nm. Another one covers the range between 550 nm to 570 nm as a maximum of the reflectance. It is identified as a backscattering feature from suspended particles in water (GITELSON 1992). The last maximum varies in the range between 685 nm to 700 nm. The latter increases in the course of higher Chl a

[1] Karlsruhe Institute of Technology (KIT), Institute of Photogrammetry and Remote Sensing, Englerstraße 7, D-76131 Karlsruhe, E-Mail: [Philipp.Maier, Stefan.Hinz, Sina.Keller]@kit.edu

concentration (GITELSON & KEYDON 1990). Thereupon, GITELSON (1992) figured out that the maximum of the reflectance between 675 nm and 730 nm shifts towards longer wavelengths with increasing Chl a concentration.

GITELSON (1992) proposed two different types of approaches to predict Chl a concentration: the first one is based on the peak height in the range between 675 nm to 730 nm. The second one relies on the ratio of specific bands. The ratio approach has been pursued and enhanced afterwards by several researchers (see e.g. SCHALLES et al. 1998). MANNHEIM et al. (2004) presented an area-based approach, where the area under the peak without the baseline correlates with the Chl a concentration. RUNDQUIST (1996), FRASER (1998), CRAIG et al. (2006) attempted to include derivatives up to the fourth order, to enhance the predictability of Chl a.

Besides the Chl a concentration, the differentiation of algae taxa is an important topic of modeling environmental processes and water quality. Especially, the cyanobacteria, commonly known as blue-green algae, is a taxon, where a distinction is necessary. It contains several species that can be harmful for animals and human beings when they disperse in drinking water reservoirs. SIMIS et al. (2007) conducted lots of research about cyanobacteria and their unique pigment phycocyanin. The latter features its typical absorption peak at 620 nm. Spectral signatures of further algae taxa were researched e.g. by GITELSON et al. (1999) and HUNTER et al. (2008) in field experiments as well as laboratory experiments such as artificial tanks. In the laboratory studies, they varied the taxa and the Chl a concentration to observe the change in the spectral reflectance signatures. According to their studies, different algae taxa can be distinguished based on the hyperspectral signature. At this point, the hyperspectral sensor demonstrates enhanced performance over the multispectral sensor (HUNTER et al. 2008).

In general, it can be stated: the higher the Chl a concentration, the more the distinguishability between the algae taxa increases. HUNTER et al. (2008) also created synthetic algae cultures by mixing two distinct taxa within different concentrations and demonstrated that they still can be distinguished, while the spectral signature appears as mixture.

GITELSON (1999) remarked that the occurrence of suspended particular organic matter (SPOM), suspended particular inorganic matter (SPIM), and colored dissolved organic matter (CDOM) challenge the prediction of Chl a concentration with spectral information. The three matters occur in most inland water bodies. They influence the backscattering behavior by interacting with the backscattering of chlorophyll and other algal pigments (GITELSON et al. (1999)). Generally, SPIM and SPOM increase the reflectance of the water body (HUNTER et al. 2008). To handle high SPIM and SPOM concentration in the context of predicting Chl a, ZHOU et al. (2013) proposed a multi-band ratio. Reviews of most of the available approaches for Chl a prediction and prediction of several taxa can be found in MATTHEWS et al. (2010) and PALMER et al. (2015).

In this contribution, we present a hyperspectral dataset with reference data, sampled on The River Elbe. We propose a basic machine learning approach to investigate the Chl a concentration in a first step. In a second step we focus on the distinction of two algae taxa: the Green Algae and the Diatoms. The preliminary regression results are conducted with a standardized machine learning regressor, the Random Forest (RF) regression.

2 Data Acquisition & Data Processing

The data, presented in this contribution, was sampled with a hyperspectral sensor mounted on a research ship (Fig. 1), along The River Elbe from Bad Schandau to Geesthacht in Germany. The data acquisition was embedded in the scope of the Elbschwimmstaffel, which took place in summer 2017 in Germany and was funded by the Federal Ministry of Education and Research (BMBF). We sequentially performed measurements along an approximately 500 km long section on The River Elbe, which result in a large set of datapoints. The applied hyperspectral sensor was a Cubert UHD 285 characterized by an amount of 125 bands in the range from 450 nm to 950 nm. Additionally, the multi-sensor system PhycoSens (invention of BBE Moldaenke) sampled in-situ water parameters. Concentration of Chl a, diatoms, and green algae as well as turbidity represent the target variables.

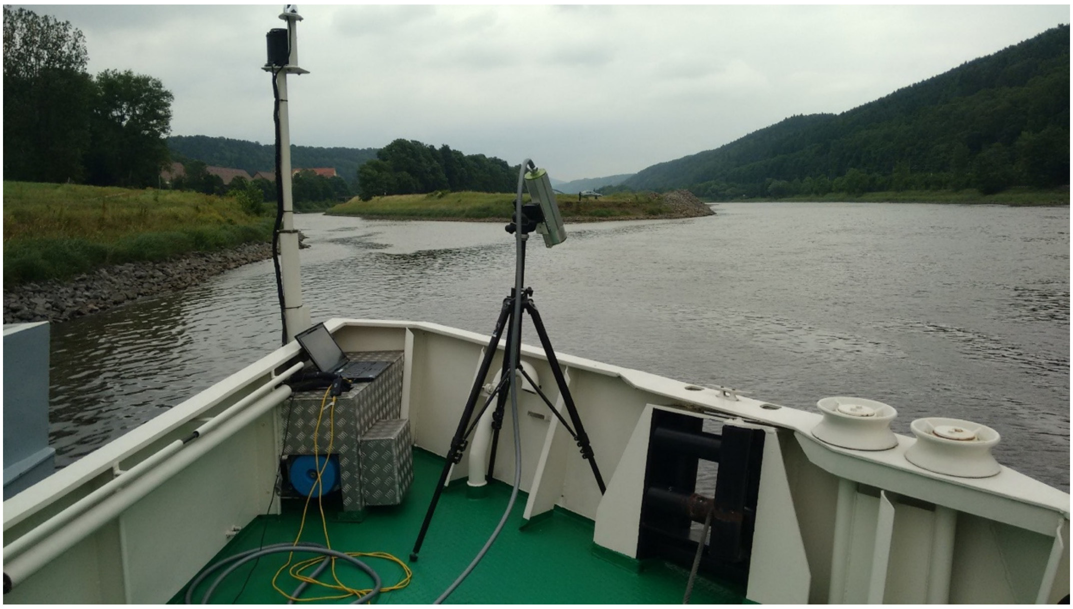

Fig. 1: Hyperspectral sensor mounted on the bow of the research ship

The hyperspectral sensor was calibrated every 20 minutes with a white reference, i.e. spectralon, to compensate the varying sun altitude. Every minute we captured a hyperspectral snapshot within a 70° angle towards the water surface. The spectralon was placed on the railing so that a part of it is visible in every snapshot to control the reflectance. In addition, we equalized minor radiative fluctuations such as slight cloud occurrences. The reference data was sampled every five minutes by the PhycoSens. To expand the dataset, we interpolated the reference data linearly. Since the measured Chl a concentration has changed in a continuous matter, the interpolation is feasible in practice.

During the data processing, the measured mean spectra was calculated by manually selecting an area, which was undisturbed by bubble formations, shadows, or waves. The effect of bubbles or waves can be seen in Fig. 2, resulting in a high variance and higher reflection values. In a next step, we applied two distinctive types of cuts.

First, based on the spectralon reflectance values of the calibration, we removed outlying images, whose reflectance on the spectralon differed around the factor of 0.3. The outlying behavior was mainly affected by shadow occurrence caused by ship turnings, bridge crossings, or sudden cloud coverages. Second, data points, sampled by the PhycoSens, with a Chl a concentration above 200 µg/l, were dismissed. These values exceed the measurement range. In total, we removed more than 500 of our former 2000 hyperspectral images.

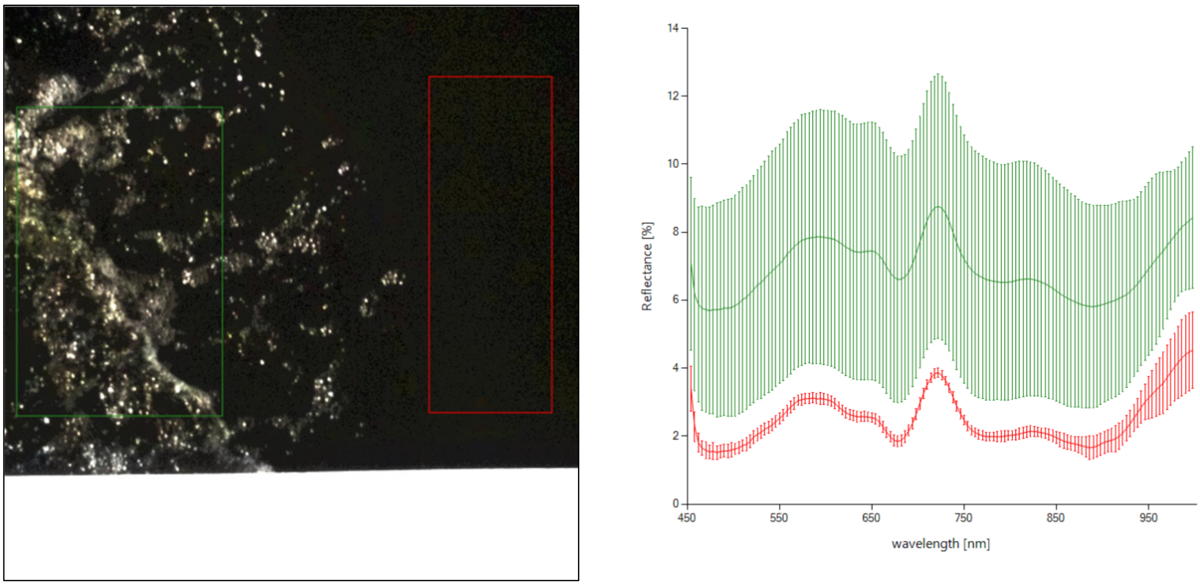

Fig. 2: Left: Hyperspectral snapshot shown as a RGB image. The white area on the bottom of the image represents the spectralon. The red and the green boxes are the selected areas to calculate the mean reflectance. Right: The charts of the selected areas (cf. left part of the figure). The green plot is characterized by "bubble" artefacts and a high variance. The red plot represents undisturbed water surface

Finally, the dataset consists of 1383 datapoints. Each of these refers to 105 hyperspectral values in the range between 486 nm to 902 nm and four measured parameters by the PhycoSens. The bands in the range between 450 nm to 485 nm and 903 nm to 950 nm were cut due to inconsistent behavior and noise.

3 Methods

For the preliminary analysis of the dataset, we perform a RF regression to predict the Chl a concentration based on hyperspectral data. The modeling of the green algae and the diatoms concentration, follows the same approach.

To perform the RF regression, we apply the *Ranger* package (WRIGHT & ZIEGLER 2017) in R. The model is tuned with a grid search employing the train function from the *caret* package (KUHN 2008) to find the best number of variables that are randomly sampled at each split. As splitting rule for the RF algorithm we process *extratrees*.

Several splits of the dataset are implemented: the training dataset is divided in 20, 50 and 80 percent of the dataset. The counterpart of the dataset represents the validation dataset respectively.

In addition to the splitting of the dataset, we evaluate the performance of the framework using a Principle Component Analysis (PCA) to reduce the dimensionality. Then, both approaches are trained with the first 20 components. The regression performance is expressed by the coefficient of determination R^2 and the mean absolute error MAE.

4 Results & Discussion

Fig. 3 shows the dataset in boxplots. Two peaks can be distinguished regarding the mean values:the first is in the range between 582 nm and 606 nm and the second one appears between 714 nm and 718 nm. The first one indicates, that the peak shifts to longer wavelengths with increasing reflectance. In GITELSON (1990) this maximum appears in the range between 550 nm and 570 nm and is affected by reflectance of Chl a as well as by backscattering on suspended particles in water.

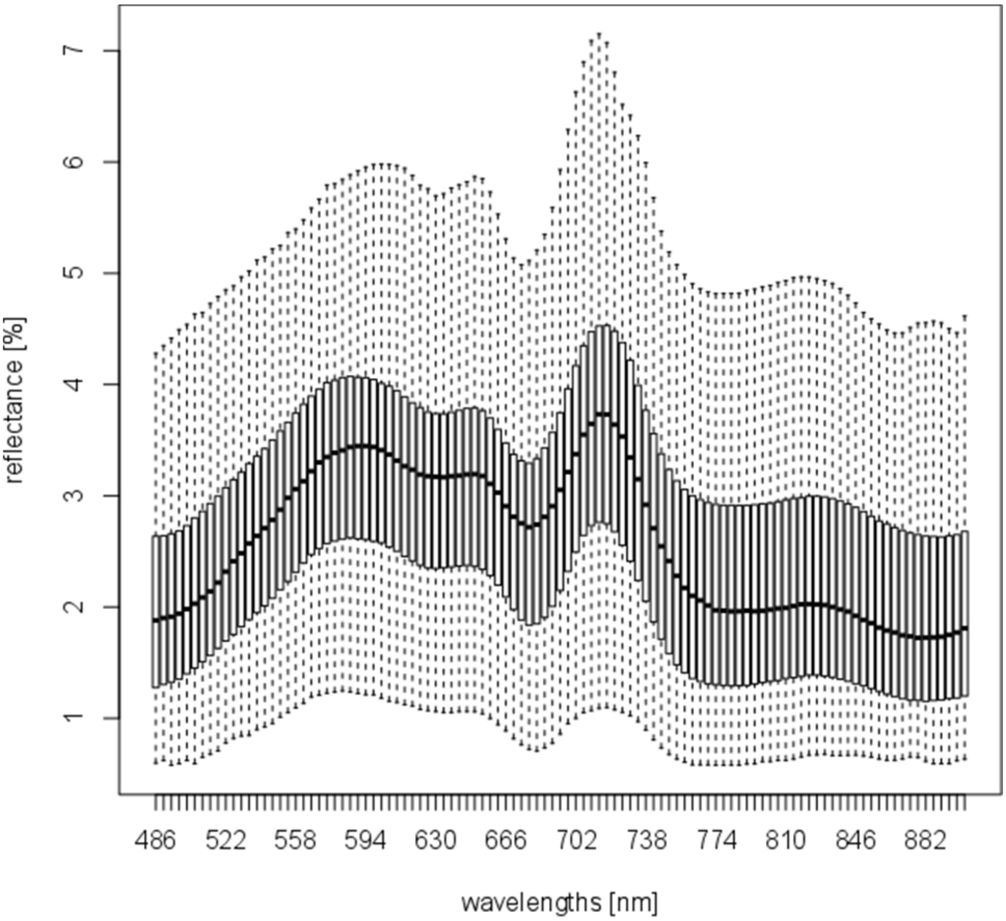

Fig. 3: The dataset is presented as boxplots between the wavelength of 486 nm and 902 nm over all 1383 datapoints. The whiskers represent the extreme values

The River Elbe is well known for its high concentration of SPIM. It can be possible that the maximum shifts even further to longer wavelengths, with a growing SPIM concentration. However, we don't have reference data to proof this behavior. The second peak is related to Chl a reflectance and is similar to values found in the literature (GITELSON 1990). The reflectance minimum at 678 nm represents a significant minimum, due to the absorption of Chl a and also suits to the literature values (GITELSON 1990).

In addition to the main extrema, slight maxima occur at 650 nm and 822 nm respectively. The general reflectance of the wavelengths varies between < 1% and 7%. The highest amplitude in the data is located in the area around the peak between 714 nm and 718 nm. All things considered, the dataset shows no inconsistent behavior.

In Fig. 4 the green line represents the mean spectrum of high Chl a concentration (>150 µg/l) and the blue line shows the mean spectrum of the datapoints with low Chl a concentration (< 50 µg/l). The reflectance of the low Chl a concentration spectrum is generally lower than the high Chl a concentration spectrum. The global peak shifts slightly from 714 nm to 718 nm with higher Chl a concentration.

Additionally, the amplitude from the minimum at 678 nm to the maximum is higher in case of the high Chl a concentration with respect to the lower concentration. In general, a trend can be observed, that an increasing concentration of Chl a, results in higher. But this may also be a consequence of higher concentration of suspended particles in the water.

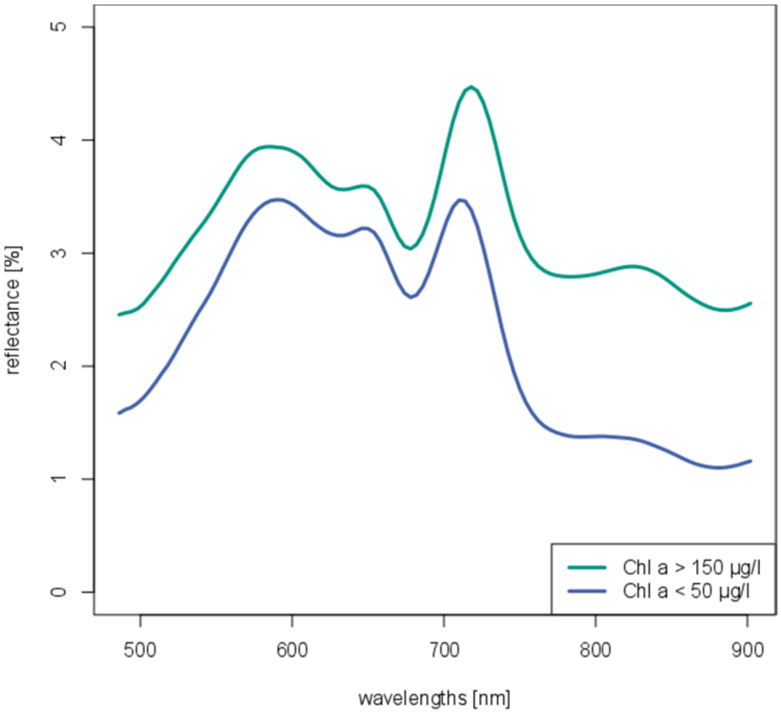

Fig. 4: Mean spectra of two different Chlo a concentration

Tab. 1 outlines the regression results for Chl a concentration with the RF model as estimator and the hyperspectral data as input vector. The regressor is trained with a different amount of data. The quality of the RF model is shown as R² and MAE.

In general, the performance increases with a higher amount of training data. But using more than 50% of the training data improves the results only slightly. Applying a PCA before the regression to reduce features enhances the performance. In this modeling, we pick the first 20 components of the PCA to fit the RF model.

Tab. 1: R^2 and MAE for LR and RF and PCA for the regression

training data	RF		RF + PCA	
[%]	R^2 [%]	MAE [µg/l]	R^2 [%]	MAE [µg/l]
20	65	15,2	72	13,0
50	80	11,1	85	9,6
70	83	10,0	87	8,3

Tab. 2 summarizes the regression results for the estimation of green algae and diatoms concentration with the RF algorithm. In general, the estimation of diatoms performs better than the estimation of the green algae, since the range of the concentration of green algae is about double the range of diatom concentration. This is an important aspect when it comes to the training of the algorithm.

PCA has a similar influence on the estimation of the Chl a concentration: performing a PCA before model fitting improves the regression results significantly because of highly correlated variables.

Tab. 2: R^2 and MAE for Green Algae and Diatom concentration operating the RF algorithm

training data	Green Algae		Green Algae + PCA		Diatoms		Diatoms + PCA	
[%]	R^2 [%]	MAE [µg/l]	R^2 [%]	MAE [µg/l]	R^2 [%]	MAE [µg/l]	R^2 [%]	MAE [µg/l]
20	57	10,5	63	9,8	72	6,5	79	5,8
50	69	8,6	76	7,6	79	5,3	86	4,5
70	77	7,4	82	6,4	82	4,7	88	4,0

5 Conclusion & Outlook

In general, the results reveal the potential of estimating Chl a concentration and concentration of different algae taxa with machine learning approaches. Although the river Elbe is a challenging inland water with high concentration in suspended particular substances, the RF performs well for all three target variables.

For prospective studies, we intend to develop a generic model approach, which is able to process various types of input data, measured at different kinds of inland waters. The model should later work with hyperspectral data collected by a UAV.

6 Acknowledgment

We thank the organizing committee of the Elbschwimmstaffel for the opportunity to participate in this field campaign. We also thank BBE Moldaenke for sharing with us the reference data. Finally, we express our gratitude to the crew of the research ship "Elbegrund" for their technical. support.

7 References

CRAIG, S.E., LOHRENZ, S.E., LEE, Z., MAHONEY, K.L., KIRKPATRICK, G.J., SCHOFIELD, O.M. & STEWARD, R.G., 2006: Use of hyperspectral remote sensing reflectance for detection and assessment of the harmful alga, Karenia brevis. Applied Optics, **45**(21), 5414-5425.

FRASER, R.N., 1998: Multispectral remote sensing of turbidity among Nebraska Sand Hills lakes. International Journal of Remote Sensing, **19**(15), 3011-3016.

GITELSON, A.A. & KEYDAN, G.P., 1990: Remote Sensing of Inland Surface Water Quality—Measurements in the Visible Spectrum. ratio, **2**, 7.

GITELSON, A., 1992: The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. International Journal of Remote Sensing, **13**(17), 3367-3373.

GITELSON, A.A., SCHALLES, J.F., RUNDQUIST, D.C., SCHIEBE, F.R., & YACOBI, Y.Z., 1999: Comparative reflectance properties of algal cultures with manipulated densities. Journal of Applied Phycology, **11**(4), 345-354.

HUNTER, P.D., TYLER, A.N., PRÉSING, M., KOVÁCS, A.W. & PRESTON, T., 2008: Spectral discrimination of phytoplankton colour groups: The effect of suspended particulate matter and sensor spectral resolution. Remote Sensing of Environment, **112**(4), 1527-1544.

KUHN, M., 2008: Caret package. Journal of Statistical Software, **28** (5).

MANNHEIM, S., SEGL, K., HEIM, B. & KAUFMANN, H., 2004: Monitoring of lake water quality using hyperspectral CHRIS-PROBA data. In Proc. of the 2nd CHRIS/PROBA workshop, ESA/ESRIN, Frascati, Italy.

MATTHEWS, M.W., BERNARD, S. & WINTER, K., 2010: Remote sensing of cyanobacteria-dominant algal blooms and water quality parameters in Zeekoevlei, a small hypertrophic lake, using MERIS. Remote Sensing of Environment, **114**(9), 2070-2087.

NEVILLE, R.A. & GOWER, J.F.R., 1977: Passive remote sensing of phytoplankton via chlorophyll α fluorescence. Journal of Geophysical Research, **82**(24), 3487-3493.

PALMER, S.C., KUTSER, T., & HUNTER, P.D., 2015: Remote sensing of inland waters: Challenges, progress and future directions.

RUNDQUIST, D.C., HAN, L., SCHALLES, J.F. & PEAKE, J.S., 1996: Remote measurement of algal chlorophyll in surface waters: the case for the first derivative of reflectance near 690 nm. Photogrammetric Engineering and Remote Sensing, **62**(2), 195-200.

SCHALLES, J.F., GITELSON, A.A., YACOBI, Y.Z., & KROENKE, A.E., 1998: Estimation of chlorophyll a from time series measurements of high spectral resolution reflectance in an eutrophic lake. Journal of Phycology, **34**(2), 383-390.

SIMIS, S.G., RUIZ-VERDÚ, A., DOMÍNGUEZ-GÓMEZ, J.A., PEÑA-MARTINEZ, R., PETERS, S.W. & GONS, H.J., 2007: Influence of phytoplankton pigment composition on remote sensing of cyanobacterial biomass. Remote Sensing of Environment, **106**(4), 414-427.

WRIGHT, M.N. & ZIEGLER, A.,2017: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, **77**(1), 1-17.