UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Fjodor Ševtšenko

CDOM-based Optical Water Types Classification

Master's Thesis (15 ECTS)

Supervisor(s): Krista Alikas, PhD
Radwa El Shawi, PhD

Tartu 2024

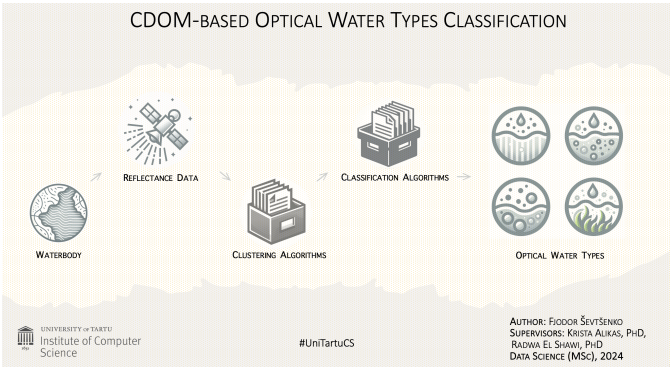
CDOM-based Optical Water Types Classification

Abstract:
A water body is a habitat, where the interaction between the living and non-living matter is observed. Any changes in the water bodies' characteristics influence the internal processes and may have a negative long-run impact on the features of ecosystem and its diversity. The main object of the current research is related to Colored Dissolved Organic Matter (CDOM). When the concentration of CDOM becomes high, then the water gets brown and the underwater light changes. The water brownification is one of that processes, that may heavily influence the features of the ecosystem and its diversity. It is important to have the information about the water bodies' states in a format of brownification scale. The water body classification topic was previously observed in early studies via establishing the term of Optical Water Type (OWT). The previous research was mostly related to creation of one optimal OWT classifier, where each class combines the combination of water parameters like CDOM, chlorophyll A, total suspended solids and secchi depth. Comparing to the previous studies, this work proposes a classification approach that could be selected based on tasks' characteristics. Also, instead of OWT, the work establishes the term of CDOM based OWT (CDOM-OWT), that classifies the water body based on CDOM relative concentration level. The CDOM-OWT classification approach is useful, when is needed to plot the CDOM relative concentration levels on location maps for various periods to monitor the spreading dynamics.

Keywords:
optical water type; remote sensing; Sentinel-2 MSI; CDOM; artificial intelligence; automl

CERCS: P170 Computer science, numerical analysis, systems, control; P176 Artificial intelligence; T181 Remote sensing

Graphical abstract[1]**:**



[1]Icons were created using ChatGPT DALL-E service.

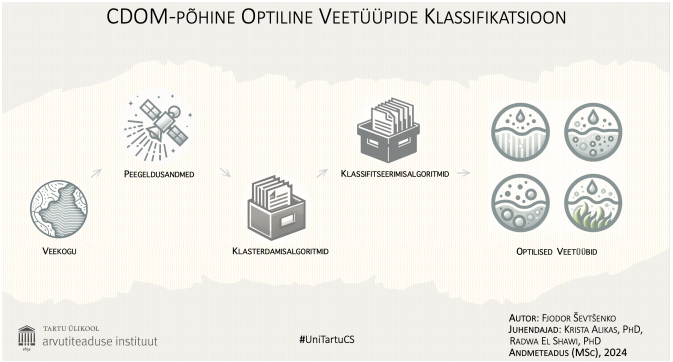
CDOM-põhine Optiline Veetüüpide Klassifikatsioon

Lühikokkuvõte:
Veekogu on elupaik, kus vaatleme elus ja eluta looduse vastasmõju. Igasugused muutused veekogude omadustes mõjutavad sisemisi protsesse ning võivad avaldada pikaajalist mõju ökosüsteemi iseärasustele ja selle mitmekesisusele. Käesoleva uurimistöö põhiobjekt on seotud värvilise lahustatud orgaanilise ainega (CDOM). Kui CDOM-i kontsentratsioon muutub kõrgeks, muutub vesi pruuniks. Vee pruunistumine on üks neist protsessidest, mis võib oluliselt mõjutada ökosüsteemi iseärasusi ja selle mitmekesisust. Oluline on omada teavet veekogude seisundite kohta pruunistumise skaala kujul. Veekogude klassifikatsiooni teemat täheldati varasemates uuringutes optilise veetüübi (OWT) termini kehtestamise kaudu. Eelnev uurimus oli peamiselt seotud ühe optimaalse OWT klassifikaatori loomisega, kus igas klassis on kombineeritud veeparameetrid nagu CDOM, klorofüll a, heljumi koguhulk ja läbipaistvus. Võrreldes varasemate uuringutega, pakub käesolev töö välja klassifikaatsiooni lähenemisviisi, mida saab valida ülesannete omaduste põhjal. Samuti kehtestatakse töös OWT asemel termin CDOM-põhine OWT (CDOM-OWT), mis klassifitseerib veekogu ainult CDOM-i suhtelise kontsentratsioonitaseme alusel. CDOM-OWT klassifitseerimisviis on kasulik, kui leviku dünaamika jälgimiseks on vaja CDOM-i suhtelised kontsentratsioonitasemed asukohakaartidele erinevate perioodide jaoks joonistada.

Võtmesõnad:
optiline veetüüp; kaugseire; Sentinel-2 MSI; CDOM; tehisintellekt; automl

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria); P176 Tehisintellekt; T181 Kaugseire

Visuaalne kokkuvõte[2]:



[2]Icons were created using ChatGPT DALL-E service.

Contents

1 Acronyms and abbreviations

Acronym/abbreviation	**Full term**
AET	**A**quatic **E**nvironmental **P**arameter
CDOM	**C**olored **D**issolved **O**rganic **M**atter
Chla	**Chl**orophyll **A**
OWT	**O**ptical **W**ater Type
TSS	**T**otal **S**uspended **S**olid

2 Introduction

The water is playing the crucial role in every aspect of life. On the micro level, the water participates in metabolic processes of organisms carrying nutrients to cells. On the macro level, the water mass is responsible for photosynthesis that occurs in aquatic plants and bacteria, guarantees the Earth's climate regulation, mitigating the effect of global warming.

Besides the general importance of the water, an arbitrary water body is a habitat, where the interaction between the living and non-living matter is observed. Any changes in the water bodies' characteristics influence the internal processes and may have a long-run impact on the features of ecosystem and its diversity.

The main object of the current research is related to Colored Dissolved Organic Matter (CDOM)[3]. CDOM consists of organic molecules got from remains of plants and animals or may be the product of the chemical reactions between non-organic molecules. The human activities related to agriculture or deforestation may increase the level of CDOM concentration. The environmental activities associated with natural processes, like soil runoff or plants decay, may also increase the volume of CDOM.

When the concentration of CDOM becomes high, then the water gets brown. The water brownification process is often associated with the increase of CDOM in water body [Blanchet et al., 2022]. The water brownification is one of that processes, that may have a long-run impact on the features of ecosystem and its diversity [Horppila et al., 2022]. CDOM particles heavily absorb the incoming light, decreasing the volumes required by a particular water body habitants, therefore negatively influencing their behavior.

It is important to have the information about the water bodies' states in a format of brownification scale. The information can be gathered *on-site*, collecting the in-situ measurements about the chemical and optical water properties like CDOM. Getting the information this way could be a very precise, but only for a particular water body's location. It is quite difficult to cover the entire water body with the appropriate samples (*spatial problem*). Also, it is problematic to guarantee the temporal dynamics' coverage for the entire water body (*temporal problem*). Among the possible constraints of *on-site* approach, there also could be mentioned the budget limits as well as the physical accessibility of the water body. Taking this into consideration, the *remote sensing* approach could be considered as a viable alternative.

The ability to apply the remote sensing approach is able to solve the mentioned spatial and temporal problems, while conducting the monitoring activities for the whole water body. In this case, the measurements are performed from the distance by the sensors located on satellites. The Landsat and Sentinel are two major programs that were designed for Earth remote observation. The data from satellite missions are available

[3]CDOM definition: `https://en.wikipedia.org/wiki/Colored_dissolved_organic_matter` Read: 10.07.2024

from 1972 for Landsat and from 2014 for Sentinel.

The status of brownification could be measured by gathering the optical water characteristics in a form of light *reflectance* arriving to the satellite sensor. The general way to calculate the remote sensing reflectance (R_{rs}) is depicted in the equation 1, where L_w and E_d correspond to water leaving radiance and downwelling irradiance respectively[4].

$$R_{rs} = \frac{L_w}{E_d} \tag{1}$$

Next, the remote sensing reflectance will be referenced as a reflectance.

Reflectance values of CDOM are mainly influenced by its heavy absorption in bands 1 (443 nm) and 2 (490 nm) and moderate and lower absorption in bands 3 (560 nm) and 4 (665 nm) of Sentinel-2 MSI sensor. As a result, the low reflectance is observed in bands 1 and 2, the moderate reflectance and relatively high are seen in bands 3 and 4 respectively. The reflectance values increase in bands 1 and 2, then the higher value is in band 3 and the highest value relates to band 4.

Omitting the atmospheric influence, the magnitude of the reflectance is determined not only by CDOM, but by the combination of optically active substances, like chlorophyll A (Chla), colored dissolved organic matter (CDOM), total suspended solid (TSS). The particles are situated in water and have the abilities to differently absorb and scatter the light, what in turn define the value of reflectance. A different water body may have a different concentrations and combinations of the particles. Although, the different combinations of particles usually lead to the different magnitude of the reflectance, it is also possible to get the same reflectance values for the different combinations. The *problem of ambiguity* could be significantly decreased by observing the reflectance in different light spectrum bands, providing us the ability to differentiate between the combinations of the particles, assigning them a proper unique label - the Optical Water Type.

The *goal* of the thesis is to develop and implement the CDOM based Optical Water Types (CDOM-OWT) classification approach, having the *hypothesis* that the reflectance measurement observed under different light spectrum bands is not posing the problem of ambiguity. The reflectance values will be used as the features for clustering and classification activities. To get the indication of CDOM concentration the *absorption coefficient of CDOM* at wavelength 440 or 442 will be used. At that point the CDOM absorption is strong and distinctable. Next, the absorption coefficient will be referenced as a CDOM.

The main idea of the approach applied in the thesis together with some of the outputs are the following. Instead of creation of the optimal OWT classification model, there

[4]ARSET - Integrating Remote Sensing into a Water Quality Monitoring Program: https://appliedsciences.nasa.gov/get-involved/training/english/arset-integrating-remote-sensing-water-quality-monitoring-program Read: 30.07.2024

will be selected a way to generate the multiple number of models using the AutoML techniques. First, the set of clustering algorithms will be applied to get the possible ways of data to be clustered. Then, each dataset will be processed by the set of classification algorithms. Next, the evaluation techniques will be introduced. The OWT to CDOM-OWT transformation method and the CDOM relative concentration computation will be established. The selection of the best classifier will be presented. Finally, having the optimal classifier and the CDOM concentration table available, the dataset from Sentinel-2 MSI satellite will be classified with CDOM-OWTs. The CDOM relative concentration levels will be plotted on the location maps.

This thesis assumes the prior knowledge about the applying of remote sensing concept to water bodies and the knowledge about the machine learning concepts in general and more specifically the techniques, like clustering and classification.

Among the number of *constraints*, there should be mentioned that the proposed OWT classification models were created using the data driven techniques. This means that the final model to use should be critically selected taking into consideration the set of requirements and the water bodies characteristics. Also, features of the datasets used in the training and validation of the models, like light reflectance as well as the informative attributes like CDOM, Chla, TSS/TSM, Secchi depth, were collected using a wide variety of tools that may add the certain bias to data. The data was used as it is without any special filters applied to eliminate the bias posed by tools or measurement techniques.

The *structure of the thesis* is the following. The chapter 3 presents the literature overview related to the developments of OWT classification taken place during the past decades. Here will be introduced the data used, methods applied as well as the results received. This will lead to the pattern, that the investigation was mostly done in a direction of having the one or limited number of OWT classification models. The chapter 4 provides the overview about the datasets utilized and methodological approaches applied. The chapter 5 and 6 presents the overall achievements and the discussion points respectively. The chapter 7 summarizes the work.

3 Background

This section provides the overview of the research that was previously done in the area of Optical Water Type (OWT) classification.

The water type classification was first defined by Nils Jerlov in [Jerlov, 1951]. The range of classes from clear to turbid were proposed to characterize oceans [Jerlov, 1976]. The downwelling diffuse attenuation coefficient (K_d) was used to get the relevant types. There were proposed 10 water types on clear-turbid scale.

The other attempt to group a water body based on the reflectance measurements at selected wavelengths was proposed by [Morel and Prieur, 1977]. The reflectance curves of oceanic water body were analysed and two extreme cases were identified. The Case 1 was marked based on the high concentration of phytoplankton compared to other particles. The Case 2, in contrast, contained the dominance of inorganic particles. The decision was mostly done based on mathematical rules derived using the reflectance data, absorption and scattering coefficients.

[Reinart et al., 2003] also made the research of the classification of water based on optical properties. The area of study was related to lakes and coastal waters of Estonia and south Finland that are representatives of a Case 2 type of water bodies. The work suggested to use five optical classes: clear, moderate, turbid, very turbid and brown. K-Means clustering method was used to get the water bodies clusters. The selected method allowed to apply the automatic classification driven by data.

The work related to optical water typing was also conducted by [Moore et al., 2009]. Instead of two Case classes the approach proposed eight water types that were received directly on the radiance measurements.

The other grouping attempt was proposed by [Spyrakos et al., 2018]. The study covered the development of topology of optical water types for inland and coastal waters. 13 spectrally distinct clusters were identified for inland water, and nine clusters - for marine environment. The decision was done based on clustering algorithm supported by functional analysis.

One other classification attempt based on reflectance spectra of in situ measurements was proposed by [Uudeberg et al., 2019] for boreal lakes and coastal areas. Five OWT classes were identified: clear, moderate, turbid, very turbid, and brown. The mathematical rules derived were used as the primary clustering technique. The results of the study were used further in [Uudeberg et al., 2020] to map the optical water quality parameters from reflectance spectra to the relevant empirical algorithms. The work targeted the issue of ineffective approach in the current water monitoring programs that are time-consuming, expensive, and what is most important may not reflect the entire state of water body.

Using in situ remote sensing reflectance data the assessment of OWT was done for Brazilian waters in [da Silva et al., 2020]. K-Means algorithm was also used as the clustering method with further manual corrections for the number of final clusters based on silhouette score. The classification algorithms were further applied for Sentinel-2 MSI

data in [da Silva et al., 2021]. Two Support Vector Machines were used for classifying the known OWTs with an attempt to integrate a novelty detection technique based on sigmoid function.

The result of the most of the observed classification approaches was a creation of one optimal classification model. The objective of this study focuses on the generation of many equivalent classification models that differ by number of proposed classes and models' hyperparameters used in training. The main idea behind is to provide a researcher with a selection option based on the task nature, needs and/or water bodies characteristics.

4 Data and methods

This section first describes the datasets used for training and evaluation. Then, the needed preprocessing steps are covered. Next, is presented the modeling approach, that consists of clustering and classification methods with the main goal to output the Optical Water Types (OWTs). The models' evaluation techniques are observed. The process of getting from OWT to CDOM based OWT (CDOM-OWT) and CDOM relative concentration computing is described. Finally, the optimal model selection process is covered.

4.1 Datasets

To achieve the thesis goal four datasets are used. Gloria [Lehmann et al., 2023], Spectra[5] and Brazil[6] datasets are in situ collected. S2A[7] (lake Võrtsjärv[8]) dataset is data collected from Sentinel-2 MSI satellite and used as a source to apply the trained model. The Gloria and Spectra datasets are entirely used for clustering algorithms training and evaluation. The train-splits of Gloria and Spectra datasets are used for classification algorithms training. The test-splits of Gloria and Spectra and the entire Brazil datasets are used for classification models evaluation and introduction of two possible ways of the optimal classifier selection. The classified S2A dataset is used to visually plot the selected model output, producing the CDOM relative concentration levels distribution on the location maps.

Each dataset contains the reflectance values. Gloria, Spectra and Brazil datasets also contain the Aquatic Environmental Parameters (AEP) values like CDOM, Chla, TSS, Secchi depth. The different instruments and methodologies were used to estimate the collected samples. To decrease the effects in reflectance values, set by different instruments and methodologies, the reflectance values across Sentinel-2 MSI bands within each sample are normalized. This technique is applied during the preprocessing steps. The normalized reflectance is used as the modeling input, where the differences in magnitude of a signal and relationships between the values of magnitude within the bands are expected to be the factors that helps to differentiate between the reflectance curves.

The AEP values are not used as part of modeling input and play the informative role, when the evaluation steps and the optimal model selection process take place. The Aquatic Environmental Parameter CDOM is the basis of OWT to CDOM-OWT

[5]Provided by University of Tartu, Tartu Observatory.

[6]Provided by Daniel Andrade Maciel, Claudio Barbosa, Edson Freirefs from Instrumentation Laboratory for Aquatic Systems (LabISA), INPE - National Institute for Space Research, São Jose dos Campos-Brazil.

[7]Loaded from ESTHub Processing Platform.

[8]Võrtsjärv, Estonia (58.3104° N, 26.0114° E): `https://en.wikipedia.org/wiki/V%C3%B5rtsj%C3%A4rv` Read: 05.07.2024

transformation technique. Also, CDOM is used in computation of CDOM order penalty, that is used in the optimal model selection step.

The next sections present a brief overview of datasets used in the study. For the reasons explained above, the variety of instruments and methodologies used for collecting the reflectance and water quality values are not covered.

4.1.1 Gloria dataset

Gloria dataset is "a globally representative hyperspectral in situ dataset for optical sensing of water quality" [Lehmann et al., 2023]. The data was collected by 59 institutions around the world for 450 various coastal and inland water bodies. The total number of collected samples equals to 7572. The reflectance measurements are done at 1 nm intervals for the 350 to 900 nm wavelength range. The Gloria dataset is opened and has a free access. Together with the reflectance measurements, the dataset provides the water quality measurements of absorption by dissolved substances[9], chlorophyll a, total suspended solids, and secchi depth.

4.1.2 Spectra dataset

Spectra dataset is also a collection of hyperspectral radiometric in situ data. It contains reflectance information from 53 different locations as well as the water quality measurements of absorption by dissolved substances[10], chlorophyll a, total suspended matter, and Secchi depth. The total number of collected samples equals to 670. The reflectance measurements are done at 1 nm intervals for the 350 to 900 nm wavelength range. Total suspended matter and total suspended solids are considered the same and presented as total suspended solid.

4.1.3 Brazil dataset

Brazil dataset is a subset of data used in [da Silva et al., 2020] and [da Silva et al., 2021]. The total number of samples available equals to 1011. The reflectance measurements are done at 1 nm intervals for the 400 to 900 nm wavelength range. The dataset also contains the water quality measurements of absorption by dissolved substances[11], chlorophyll a, total suspended solids, and secchi depth.

4.1.4 S2A dataset

S2A dataset is a collection of Sentinel-2 MSI multispectral data. It contains the reflectance information about the area of lake Võrtsjärv organized in Sentinel-2 MSI bands:

[9] Absorption coefficient of CDOM at wavelength 440.
[10] Absorption coefficient of CDOM at wavelength 442.
[11] Absorption coefficient of CDOM at wavelength 440.

B1, B2, B3, B4, B5, B6, B7, B8, B8A that correspond to the range from 412 to 907 nm. The data is collected from ESTHub Processing Platform by months for the period of 2021 - 2024. The collection period for each year is April to October. Data for 2024 is limited to the period from April to July. The Sentinel-2 MSI L1C (Input File Set) data was processed with Polymer processor (v4.16.1.) to obtain reflectance values over target water bodies.

4.1.5 Datasets preprocessing steps

The datasets preprocessing steps mainly involve the elimination of samples containing either noisy or missing elements as well as applying limits to wavelength to be used further in modeling.

Gloria dataset contains a set of flags that marks samples by inconsistency state. To get only non-noisy samples 0-flag is applied. Additionally, for Gloria, Spectra and Brazil datasets the range of wavelengths corresponds to the interval of 412 to 900 nm only. Also, the related reflectance measurements for that range should not contain any missing values or values below 0. The resulted statistics of water bodies characteristics for datasets are presented in Table 1.

Table 1. The statistics of water bodies characteristics for Gloria, Spectra and Brazil datasets.

Water body characteristic / Statistic	**Gloria, (G)**	**Spectra, (S)**	**Gloria+Spectra, (G+S)**	**Brazil, (B)**
Reflectance				
- nr of nonmissing	1746	669	2415	1003
CDOM				
- nr of nonmissing	1120	410	1530	323
- mean ± std	0.8 ± 1.0	3.4 ± 5.4	1.5 ± 3.1	1.7 ± 0.9
- median; q25 - q75	0.5; 0.2 - 0.9	2.0; 1.3 - 3.3	0.8; 0.3 - 1.7	1.6; 1.2 - 2.1
Chla				
- nr of nonmissing	1226	410	1636	708
- mean ± std	41.7 ± 398.2	19.5 ± 26.9	36.1 ± 345.0	22.2 ± 81.7
- median; q25 - q75	9.5; 2.9 - 25.2	12.3; 6.4 - 24.9	10.6; 3.7 - 25.0	9.9; 4.8 - 22.8
TSS				
- nr of nonmissing	1256	410	1666	691
- mean ± std	40.7 ± 156.0	7.8 ± 6.9	32.6 ± 136.2	75.2 ± 180.2
- median; q25 - q75	10.8; 4.2 - 28.8	5.6; 3.2 - 10.7	8.8; 3.7 - 20.6	14.5; 6.4 - 34.9
Secchi depth				
- nr of nonmissing	955	582	1537	1003
- mean ± std	1.6 ± 1.7	1.8 ± 1.3	1.7 ± 1.5	0.7 ± 0.8
- median; q25 - q75	0.9; 0.4 - 2.2	1.4; 0.8 - 2.3	1.2; 0.6 - 2.3	0.5; 0.25 - 0.8

The clustering and classification algorithms are trained based on Sentinel-2 MSI bands format. Based on wavelength-band mappings, presented in Sentinel-2 MSI Spectral Response Functions document, (S2A, 2022, version 3.2)[12] the transformation of

[12]S2 Documents: `https://sentiwiki.copernicus.eu/web/s2-documents` Read: 26.06.2024

the reflectance wavelength data into the corresponding bands is done. The observed wavelength upper limit for our datasets is 900 nm. Sentinel-2 MSI B8 band contains the range between 760 and 907. For Sentinel-2 MSI B8 band the aggregated range is limited to interval from 760 to 900.

The datasets resulted from the preprocessing steps are further used in modeling and evaluation phases. The set of attributes {"S2A_B1", "S2A_B2", "S2A_B3", "S2A_B4", "S2A_B5", "S2A_B6", "S2A_B7", "S2A_B8", "S2A_B8A"} are used in modeling activities. The set of attributes {"cdom", "chla", "tss", "secchi_depth"} are used as the additional information sources in the evaluation and the optimal model selection activities. The attributes together with examples of values and possible data types are presented in Table 2.

Table 2. The formats of datasets after applying the preprocessing steps.

Attribute	**Value**	**Type**
id	{"GID_24", "SID_32", "BID_14"}	{String}
source	{"gloria", "spectra", "brazil"}	{String}
cdom	{2.4, missing}	{Float64, Missing}
chla	{4.19, missing}	{Float64, Missing}
tss	{11.0, missing}	{Float64, Missing}
secchi_depth	{0.7, missing}	{Float64, Missing}
S2A_B1	{0.00641531}	{Float64}
S2A_B2	{0.00901746}	{Float64}
S2A_B3	{0.0151849}	{Float64}
S2A_B4	{0.0169478}	{Float64}
S2A_B5	{0.0139325}	{Float64}
S2A_B6	{0.00465688}	{Float64}
S2A_B7	{0.00467691}	{Float64}
S2A_B8	{0.00376684}	{Float64}
S2A_B8A	{0.0025264}	{Float64}

The resulted distributions of unnormalized and normalized reflectance for each dataset across Sentinel-2 MSI bands are presented in Figure 1. The reflectance curves are formed drawing the imaginary line through the median points of each band. The reflectance curves for Gloria and Spectra datasets are very similar, forming the hill within the range from B1 to B5 and the plateau within the range from B6 to B8A. Brazil data medians variation is less comparing to Gloria and Spectra data. The modest hill within the range from B1 and B5 and the plateau within the range from B6 to B8A are also observed. The overall reflectance signal across all bands is higher for Brazil dataset comparing to Gloria and Spectra datasets.

The samples distribution across the globe[13] is presented in Figure 2. Gloria dataset samples (dark blue) are enriched with Spectra (orange) dataset samples. The vast majority of samples come from Europe and Asia. Significantly less representatives are observed

[13]Created using `https://mobisoftinfotech.com/tools/plot-multiple-points-on-map/` Read: 25.06.2024

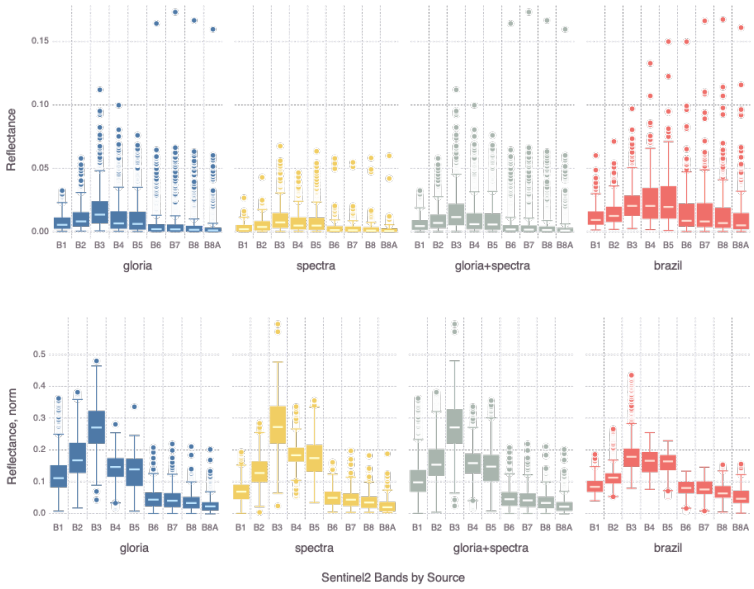


Figure 1. The reflectance and normalized reflectance distribution by Sentinel-2 MSI bands for Gloria (blue), Spectra (yellow), Gloria+Spectra (grey) and Brazil (red) datasets.

from North/South America, Africa and Australia.

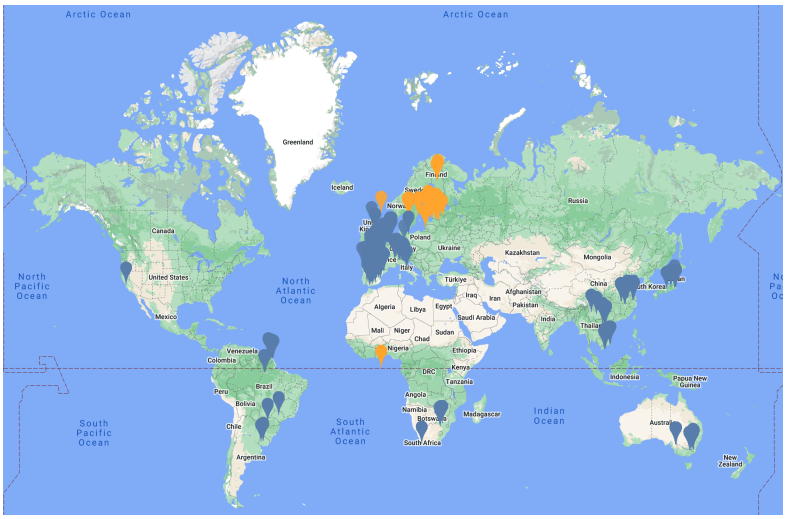


Figure 2. The global distribution of samples selected from Gloria (dark blue markers) and Spectra (orange markers) datasets.

S2A datasets are loaded as nc-extention files (NetCDF[14]). The variables we use are 'bitmask', 'latitude', 'longitude', 'Rw443', 'Rw490', 'Rw560', 'Rw665', 'Rw705', 'Rw740', 'Rw783', 'Rw842', 'Rw865'. The Rw* related attributes are central wavelengths (nm) of bands B1, B2, B3, B4, B5, B6, B7, B8, B8A respectively. The suggestion of the nc-file comments to remove the pixels such that 'bitmask' & 1023 != 0 are followed. All S2A datasets within a particular month are concatenated, replacing all the negative reflectance values for all bands with 'missing', grouping the data by 'latitude', 'longitude' and computing the median for each reflectance band value ('missing' excluded). Finally, for entire dataset any rows containing 'missing' values and duplicates are dropped.

4.2 Methods

This section first describes the modeling approach chosen to perform the optical water classification task. The clustering and classification methods form the backbone of the approach. The description of them are given in separate subsections. Another subsection describes the evaluation techniques applied. Separately presented the technique of getting from Optical Water Types to CDOM-based Optical Water Types and CDOM concentration computing. Finally, the optimal model selection steps are presented.

4.2.1 Modeling approach

The approach coming from Automated Machine Learning (AutoML) field of research is used to perform the optical water classification task. Despite of no single contributor to AutoML, there are several projects that have significant influence on automated learning area. One of the most known is the Auto-WEKA [Thornton et al., 2013] project that proposed the automation to model selection and hyperparameter optimization. The tutorial-level overview of the AutoML methods in general can be seen in [Hutter et al., 2019]. Another comprehensive survey for the state-of-the-art efforts in tackling the CASH problem (Combined Algorithm Selection and Hyper-parameter tuning) is [Elshawi et al., 2019].

The AutoML approach is considered as a good strategy to be applied to the modeling process. Comparing to a single or few models mode, the many models way of the problem solving has a clear advantages.

Many models approach may lead to significant decrease of the ambiguity through the increasing of the amount of possible ways of datasets clustering. It is supposed that different combinations of four main water quality contributors (cdom, chla, tss, secchi depth) may form the same reflectance signal output within the same band. In other words, the many-to-one relationships may be seen. At the same time it is expected that, having the reflectance curve at hand (reflectance signal by many bands), there may be found the output signal that is uniquely described by water quality parameters. In other words,

[14]NetCDF: `https://en.wikipedia.org/wiki/NetCDF` Read: 05.07.2024

there is the one-to-one relationships or at least there is the ability to differentiate between the reflectance curves, significantly decreasing the level of ambiguity.

The efficiency and simplicity of the given approach may form the other advantages. Depending on the task or during the process of investigation, a researcher as a domain expert may wish to operate with many models, where each model may classify the curves based on different combination of features. Some problems may expect the one combinations, but the other may expect the different combinations. The diversity of models significantly increase the speed of investigation. Also, the automated way of models generation may require less technical knowledge and skills.

The steps of proposed approach is the following:

1. Select the number of ***clustering algorithms*** with the set of hyperparameters. Apply the algorithms to initial dataset. The output of each algorithm forms a new dataset with a clusters (OWTs) assigned. Split each dataset into training and testing.

2. Select the number of ***classification algorithms*** with the set of hyperparameters. Apply the algorithms to the train-split of each dataset received on the previous step. The output of each algorithm forms a new dataset with a predicted cluster (OWT) assigned.

3. ***Evaluate*** the modeling results based on test-split received on the first step. Evaluate the modeling results based on entirely new dataset.

4. Get the OWTs to ***CDOM-OWT***s transformed.

5. Select the ***optimal model*** applying the evaluation metrics and domain specific knowledge.

Every step is separately described in the next subsections.

The technical work is done using Julia[15] programming language version 1.10.4 and Python[16] programming language version 3.11.9. The clustering algorithms implementation is taken from Conda.jl[17] package. The AutoGluon[18] solution is used to run the classification algorithms. The link to GitHub repository to the notebooks used are located in appendix I.

4.2.2 Clustering methods

The Clustering algorithms applied together with hyperparameters space are presented in Table 3.

[15]Julia: `https://julialang.org/` Read: 27.06.2024
[16]Python: `https://www.python.org/` Read: 27.06.2024
[17]Conda.jl: `https://github.com/JuliaPy/Conda.jl` Read: 27.06.2024
[18]AutoGluon: `https://auto.gluon.ai/stable/index.html` Read: 27.06.2024

Table 3. Clustering algorithms together with the space of hyperparameters.

Algorithm / Hyperparameter	**Ranges**
Agglomerative clustering	
- n_clusters	5 - 12
- affinity	["euclidean", "l1", "l2", "manhattan", "cosine"]
- linkage	["ward", "complete", "average", "single"]
combinations, total/valid	160/128
Bisecting K-Means	
- n_clusters	5 - 12
- init	["k-means++", "random"]
- bisecting_strategy	["biggest_inertia", "largest_cluster"]
- algorithm	["lloyd", "elkane"]
combinations, total/valid	64/32
K-Means	
- n_clusters	5 - 12
- init	["k-means++", "random"]
- algorithm	["lloyd", "elkane"]
combinations, total/valid	32/16
MiniBatch K-Means	
- n_clusters	5 - 12
- init	["k-means++", "random"]
- batch_size	10:20:200
combinations, total/valid	160/160
Spectral clustering	
- n_clusters	5 - 12
- affinity	["nearest_neighbors", "rbf"]
- n_neighbors	5:5:15
- assign_labels	["kmeans", "discretize", "cluster_qr"]
combinations, total/valid	144/144

Not all the hyperparameters combinations form a valid algorithm's setup. Out of 560 possible combinations, 480 valid combinations are used. The hyperparameters space is formed by referencing the ScikitLearn[19] and MLJScikitLearnInterface.jl clustering manuals[20]

The main selection criterion for the algorithm is a presence of a "number of clusters" among the hyperparameters. Five clustering algorithms that contain the "number of clusters" parameter are selected. They are Agglomerative clustering, Bisecting K-Means, K-Means, MiniBatch K-Means, and Spectral clustering.

Agglomerative clustering is the algorithm from hierarchical clustering family. The mechanics of algorithm is around the idea of tree building, where leaves are treated as the individual items and branches are groups of these items. There are no a single creator. Peter Sneath and Robert R. Sokal [Sneath and Sokal, 1973] can be named among noticeable contributors who developed and spread the hierarchical clustering method.

The main idea behind the *K-Means* algorithm [Lloyd, 1982], [McQueen, 1967] is

[19]ScikitLearn: `https://scikit-learn.org/stable/modules/clustering.html` Read: 27.06.2024

[20]MLJScikitLearnInterface.jl: `https://github.com/JuliaAI/MLJScikitLearnInterface.jl/blob/master/src/models/clustering.jl` Read: 27.06.2024

to iteratively set the points called centroids and assign the items to the nearest one. The *MiniBatch K-Means* [Sculley, 2010] is the variant of the same algorithm with the difference that it uses the mini-batches to reduce the computational time. Referencing the technical manual, it converges faster comparing to K-Means, but the resulted quality may be lower. Another variant is *Bisecting K-Means* algorithm. Referencing the technical manual, the clusters from Bisecting K-Means have the order and create a visible hierarchy.

The mechanics behind the *Spectral clustering* [Shi and Malik, 2000] algorithm is around the graph formation where nodes are mapped to low-dimensional embedding space with further of clusters assignment. Referencing the technical manual, it performs well when the number of clusters is relatively small.

4.2.3 Classification methods

The selected classification methods are a part of the tabular type models of AutoGluon solution [Erickson et al., 2020] that is an open-source AutoML framework. The Python code of the basic setup and methods that are used to classify the data are presented in Figure 3.

```
predictor = TabularPredictor(
  label=label_name, eval_metric='balanced_accuracy',
  path=f'{PATH_SAVE}').fit(
    train_data=train_data, time_limit=2*60, presets='best_quality'
    included_model_types=[
      'GBM', 'CAT', 'XGB', 'RF', 'XT', 'NN_TORCH', 'FASTAI', 'KNN'])
```

Figure 3. AutoGluon TabularPredictor's basic setup.

The predictor is defined together with the evaluation metric that is used to evaluate the result and to select the best model. The fit method also has the proper setup. Inside the fit method the time budget is defined. The time budget means the amount of time that is spent by the fit method for execution. Also, inside the fit method the *presents* property is set. The *presents* property relates to the predefined configurations for various arguments in fit method. The 'best_quality' configuration[21] is chosen that corresponds to the best predictive accuracy. Each configuration has the default hyperparameters search space defined. The 'best_quality' configuration has the hyperparameters space named 'zeroshot'[22], that for each model type has a set of the default AutoGluon-Tabular

[21]Autogluon, 'best_quality' configuration: https://github.com/autogluon/autogluon/blob/master/tabular/src/autogluon/tabular/configs/presets_configs.py Read: 06.08.2024

[22]Autogluon, hyperparameters search space ('zeroshot'): https://github.com/autogluon/autogluon/blob/master/tabular/src/autogluon/tabular/configs/zeroshot/zeroshot_portfolio_2023.py Read: 06.08.2024

hyperparameters. The list of model types is also defined within the fit method under *included_model_types* argument. The algorithms[23] to be used together with the implementation references are presented in Table 4.

Table 4. Classification algorithms, AutoGluon.

Abbreviation	**Algorithm**
CAT	CatBoost model (catboost.ai)
GBM	LightGBM model (lightgbm.readthedocs.io)
KNN	KNearestNeighbors model (scikit-learn)
RF	Random Forest model (scikit-learn)
XGB	XGBoost model (xgboost.readthedocs.io)
XT	Extremely Randomized Trees (scikit-learn)
FASTAI	Neural Network with FastAI backend
NN_TORCH	Neural Network implemented in PyTorch

The *KNearestNeighbors* [Fix and Hodges, 1951] algorithm makes the predictions based on a number of closest items. Instead of relying on one model the *Random Forests* [Cutler et al., 2012] algorithm uses many small decision trees to output the final decision. The *Extremely Randomized Trees* [Geurts et al., 2006] algorithm is the extension of Random Forest algorithm. Similar to Random Forest it is an ensemble learning method with the introduction of extra randomization in trees building process. Another representative of ensemble algorithm is the *XGBoost* [Chen and Guestrin, 2016] algorithm that is a scalable tree boosting system. Another gradient boosting decision tree algorithm is the *LightGBM* [Ke et al., 2017]. It provides additional internal techniques to deal with the large number of data points and features. The *CatBoost* [Prokhorenkova et al., 2017] algorithm introduced the ordering boosting technique.

Also, neural networks classifiers based on *FASTAI* [Howard and Gugger, 2020] and *PyTorch* [Paszke et al., 2019] libraries form two final selections run by AutoGluon framework.

4.2.4 Evaluation techniques

The evaluation step consists of a set of actions that contain the computation of common machine learning metrics.

Each clustering algorithm's output is evaluated using the *silhouette score* [Rousseeuw, 1987] that provides a measure of object's similarity within the cluster. Silhouette score has a range between -1 and 1. A score value near 1 indicates the good matching level of an object to its own cluster.

[23]Autogluon, classification algorithms: `https://auto.gluon.ai/stable/api/autogluon.tabular.models.html` Read: 28.06.2024

Each classification algorithm is evaluated using the *balanced accuracy score*[24] that is a good option for imbalanced datasets and computed as the average of recall received on every separate class.

There is additionally observed a number of evaluation metrics calculated by AutoGluon like *f1 scores*[25] the 'micro' and 'macro' versions. Also, the *features importance analysis* is provided by the same framework.

The results of each classifier are also evaluated using a *confusion matrix*[26].

4.2.5 OWT to CDOM-OWT transformation technique. CDOM concentration computing

The result of clustering/classification modeling is the dataset assigned with the field of clusters. These clusters are treated as Optical Water Types (OWTs). To get the Colored Dissolved Organic Matter based OWT or CDOM-OWT and compute the CDOM relative concentration for each CDOM-OWT, the following steps are done:

1. From training dataset select only the rows having the non-missing values of CDOM.
2. Group the training dataset by OWTs and compute the median of CDOM.
3. Order the OWTs by CDOM median in reverse and create the CDOM-OWT field that has the labels starting from 1 to N (count of unique OWTs). The smallest label corresponds to a highest value of CDOM median. The largest label corresponds to a smallest value of CDOM median.
4. Compute the CDOM relative concentration field that is calculated for each CDOM-OWT as CDOM median divided by CDOM median corresponding to CDOM-OWT label marked as 1.

4.2.6 Optimal model selection

The classifiers' related evaluation metrics' values and the additional information provided by the values of CDOM - Aquatic Environment parameter (AEP) are used to select the optimal model. The description of the optimal model selection together with the general overview of required computations are provided here.

There is the expectation that the evaluation dataset is supported by the CDOM value.

[24]Balanced accuracy score: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html` Read: 28.06.2024

[25]F1 Scores: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html` Read: 28.06.2024

[26]Confusion matrix: `https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html` Read: 28.06.2024

The additional metric that is based on CDOM values called ***CDOM Order Penalty*** is added. The idea behind the metric is the following.

Given the train-split and evaluation datasets with the fields of CDOM median grouped by CDOM-OWT. Two datasets are joined together by CDOM-OWT and ordered by train-split CDOM-OWT in reverse. This is equivalent to the reverse order by train-split CDOM median. If each next evaluation dataset related CDOM-OWT label is greater than the previous label (meaning that each next CDOM median is less than previous), then no penalty is assigned. If there are mismatches, then each mismatch adds the penalty amount equal to 1 divided by the total number of unique labels observed in train-split.

Now, the step of the ***optimal model selection*** is the following:

1. Order all classifier characteristics by Number of clusters (descending), Balanced score on training or evaluation dataset (descending), CDOM Order Penalty (ascending).

2. Visualize the train-split and evaluation dataset using boxplots where x-axis contains the CDOM-OWTs and y-axis contains the CDOM median values.

3. Starting from the first model in the ordered list, select the most suitable model taking into consideration the visual output.

5 Results

This section presents the results of the work.

First, the clustering algorithms' performance is observed. This is done using two ways. The performance is estimated directly using a silhouette score and indirectly based on the results of classifiers applied to the labeled datasets (the output of the clustering algorithms).

Then, the results of optimal model selection step based on classifiers' performance on Gloria+Spectra dataset are presented. The same is done based on classifiers' performance on Gloria+Spectra and Brazil datasets.

The set of evaluations activities for both selection options includes the observation of Balanced accuracy, Macro F1 and Micro F1 scores, applying of CDOM order penalty computation, the investigation of results of Confusion matrix and Feature importance, and the visual comparison of train and test reflectance curves by CDOM-OWTs.

Both options produce two main outputs. The first output is the optimal classifier that is used to classify a Sentinel-2 MSI formatted reflectance data. The second output is concentration of CDOM for each CDOM-OWT. Both outputs will be further used in getting the CDOM relative concentration levels distribution on the location maps.

5.1 Clustering algorithms overall performance evaluation

This section evaluates the clustering algorithms' overall performance.

The variety of clustered datasets were prepared by a number of clustering algorithms. How well the samples are differentiated between clusters is evaluated using a silhouette score. The statistics of the silhouette score grouped by models is presented in the first part of Table 5.

The maximum silhouette score is achieved by Agglomerative clustering (0.408), the next one belongs to K-Means (0.308). The maximum median results we see for K-means (0.288), the next one belongs to MiniBatch K-Means (0.263). Based on silhouette score observed it is stated that the differentiation between clusters are far from perfect. Silhouette scores are much less than 1.0.

There is also observed the performance of clustering algorithms indirectly by applying a classification algorithms to clustered dataset created by clustering algorithms. The total number of classifiers applied to clustered datasets are presented in column 'count'. Gloria+Spectra dataset is divided into train and test splits and the performance of the classifier is evaluated based on balanced accuracy. The statistics of the balanced accuracy grouped by clustering models is presented in the last two parts of Table 5.

The best performance on train-split is achieved on datasets labeled by Agglomerative clustering (1.000). The next best result belongs to Spectral clustering (0.994). The best median performance is done on datasets labeled by Agglomerative clustering (0.978). The next best median performance belongs to K-Means (0.976).

Table 5. The clustering algorithms overall performance. Descriptive statistics.

Metric / Model	count	mean	std	min	q25	median	q75	max
Silhouette score								
- Agglomerative clust.	128	0.218	0.074	-0.029	0.202	0.231	0.261	**0.408**
- Bisecting K-means	32	0.222	0.036	0.163	0.180	0.232	0.248	0.276
- K-Means	16	0.289	**0.010**	0.277	0.282	**0.288**	0.293	**0.308**
- MiniBatch K-Means	160	0.263	**0.019**	0.201	0.249	**0.263**	0.275	0.302
- Spectral clustering	144	0.201	0.072	0.042	0.137	0.232	0.258	0.303
Balanced accuracy, G+S train dataset								
- Agglomerative clust.	128	0.882	0.263	0.111	0.961	**0.978**	0.991	**1.000**
- Bisecting K-means	32	0.964	0.013	0.919	0.958	0.963	0.972	0.986
- K-Means	16	0.976	**0.005**	0.968	0.971	**0.976**	0.982	0.983
- MiniBatch K-Means	160	0.969	**0.010**	0.940	0.964	0.969	0.978	**0.989**
- Spectral clustering	144	0.965	0.023	0.892	0.957	0.971	0.982	**0.994**
Balanced accuracy, G+S test dataset								
- Agglomerative clust.	128	**0.673**	0.196	0.200	0.621	**0.716**	0.796	0.990
- Bisecting K-means	32	0.960	0.016	0.933	0.946	0.959	0.975	0.987
- K-Means	16	0.973	0.012	0.945	0.967	0.974	0.979	**0.993**
- MiniBatch K-Means	160	0.965	0.013	0.924	0.958	0.967	0.976	**0.992**
- Spectral clustering	144	0.959	0.025	0.867	0.953	0.965	0.977	**0.993**

The best performance on test-split is achieved on datasets labeled by K-Means (0.993) and Spectral clustering (0.933). The best median performance is done on datasets labeled by K-Means (0.974) and MiniBatch K-Means (0.967). It is also noticed that despite of the best performance on train dataset, the Agglomerative clustering performs the worse on test dataset (0.716).

5.2 Optimal model selection based on classifiers performance on Gloria+Spectra dataset

Previous section provided the overall performance information for the clustering algorithms. In this section the optimal classifier trained and tested on Gloria+Spectra dataset is selected. Applying the optimal model selection steps described in section 4.2, the classifier with the characteristics presented in Table 6 is chosen.

It is planed to get the location map visualization based on eight CDOM-OWT labels. The classifier is selected out of group that has a number of clusters (OWT) equals to eight. The best performing classifiers are not limited to group eight. Each group has a number of well performing classifiers. A researcher is free to choose the group that responds best to the given task.

The classifier is trained on dataset labeled by K-Means clustering algorithm. The value of Silhouette score is 0.308 that is greater than median (0.288). The Balanced accuracy for train- and test-splits are 0.979 and 0.963 respectively. The model's Macro F1-Score and Micro F1-Score have also the high values of 0.970 and 0.971 respectively.

CDOM order penalty is 0.0, telling us that the set of clusters by CDOM medians has

Table 6. The selected model's general characteristics for Gloria+Spectra dataset.

Metric	**Value**
- Model name	**K-Means**
- Model ID	kmeans_2
- Classifier	WeightedEnsemble_L3
- Number of clusters (OWT)	**8**
- Silhouette score	0.308
- Balanced accuracy, train-split (G+S)	0.979
- Balanced accuracy, test-split (G+S)	**0.963**
- Macro F1-Score	0.970
- Micro F1-Score	0.971
- CDOM order penalty	**0.000**
- Chla order penalty	0.125
- TSS order penalty	0.125
- Secchi Depth order penalty	0.000

the same order for train and test outputs. There is also computed the Chla/TSS/Secchi depth order penalties using the same method, but replacing the CDOM with appropriate aquatic water parameter. This is done for informative purposes only.

Visually, the CDOM order penalty based performance is observed in Figure 4.

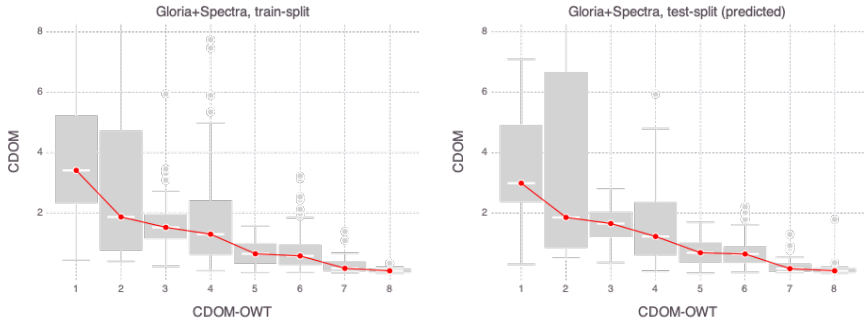


Figure 4. The selected model's performance evaluated by CDOM order penalty on Gloria+Spectra dataset. The predicted CDOM-OWTs on test-split are supported by CDOM median values (m^{-1}). Every next predicted CDOM-OWT has the lower CDOM median value compared to previous (declining red curve). This is the same behaviour as is seen for train-split.

On the left side of the plot there is the distribution of CDOM values by true CDOM-OWT observed in Gloria+Spectra train-split. On the right side there is the distribution of CDOM values by predicted CDOM-OWT observed in Gloria+Spectra test-split. There is observed that the predictions of the classifier are also supported by order of CDOM medians. The highest value of CDOM for both sides is related to CDOM-OWT equals to 1. The next highest value is to CDOM-OWT equals to 2 etc. The CDOM-OWT

equals to 8 is marked by the lowest CDOM median. The observation of the same order is considered as an important indicator of the selected model's performance evaluation.

One additional way of the selected model evaluation is based on the reflectance curves visual comparison by CDOM-OWT. The normalized reflectance curves match is observed in Figure 5. The reflectance curves match is also presented in Figure 6.

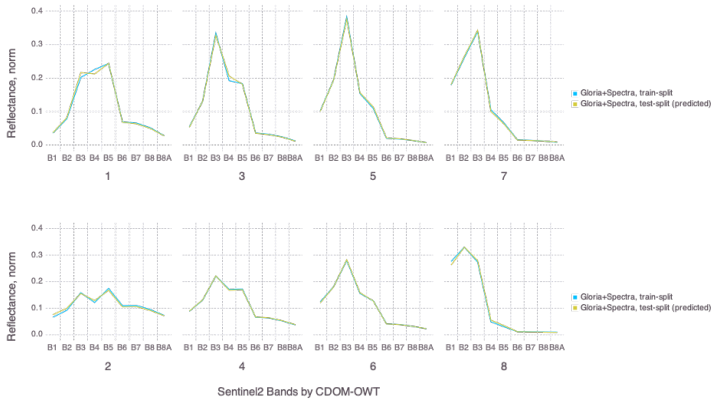


Figure 5. The selected model's performance evaluated by normalized reflectance curves matching for Gloria+Spectra dataset. The classifier is able almost perfectly to assign a test-split curve with the correct CDOM-OWT.

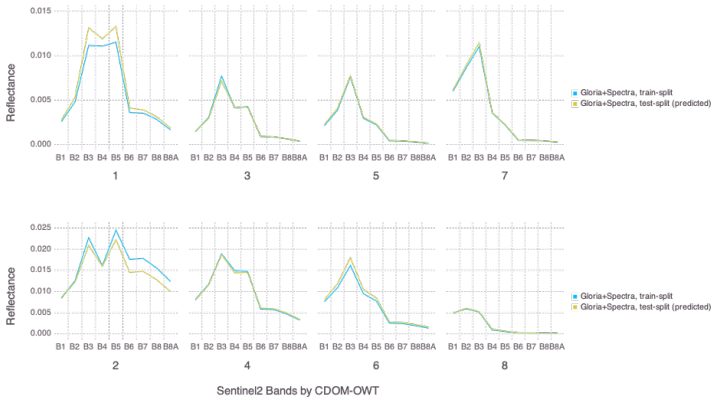


Figure 6. The selected model's performance evaluated by reflectance curves matching for Gloria+Spectra dataset. The classifier is able almost perfectly to assign a test-split curve with the correct CDOM-OWT. Small differences are observed for CDOM-OWTs 1 and 2.

The classifier is trained on input of normalized reflectance values. In Figure 5 the selected model performance is seen. The classifier is able almost perfectly to assign a curve with the correct CDOM-OWT. The model uses the relative proportions of reflectance values between the bands as the main differentiation drivers. The true reflectance curves matching output (Figure 6) shows also the match not only between the relative proportions, but also the match between the reflectance values. Almost perfect match is observed for clusters 3, 4, 5, 6, 7, 8. Near perfect match is seen for clusters 1 and 2. For clusters 1 and 2, the small differences in reflectance values are seen for some number of Sentinel-2 MSI bands. Despite of the differences, the overall shape of the train-curve is followed by the test-curve. The true reflectance values match is manly explained by the fact that train-split and test-split came from the same dataset's distribution.

The confusion matrix output helps to observe how well the selected model predicts the CDOM-OWTs. Looking at Figure 7, it is seen that the model is able to correctly predict the majority of labels with a very few mistakes.

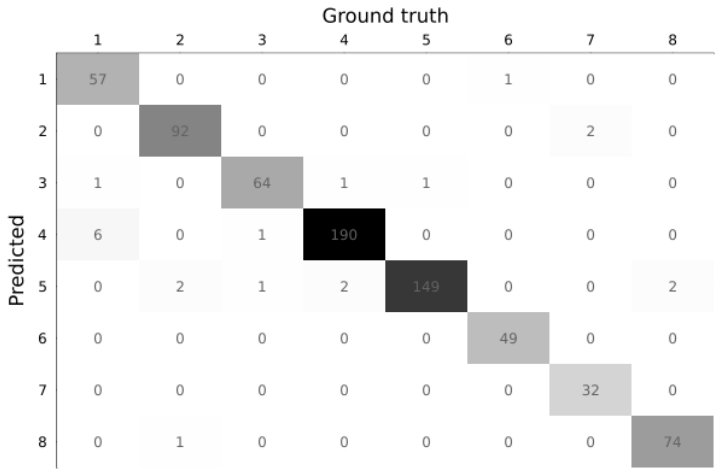


Figure 7. The selected model's performance evaluated using the confusion matrix for Gloria+Spectra dataset.

A CDOM-OWT and a cluster terms are used interchangeably. The most confusion combination of CDOM-OWTs is 1 and 4. Sometimes the model cannot correctly predict cluster 1, mixing it with cluster 4 (6 mistakes). One other pattern of mistakes observed is that the model instead of cluster 2 (2 mistakes), cluster 4 (2 mistakes) and cluster 8 (2 mistakes) predicts the cluster 5. Also, cluster 7 sometimes is predicted as cluster 2 (2 mistakes).

All features, except S2A_B8A, are considered statistically significant, having the p-values less than 0.001. The most important features are S2A_B3, S2A_B2 and S2A_B1.

These bands support the correspondence to high values of CDOM that was described in section 2. In other words, the classification decision is mostly driven by CDOM related features. The less important features are S2A_B8A and S2A_B8. The information related to feature importance indication is presented in Table 7.

Table 7. Selected model: feature importance, Gloria+Spectra train-split.

Feature	**Importance**	**P-value**
S2A_B1	**0.229**	1.74857e-5
S2A_B2	**0.257**	2.01764e-6
S2A_B3	**0.309**	1.39935e-6
S2A_B4	0.133	2.60306e-5
S2A_B5	0.198	8.16122e-7
S2A_B6	0.071	8.85386e-6
S2A_B7	0.070	1.10135e-5
S2A_B8	0.026	**0.00074245**
S2A_B8A	0.012	**0.00865935**

CDOM concentration output by model's OWT/CDOM-OWT for Gloria+Spectra train-split and test-split (predicted) is presented in Table 8.

Table 8. CDOM concentration by OWT/CDOM-OWT for Gloria+Spectra train- and test-splits.

OWT	**CDOM-OWT**	**Samples, G+S, train-split**	**CDOM, G+S, train-split**	**CDOM rel. conc., G+S, train-split**	**Samples, G+S, test-split (predicted)**	**CDOM, G+S, test-split (predicted)**	**CDOM conc., G+S, test-split (predicted)**
1	1	91	3.41	1.000	43	2.99	1.000
6	2	59	1.87	0.547	28	1.86	0.622
3	3	99	1.52	0.446	40	1.66	0.554
4	4	287	1.29	0.379	106	1.23	0.411
8	5	119	0.65	0.189	51	0.69	0.231
5	6	220	0.58	0.169	105	0.65	0.216
2	7	142	0.16	0.048	60	0.16	0.055
7	8	57	0.08	0.024	23	0.09	0.031

The output of this section is the optimal classifier that may be used as a model to assign CDOM-OWT labels to Sentinel-2 MSI reflectance dataset. To analyze the levels of CDOM distribution, each Sentinel-2 MSI reflectance sample should be assigned with CDOM value. Having the water body geographic coordinates available, the CDOM relative concentration values could be marked with appropriate color. Colored data together with coordinates is used to visualize the shape of water body filling in with CDOM relative concentration levels.

This section observed, how to make the optimal model selection based on test-split, where the OWTs were available. There also were received the CDOM-OWT/CDOM concentration table as output. Next section explains, how to select the optimal classifier

and computes CDOM-OWT/CDOM concentration table for dataset that does not have the related OWTs assigned, but do have the CDOM values.

5.3 Optimal model selection based on classifiers performance on Gloria+Spectra and Brazil datasets

In the previous section, to select the optimal classifier, the test-split from the same dataset was used. The train-split and test-split had the same distribution. In this section the optimal model is selected, using the dataset coming from a different distribution (Brazil dataset). Brazil dataset also has the aquatic environmental parameters like CDOM, Chla, TSS and Secchi depth. The same optimal model selection steps described in section 4.2 are followed. In this case the classifier with the characteristics presented in Table 9 is chosen.

Table 9. The selected model's general characteristics for Gloria+Spectra and Brazil datasets.

Metric	**Value**
- Model name	**K-Means**
- Model ID	kmeans_32
- Classifier (OWT)	WeightedEnsemble_L3
- Number of clusters	**8**
- Silhouette score	0.308
- Balanced accuracy, train-split (G+S)	0.976
- Balanced accuracy, test-split (G+S)	**0.974**
- Macro F1-Score	0.972
- Micro F1-Score	0.975
- CDOM order penalty	**0.000**
- Chla order penalty	0.286
- TSS order penalty	0.286
- Secchi Depth order penalty	0.143

As was told in previous section, the aim is to get the location map visualization based on eight CDOM-OWT levels. Again, the classifier out of group that has a number of clusters (OWT) equals to eight was selected. The best performing classifiers are not limited to group eight. Each group has a number of well performing classifiers as well. A researcher is free to choose the group that corresponds best to the task under investigation.

The classifier is trained on dataset labeled by K-Means clustering algorithms. The value of Silhouette score is 0.308 that is greater than median (0.288). Comparing to Gloria+Spectra test-split, Brazil dataset does not have true clusters (OWT) assigned. This means that there is no possibility to compute the evaluation metrics for Brazil dataset. The selection of optimal model here is mostly done based on Balanced accuracy value calculated on Gloria+Spectra train-split (0.976) and CDOM order penalty (0.0). CDOM

order penalty tells that the set of clusters by CDOM meadians, has the same order for Gloria+Spectra train-split and Brazil dataset.

For informative purposes there are also computed the Balanced accuracy, Macro F1-Score and Micro F1-Score on Gloria+Spectra test-split that are respectively equal to 0.974, 0.972 and 0.975. The Chla/TSS/Secchi depth order penalties are also computed using the same method, but replacing the CDOM with appropriate aquatic water parameter.

Visually, the CDOM order penalty based performance is observed in Figure 8.

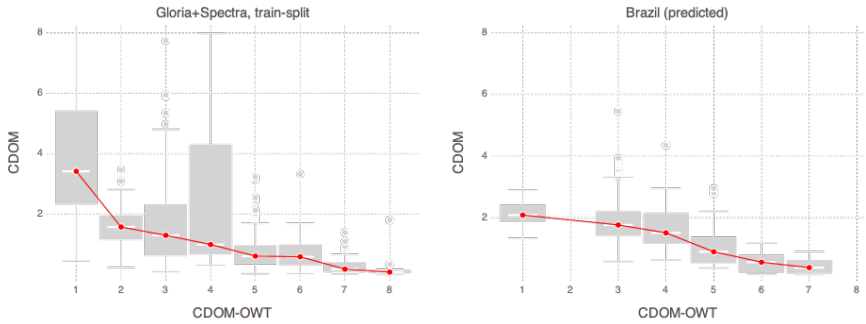


Figure 8. The selected model's performance evaluated by CDOM order penalty on Gloria+Spectra and Brazil datasets. The predicted CDOM-OWTs on Brazil dataset are supported by CDOM median values (m^{-1}). Every next predicted CDOM-OWT has the lower CDOM median value compared to previous (declining red curve). This is the same behaviour as is seen for Gloria+Spectra train-split.

On the left side of the plot there is the distribution of CDOM values by true CDOM-OWT observed in Gloria+Spectra train-split. On the right side there is the distribution of CDOM values by predicted CDOM-OWT observed in Brazil dataset. There is observed that the predictions of the classifier are also supported by order of CDOM medians. The highest value of CDOM for both sides is related to CDOM-OWT equals to 1. The next highest value is to CDOM-OWT equals to 3. We do not have CDOM-OWT equals to 2 representatives in Brazil dataset. The CDOM-OWT equals to 7 is marked by the lowest CDOM median. CDOM-OWT representatives equals to 8 are not presented in Brazil dataset. The observation of the same order is considered as an important indicator of the selected model's performance evaluation.

The additional way of the selected model evaluation is based on the reflectance curves visual comparison by CDOM-OWT. The normalized reflectance curves match is observed in Figure 9. The reflectance curves match is presented in Figure 10.

The classifier is trained on input of normalized reflectance values. In Figure 9 the selected model performance is seen. The classifier is sufficiently able to assign a curve with the correct CDOM-OWT. The model uses the relative proportions of reflectance

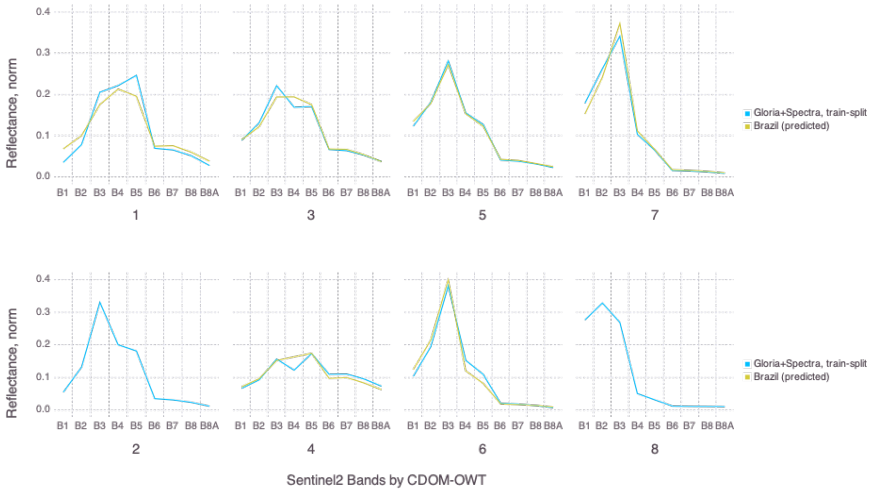


Figure 9. The selected model's performance evaluated by normalized reflectance curves matching for Gloria+Spectra and Brazil datasets. The classifier is able to assign Brazil curve with the correct CDOM-OWT.

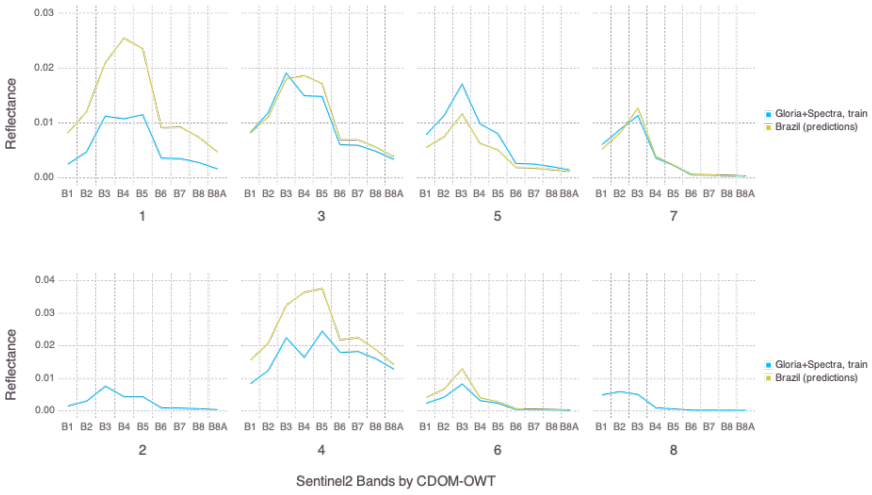


Figure 10. The selected model's performance by reflectance curves matching for Gloria+Spectra and Brazil datasets. The classifier is able to assign Brazil curve with the correct CDOM-OWT for types 6 and 7. There is not seen the perfect match in values, but only the match by the relative proportions (shape) for types 1, 3, 4 and 5.

values between the bands as the main differentiation drivers. The true reflectance curves matching output (Figure 10) also shows the match between the reflectance values for CDOM-OWTs equal to 6 (except B3 band) and 7. For the reflectance curves related to CDOM-OWTs 1, 3, 4 and 5 there is not seen the perfect match in values, but only the match by the relative proportions (shape).

The true CDOM-OWTs for Brazil dataset are not available, but these values are available for Gloria+Spectra test-split. Based on confusion matrix output, Figure 11, the selected classifier predictions quality can be evaluated based on Gloria-Spectra test-split.

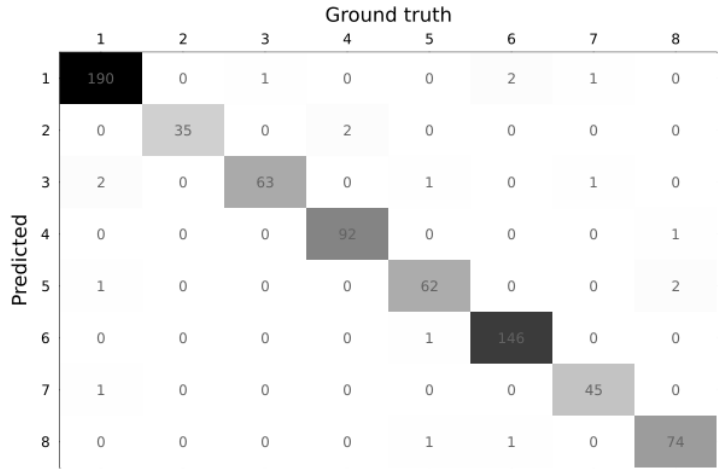


Figure 11. The selected model's performance evaluated by confusion matrix for Gloria+Spectra dataset.

A CDOM-OWT and a cluster terms are used interchangeably. The classifier is able to correctly predict the majority of CDOM-OWTs with very few mistakes. Sometimes the model cannot correctly predict clusters 1, 4, 6 and 8 mixing them respectively with clusters 3, 2, 1 and 5.

All features are considered statistically significant, having the p-values less than 0.001. The most important features are S2A_B3, S2A_B2 and S2A_B1. These bands support the correspondence to high values of CDOM that was described in section 2. In other words, the classification decision is mostly driven by CDOM related features. The less important features are S2A_B8, S2A_B8A and S2A_B7. The information related to feature importance indication is presented in Table 10.

CDOM concentration output by model's OWT/CDOM-OWT for Gloria+Spectra train-split and Brazil (predicted) is presented in Table 11.

The output of this section is the optimal model selected that may also be used as a classifier to assign CDOM-OWT labels to Sentinel-2 MSI reflectance dataset. To analyze the levels of CDOM distribution, each Sentinel-2 MSI reflectance sample should be

Table 10. The selected model's feature importance for Gloria+Spectra train-split.

Feature	Importance	P-value
S2A_B1	**0.236**	2.01501e-6
S2A_B2	**0.258**	4.83161e-8
S2A_B3	**0.316**	1.57237e-7
S2A_B4	0.135	3.80227e-6
S2A_B5	0.211	2.38797e-7
S2A_B6	0.045	4.03408e-5
S2A_B7	0.037	6.04632e-5
S2A_B8	0.023	**0.00064474**
S2A_B8A	0.024	**0.00092497**

Table 11. CDOM concentration by OWT/CDOM-OWT for Gloria+Spectra train-split and Brazil (predicted) dataset.

OWT	CDOM-OWT	Samples, G+S, train-split	CDOM, G+S, train-split	CDOM rel. conc., G+S, train-split	Samples, B (predicted)	CDOM, B (predicted)	CDOM conc., B (predicted)
3	1	95	3.41	1.000	21	2.08	1.000
5	2	94	1.57	0.459	-	-	-
1	3	275	1.29	0.378	198	1.76	0.846
7	4	58	0.99	0.289	32	1.50	0.722
6	5	228	0.60	0.176	35	0.89	0.428
8	6	123	0.58	0.169	17	0.55	0.264
4	7	138	0.17	0.049	20	0.37	0.179
2	8	56	0.07	0.021	-	-	-

assigned with CDOM value. Having the water body geographic coordinates available, the CDOM relative concentration values could be marked with appropriate color. Colored data together with coordinates is used to visualize the shape of water body filling in with CDOM relative concentration levels.

This section observed, how to make the optimal model selection based on a new dataset that does not have any OWTs assigned. The CDOM-OWT/CDOM concentration table was also received as output. Next, will be shown how to use the selected classifier and CDOM-OWT/CDOM concentration table outputs to plot the CDOM relative concentration distribution on a location map.

5.4 Visualization of CDOM relative concentraion distribution on the location maps

In this section, the selected model and CDOM-OWT/CDOM concentration table received in section 5.2 is applied to classify S2A dataset (section 4.1.4) and plot the CDOM relative concentration distribution on a location map (lake Võrtsjärv, Estonia). The plotting for each month of the available year from 2021 to 2024 is done.

The visualization and evaluation activities for the subset related to the period of

2021 year, August is presented here. Gloria+Spectra and S2A datasets reflectance and normalized reflectance distributions comparison is presented in Figure 12.

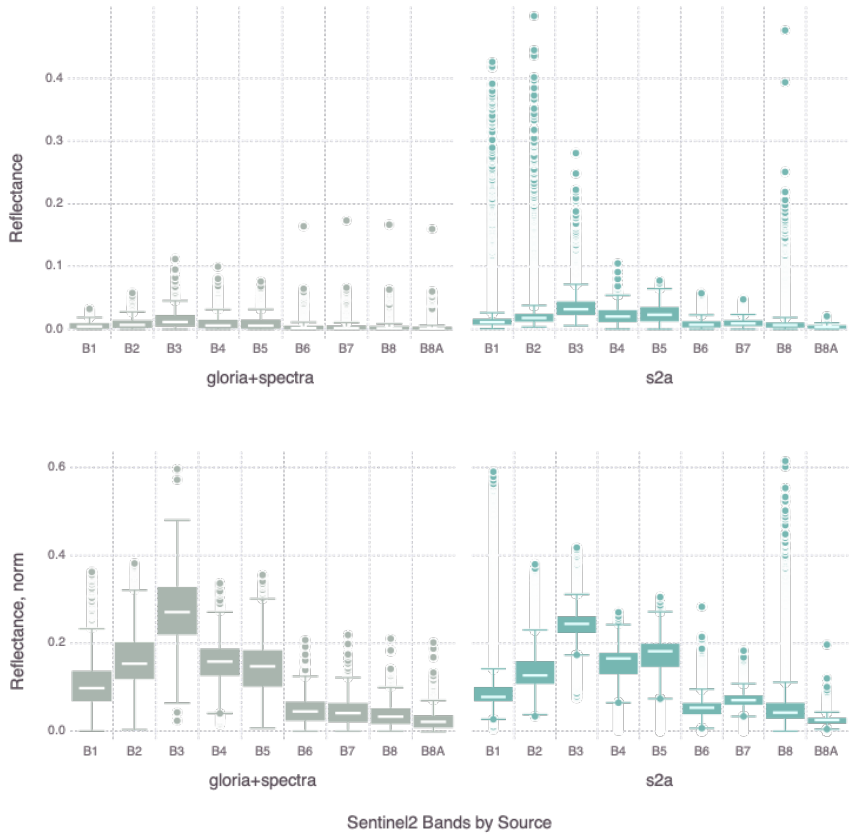


Figure 12. The reflectance and normalized reflectance distribution by Sentinel-2 MSI bands for Gloria+Spectra (grey) and S2A (green, Võrtsjarv, Sentinel-2 MSI) datasets.

The data distributions are almost similar. S2A dataset has a little bit higher the reflectance values compared to Gloria+Spectra dataset (upper part of the plot). High reflectance values in B2 (blue range) are not correspond to true values that observed for lake Võrtsjärv. The differences are posed due to atmospheric correction error. The relative proportions (normalized reflectance) are also similar, having negligible differences in median values of Sentinel-2 MSI B5 and B7 bands. The classifier trained on Gloria+Spectra train-split may be applied to S2A dataset to get CDOM-OWT classification and CDOM concentration outputs.

The conclusion is also supported by the reflectance curves visual comparison by CDOM-OWT. The normalized reflectance curves match is seen in Figure 13. The

reflectance curves match is observed in Figure 14.

The classifier is trained on input of normalized reflectance values. In Figure 13 is observed that the classifier is able to assign a curve with the CDOM-OWT based on its shape characteristics. It is seen that the S2A curves are distributed between CDOM-OWTs following the relative proportions. The true reflectance curves matching output (Figure 12) confirming the shape match, but not the reflectance values.

The match between the reflectance values is not important for the given task of CDOM relative concentration plotting. The most important indicator of the relevance of CDOM-OWTs assignment is that the normalized reflectance curves have the very high match. The matching between the relative proportions tells that the curves are sufficiently distributed between CDOM-OWTs based on the ability of the combinations of aquatic environmental parameters to reflect the light within the different spectral bands.

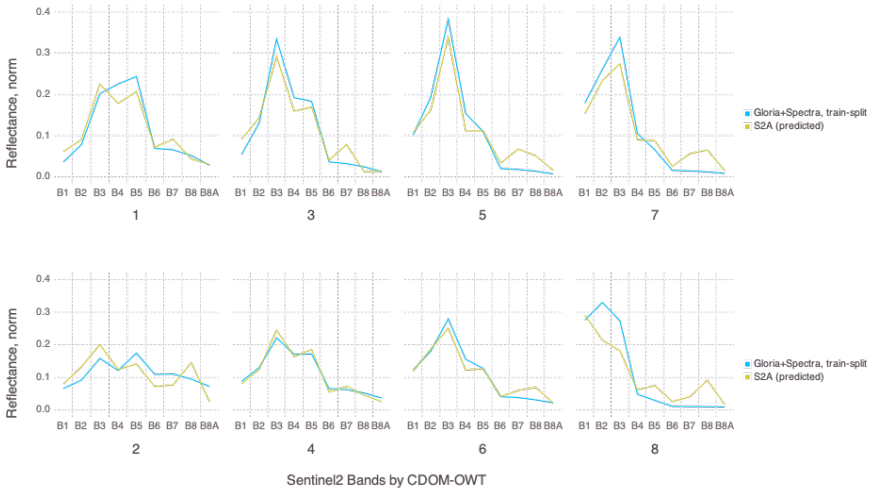


Figure 13. The selected model's performance evaluated by normalized reflectance curves matching for Gloria+Spectra and S2A datasets. The classifier is able to assign S2A curve with the CDOM-OWT based on its shape characteristics.

S2A dataset is received from satellite. The CDOM values are not provided. For CDOM distribution plotting the CDOM concentration values coming from the selected classifier are used. The colors to CDOM concentration values are assigned based on white-brown gradient color palette, where white color corresponds to 0-valued CDOM relative concentration and brown color corresponds to 1-valued CDOM relative concentration.

CDOM concentration output by model's OWT/CDOM-OWT for Gloria+Spectra train-split and S2A (predicted) with related colors assigned is presented in Table 12.

The CDOM-OWTs can be divided into four groups based on CDOM relative concentration values. The highest CDOM relative concentration belongs to CDOM-OWT

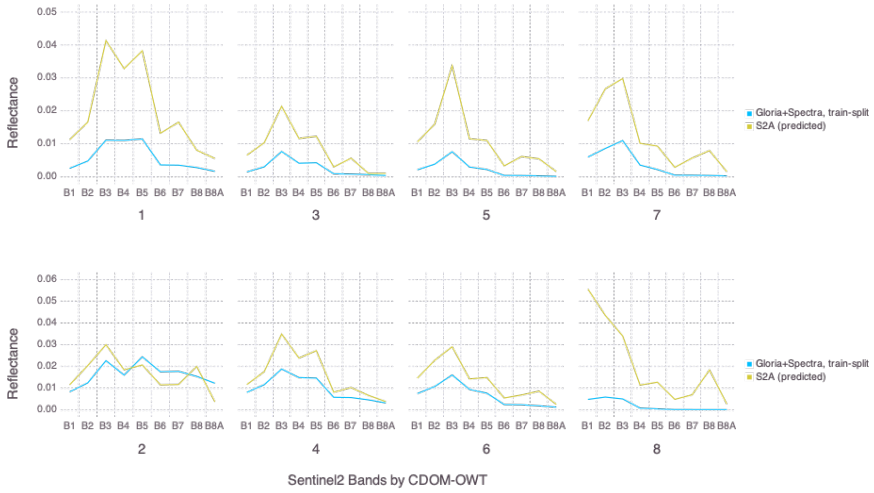


Figure 14. The selected model's performance evaluated by reflectance curves matching for Gloria+Spectra and S2A datasets. The match is confirmed by shape, but not by the reflectance values.

equals to 1. The next high values of 54.7%, 44.6% and 37.9% respectively correspond to CDOM-OWTs equal to 2, 3 and 4. Next group has the values of 18.9% (CDOM-OWT 5) and 16.9% (CDOM-OWT 6). The lowest relative concentration has CDOM-OWTs equal to 7 (4.8%) and 8 (2.4%).

Table 12. CDOM concentration by OWT/CDOM-OWT for Gloria+Spectra train-split and S2A (predicted) dataset.

OWT	**CDOM-OWT**	**Samples, G+S, train-split**	**CDOM, G+S, train-split**	**CDOM conc., G+S, train-split**	**Samples, S2A (predicted)**	**CDOM conc., B (predicted)**
1	1	91	3.41	1.000	9471	#A52A2A
6	2	59	1.87	0.547	1548	#CE8A8A
3	3	99	1.52	0.446	1882	#D7A0A0
4	4	287	1.29	0.379	37800	#DDAEAE
8	5	119	0.65	0.189	144	#EED7D7
5	6	220	0.58	0.169	5848	#F0DBDB
2	7	142	0.16	0.048	735	#FBF5F5
7	8	57	0.08	0.024	1704	#FDFAFA

Having S2A dataset classified and each sample assigned with the color that corresponds to CDOM relative concentration level, the data on a location map are plotted. The result for lake Võrtsjärv, Estonia for the period of August 2021 is seen in Figure 15.

Following the same steps observed, the same output for S2A data entire period is

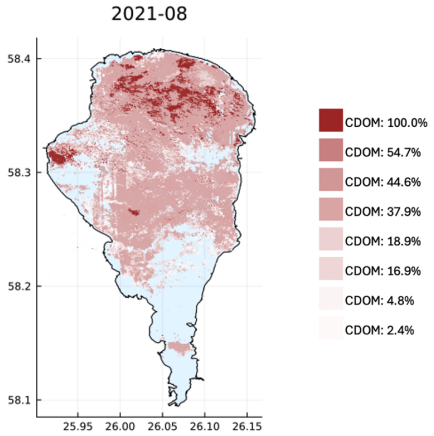


Figure 15. CDOM relative concentration levels of lake Võrtsjärv for the period of 2021/08.

received. The 2021, 2022, 2023 and 2024 CDOM relative concentration levels dynamics for lake Võrtsjärv is seen in respective appendices II, III, IV and V.

The visual outputs received are supported by the collected feedback from the expert. April and October plots indeed should contain the higher relative concentration of CDOM compared to the periods in between. Also, the south side of the lake generally has the lower CDOM relative concentration due to the river's influence.

6 Discussion

Comparing to previous research, the AutoML solution is proposed. Many clustering algorithms and many classifiers were trained. A researcher got an option to use them all, selecting the optimal one based on the needs of the task. Also, there were the flexible limits set to the number of Optical Water Types to get. Instead, the range from 5 to 12 was used, providing a researcher the option to make a selection of the most appropriate.

Some discussion points to notice are provided below.

The selection of the optimal model for Sentinel-2 MSI dataset was mostly driven by CDOM order penalty computed on Gloria+Spectra test-split. The first presented option was used, because the CDOM values for Sentinel-2 MSI dataset were not available. If CDOM values are available, it is recommended to select the optimal model, using CDOM order penalty of the target dataset. This approach was shown on Brazil dataset separately.

Another point to notice is related to the evaluation technique based on reflectance curves matching. The high predictive characteristics of the selected model based on normalized reflectance curves matching results were confirmed. There was not used just the reflectance version, because there was not a goal to make the true CDOM predictions, but only the predictions of the CDOM relative concentration levels. The goal was to differentiate between the CDOM concentrations expressed in percentages. The high match between the relative proportions (normalized reflectance curves) was the indicator of the goal's fulfilling.

There was not the goal to predict the true CDOM levels. The CDOM concentration levels were only estimated as the percentage from the maximum median value related to the first cluster (ascending order by CDOM median). Therefore, the location maps visual comparison between multiple water bodies are possible only, if the same classifier and the same CDOM concentration table output are used.

7 Conclusion

The goal of the thesis was to contribute to the field of Optical Water Types classification, developing and implementing the CDOM based Optical Water Types (CDOM-OWT) classification approach.

The progress was divided into the layers. The first layer presented the clustering methods that were applied to dataset. The clustering algorithms were organized by types. Each type run multiple times based on the unique set of hyperparameters. The clustered dataset results were passed to the next layer - the layer of classification. The classification layer presented the multiple classification methods that were applied to clustered datasets. The output of the methods from both layers were evaluated.

The evaluation layer provided the basis for optimal classifier selection. Two selection options were presented. The first option was based on a test-split dataset that contained OWT assigned. The second option was based on a new dataset that did not contain OWT assigned. Both options used a newly created CDOM order penalty computation to get the optimal classifier together with the CDOM concentration table containing the OWT to CDOM-OWT mapping. The evaluation step confirmed the hypothesis that the reflectance measurements observed under different light spectrum bands in general is not posing the problem of ambiguity and could be clustered based on their unique characteristics.

Having the optimal classifier and CDOM concentration table at hand, there also was presented the application layer, where the Sentinel-2 MSI dataset was classified and CDOM relative concentration was plotted to the location maps to visually follow the CDOM dynamics by time periods.

The work done presents the important technical solution that could be added to the toolset of CDOM relative concentration dynamics monitoring to be able in timely manner to mitigate the risks posed by the process of water brownification.

The further development may include the creation of other versions of Optical Water Types that are based on Chlorophyll A, Total Suspended Matters/Solids or Secchi depth values. Also, to automate the monitoring process to near real-time, it is considered the adaptation of the developed approach to the deployment requirements posed by the cloud computing services.

8 Acknowledgements

I would like to thank my thesis supervisors, Dr. Krista Alikas and Dr. Radwa El Shawi, who provided me with great support and valuable expertise. Also, I would like to acknowledge Daniel Andrade Maciel, Claudio Barbosa, Edson Freirefs from instrumentation Laboratory for Aquatic Systems (LabISA), INPE - National Institute for Space Research, São Jose dos Campos- Brazil, for providing the additional dataset for models' evaluation.

References

[Blanchet et al., 2022] Blanchet, C. C., Arzel, C., Davranche, A., Kahilainen, K. K., Secondi, J., Taipale, S., Lindberg, H., Loehr, J., Manninen-Johansen, S., Sundell, J., Maanan, M., and Nummi, P. (2022). Ecology and extent of freshwater browning - what we know and what should be studied next in the context of global change. Science of the Total Environment 812 152420. https://doi.org/10.1016/j.scitotenv.2021.152420.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. University of Washington. https://doi.org/10.48550/arXiv.1603.02754.

[Cutler et al., 2012] Cutler, A., Stevens, J. R., and Cutler, D. R. (2012). Random forests. Ensemble Machine Learning. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9326-7_5.

[da Silva et al., 2021] da Silva, E. F. F., de Moraes Novo, E. M. L., de Lucia Lobo, F., Barbosa, C. C. F., Cairo, C. T., Noernberg, M. A., and da Silva Rotta, L. H. (2021). A machine learning approach for monitoring brazilian optical water types using sentinel-2 msi. Remote Sensing Applications: Society and Environment 23 100577. https://doi.org/10.1016/j.rsase.2021.100577.

[da Silva et al., 2020] da Silva, E. F. F., de Moraes Novo1, E. M. L., de Lucia Lobo, F., Barbosa3, C. C. F., Noernberg, M. A., da Silva Rotta, L. H., Cairo, C. T., Maciel, D. A., and Júnior, R. F. (2020). Optical water types found in brazilian waters. Limnology 22(1). https://doi.org/10.1007/s10201-020-00633-z.

[Elshawi et al., 2019] Elshawi, R., Maher, M., and Sakr, S. (2019). Automated machine learning: State-of-the-art and open challenges. https://doi.org/10.48550/arXiv.1906.02287.

[Erickson et al., 2020] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. https://doi.org/10.48550/arXiv.2003.06505.

[Fix and Hodges, 1951] Fix, E. and Hodges, J. L. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field.

[Geurts et al., 2006] Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Machine Learning 63(1):3-42. http://dx.doi.org/10.1007/s10994-006-6226-1.

[Horppila et al., 2022] Horppila, J., Pippingsköld, E., and Estlander, S. (2022). Effects of water colour on the pigment content of a floating-leaved macrophyte - implications of lake brownification. Aquatic Botany 181 103540. https://doi.org/10.1016/j.aquabot.2022.103540.

[Howard and Gugger, 2020] Howard, J. and Gugger, S. (2020). fastai: A layered api for deep learning. University of San Francisco. https://doi.org/10.48550/arXiv.2002.04688.

[Hutter et al., 2019] Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). Automated machine learning methods, systems, challenges. Springer. https://doi.org/10.1007/978-3-030-05318-5.

[Jerlov, 1951] Jerlov, N. G. (1951). Optical studies of ocean water. Reports of the Swedish Deep-Sea Expedition 1947–1948 Vol. 3 Physics and Chemistry No. 1. Göteborgs Kungl.

[Jerlov, 1976] Jerlov, N. G. (1976). Marine optics. Amsterdam New York: Elsevier Scientific Pub. Co.

[Ke et al., 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Microsoft Research, Peking University, Microsoft Redmond. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[Lehmann et al., 2023] Lehmann, M. K., Gurlin, D., Pahlevan, N., Alikas, K., Conroy, T., Anstee, J., Balasubramanian, S. V., Barbosa, C. C. F., Binding, C., Bracher, A., Bresciani, M., Burtner, A., Cao, Z., Dekker, A. G., Vittorio, C. D., Drayson, N., Errera, R. M., Fernandez, V., Ficek, D., Fichot, C. G., Gege, P., Giardino, C., Gitelson, A. A., Greb, S. R., Henderson, H., Higa, H., Rahaghi, A. I., Jamet, C., Jiang, D., Jordan, T., Kangro, K., Kravitz, J. A., Kristoffersen, A. S., Kudela, R., Li, L., Ligi, M., Loisel, H., Lohrenz, S., Ma, R., Maciel, D. A., Malthus, T. J., Matsushita, B., Matthews, M., Minaudo, C., Mishra, D. R., Mishra, S., Moore, T., Moses, W. J., Nguyn, H., Novo, E. M. L. M., Novoa, S., Odermatt, D., O'Donnell, D. M., Olmanson, L. G., Ondrusek, M., Oppelt, N., Ouillon, S., Filho, W. P., Plattner, S., Verdú, A. R., Salem, S. I., Schalles, J. F., Simis, S. G. H., Siswanto, E., Smith, B., Somlai-Schweiger, I., Sopp, M. A., Spyrakos, E., Tessin, E., van der Woerd, H. J., Woude, A. V., Vandermeulen, R. A., Vantrepotte, V., Wernand, M. R., Werther, M., and Yue, K. Y. . L. (2023). Gloria - a globally representative hyperspectral in situ dataset for optical sensing of water quality. Sci Data 10, 100. https://doi.org/10.1038/s41597-023-01973-y.

[Lloyd, 1982] Lloyd, S. P. (1982). Least squares quantization in pcm. IEEE Transactions on Information Theory, 28:129–137.

[McQueen, 1967] McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Berkeley, University of California Press, 1:281-297.

[Moore et al., 2009] Moore, T. S., Campbell, J. W., and Dowell, M. D. (2009). A class-based approach to characterizing and mapping the uncertainty of the modis ocean chlorophyll product. Volume 113, Issue 11, Pages 2424-2430. https://doi.org/10.1016/j.rse.2009.07.016.

[Morel and Prieur, 1977] Morel, A. and Prieur, L. (1977). Analysis of variations in ocean color. Limnology and Oceanography 4. https://doi.org/10.4319/lo.1977.22.4.0709.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. https://doi.org/10.48550/arXiv.1912.01703.

[Prokhorenkova et al., 2017] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2017). Catboost: unbiased boosting with categorical features. Yandex, Moscow, Russia. https://doi.org/10.48550/arXiv.1706.09516.

[Reinart et al., 2003] Reinart, A., Herlevi, A., Arst, H., and Sipelgas, L. (2003). Preliminary optical classification of lakes and coastal waters in estonia and south finland. Journal of Sea Research 49 (2003) 357–366. https://doi.org/10.1016/S1385-1101(03)00019-4.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Juornal of Computational and Applied Mathematics 20, 53-67. https://doi.org/10.1016/0377-0427(87)90125-7.

[Sculley, 2010] Sculley, D. (2010). Web-scale k-means clustering. Google, Inc. Pittsburgh. PA USA. https://citeseerx.ist.psu.edu/documentrepid=rep1type=pdf&doi=b452a856a3e3d4d37b1de837996aa6813bedfdcf.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 22, NO. 8. https://people.eecs.berkeley.edu/~malik/papers/SM-ncut.pdf.

[Sneath and Sokal, 1973] Sneath, P. and Sokal, R. R. (1973). Numerical taxonomy. W. H. Freeman and Company.

[Spyrakos et al., 2018] Spyrakos, E., O'Donnell, R., Hunter, P. D., Miller, C., Scott, M., Simis, S. G. H., Neil, C., Barbosa, C. C. F., Binding, C. E., Bradt, S., Bresciani, M., Dall'Olmo, G., Giardino, C., Gitelson, A. A., Kutser, T., Li, L., Matsushita, B., Martinez-Vicente, V., Matthews, M. W., Ogashawara, I., Ruiz-Verdu, A., Schalles, J. F., Tebbs, E., Zhang, Y., and Tyler, A. N. (2018). Optical types of inland and coastal waters. Limnology and Oceanography, 63(2), 846-870. https://doi.org/10.1002/lno.10674.

[Thornton et al., 2013] Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. Department of Computer Science, University of British Columbia. https://doi.org/10.48550/arXiv.1208.3719.

[Uudeberg et al., 2020] Uudeberg, K., Aavaste, A., Kõks, K.-L., Ansper, A., Uusnõue, M., Kangro, K., Ansko, I., Ligi, M., Toming, K., and Reinart, A. (2020). Optical water type guided approach to estimate optical water quality parameters. Remote Sensing 12, no. 6: 931. https://doi.org/10.3390/rs12060931.

[Uudeberg et al., 2019] Uudeberg, K., Ansko, I., Põru, G., Ansper, A., and Reinart, A. (2019). Using optical water types to monitor changes in optically complex inland and coastal waters. Remote Sensing 11, no. 19: 2297. https://doi.org/10.3390/rs11192297.

Appendix

I. Code repository

The computations done for the work are located in GitHub repository:
`https://github.com/fjodorsevtsenko/LTAT00019_MT`

II. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2021/04-10

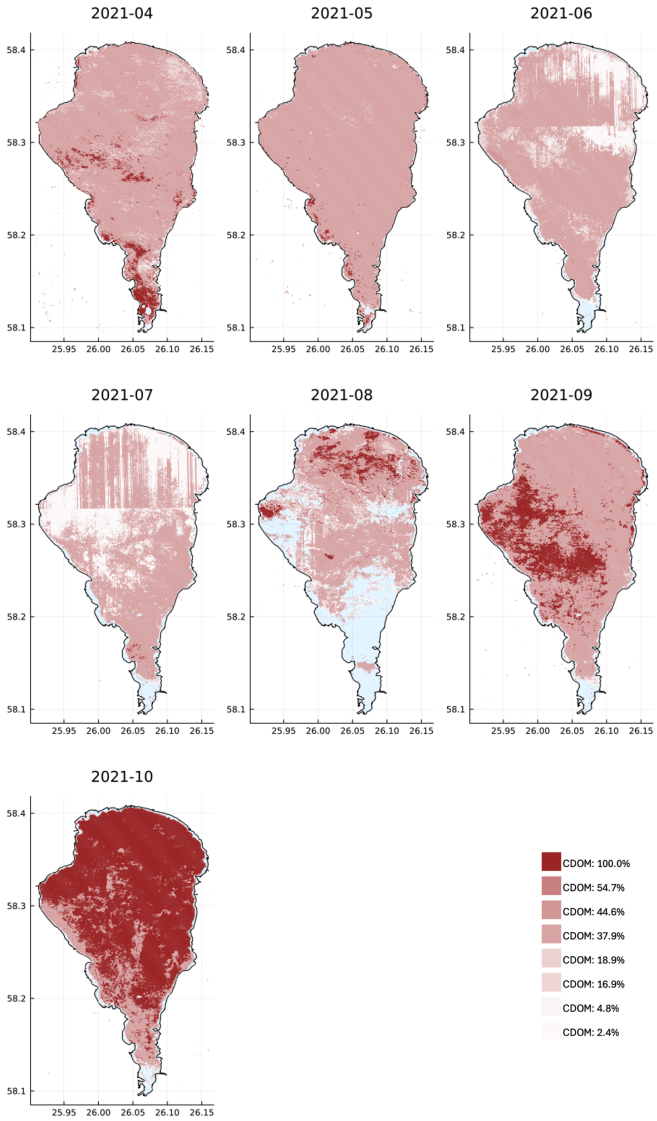


Figure 16. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2021/04-10

III. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2022/04-10

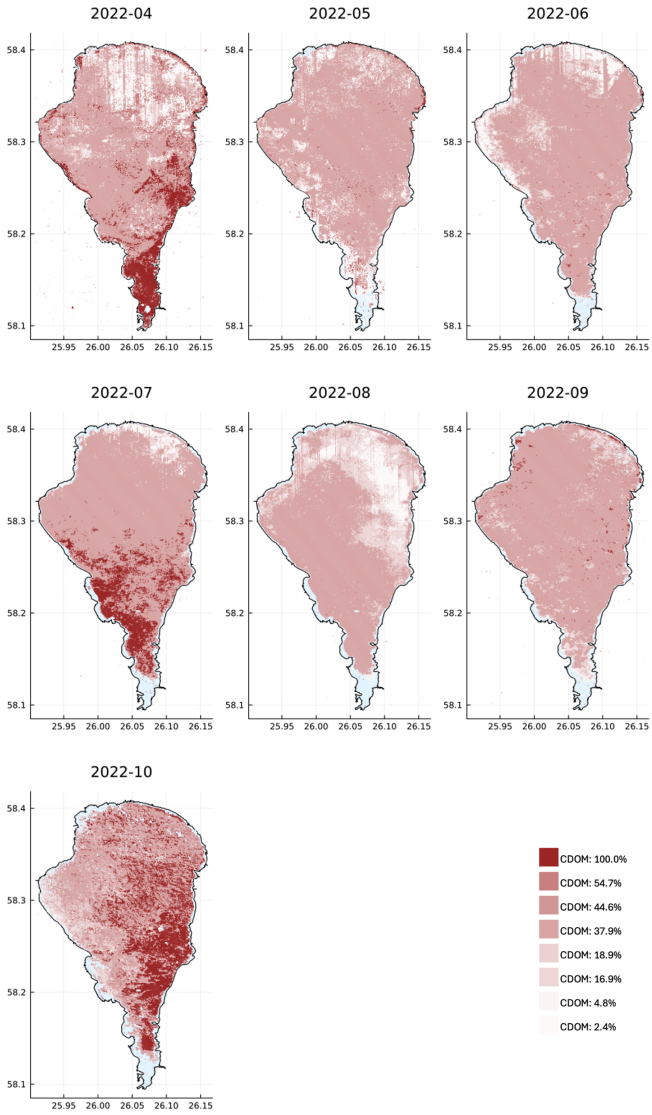


Figure 17. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2022/04-10

IV. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2023/04-10

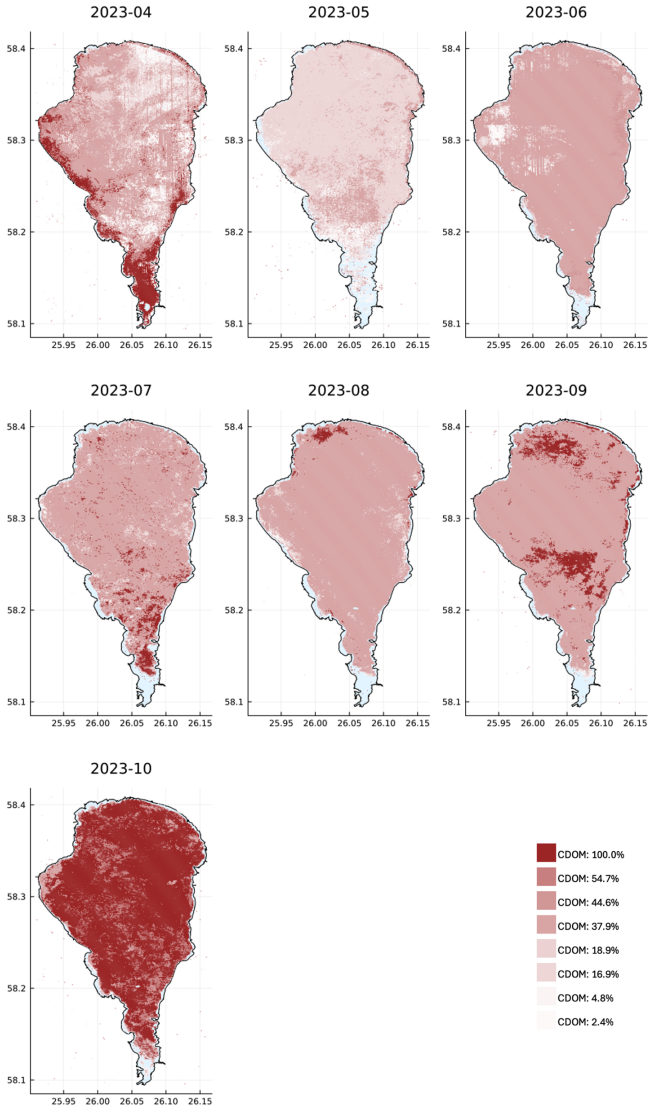


Figure 18. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2023/04-10

V. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2024/04-07

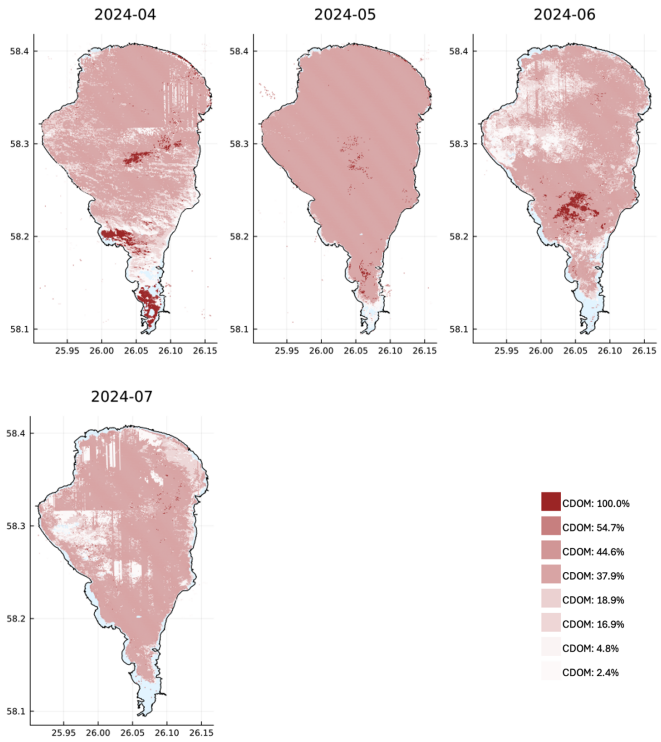


Figure 19. CDOM relative concentration levels dynamics of lake Võrtsjärv for the period of 2024/04-07

VI. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Fjodor Ševtšenko**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   CDOM-based Optical Water Types Classification,
   *(*title of thesis)

   supervised by Krista Alikas and Radwa El Shawi.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Fjodor Ševtšenko
08/08/2024