Larhyss Journal, ISSN 1112-3680, n°56, Dec 2023, pp. 179-192

USE OF MACHINE LEARNING ALGORITHMS AND IN SITU DATA FOR ESTIMATING PARTICULATE ORGANIC CARBON FROM THE MEDITERRANEAN SEA

FELLOUS S.[1*], BENDJAMA A.[2], BENZAOUI Y.[3]

[1] Laboratory of Hydrobiology and Pollution, Djillali Liabes University of Sidi Bel Abbes, 22000, Algeria
[2] Faculty of Life Sciences, University 20 août 55 of Skikda, 21000, Algeria
[3] Environmental Monitoring Network Laboratory, Sciences Faculty, Oran University, 31000, Algeria

(*) fellousuniv@gmail.com

Research Article – Available at http://larhyss.net/ojs/index.php/larhyss/index
Received June 6, 2023, Received in revised form November 28, 2023, Accepted December 4, 2022

ABSTRACT

Water quality indicators, including biological, chemical and physical properties, are usually determined by collecting data from the field and analyzing them in the laboratory. Although these in situ measurements are costly and time-consuming, they offer high accuracy. This study focuses on the estimation of particulate organic carbon (POC) as a water quality parameter using a combination of machine learning algorithms and hyperspectral in situ data. A data-driven approach that does not need any domain knowledge was used. We were interested in POC generated by bacteria, phytoplankton, zooplankton, detritus and sediments in the Mediterranean Sea from 15 May to 10 June 2017. Therefore, the objective of this study was to use five regression frameworks from machine learning algorithms to estimate POC with hyperspectral in situ data and evaluate their performance. Based on the coefficient of determination R^2, the best-performing modes were nearest neighbors (KNN), gradient boosting (GB) and random forest (RF), with an R^2 in the range of 72.33 to 74.7%. These machine learning models can be used to investigate more water quality parameters, as they reveal the great potential of this approach.

Keywords: POC, Machine learning, In situ measurement, Phytoplankton, Hyperspectral

INTRODUCTION

Particulate organic carbon (POC) is composed of living biomass, e.g., bacteria, zooplankton, phytoplankton, and nonliving organic matter, i.e., sediments, facal pellets and detritus; it is of substantial interest because it represents one carbon reservoir. Concentration changes in POC in surface waters result from its transformation (e.g., excretion of organic carbon, remineralization), biological production and export inside the ocean (Stramska, 2005). In fact, POC sinks from surface waters and works as a biological pump that provides a mechanism for storing carbon in the deep ocean (Stramski1, 2008). To understand the effect of climate change on the impacts of biological carbon pumps, it is important to know the spatial and temporal distribution of POC in the global ocean (Bopp et al. 2001; Xie, 2019).

There are few studies on the utility of total particulate absorption (ap) as a POC proxy (Stramski et al., 2008; Allison et al., 2010a), as most in situ studies have focused on the relationship between the backscattering coefficient of particles (bbp) or the beam attenuation coefficient of particles (cp) and POC (Stramski et al., 2008; Allison et al., 2010a). It is known that particulate absorption coefficient variability depends on the variability in particulate inorganic and organic components (biogenic detritus, phytoplankton, bacteria, biomass, atmospheric dust) contained in POC (Bricaud et al., 1998; Stramski et al., 2004b).

Conventional monitoring techniques measure point-based water quality parameters either in situ or later in a laboratory. These measurements are precise at a specific location and provide a detailed depth profile compared to remote sensing. Machine learning data-driven approaches are generally able to address complex problems, and without any field knowledge, they only require access to a sufficient amount of input data (Ifarraguerri, 2000), such as high-dimensional hyperspectral data from both in situ measurements and remote sensing.

This study was inspired by the work of Keller (2018) on hyperspectral data and machine measurements for estimating chromophoric dissolved organic matter (CDOM), chlorophyll a, diatoms, green algae and turbidity. We used a hyperspectral in situ dataset from the SeaWiFS Bio-optical Archive and Storage System (SeaBASS) (Werdell et al. 2003), measured under real-world conditions in the Mediterranean Sea for the period between May and June 2017. For this purpose, five machine learning regression algorithms were applied to capture the best model for monitoring POC. The hyperspectral bands have been used as a function of POC target data for supervised machine learning models. This framework has been applied and evaluated in the context of both linear and nonlinear regression problems. The aim of this study was to use hyperspectral in situ data with machine learning algorithms and test the power of these models for the estimation of particulate organic carbon (POC) in the Mediterranean Sea. The variable inflation factor (VIF) and PCA reducing dimensionality will be used for feature selection before performing the model prediction.

DATA AND METHODS

Dataset and instrumentation

Since 1997, through various projects and programs, NASA has funded the collection of ocean in situ measurements for satellite data product validation, algorithm development, satellite data comparison and intercalibration, data merger studies and time series analyses. The SeaWiFS Bio-optical Archive and Storage System (SeaBASS) (Werdell et al., 2003; Werdell and Bailey, 2005) maintains a local repository of in situ ocean bio-optical and optical data to support regular scientific analyses. Specifically, the database includes in situ ocean biological (e.g., chlorophyll-a/b, CDOM, POC concentration), optical (e.g., ocean water-leaving radiance spectra), and other related oceanographic and atmospheric data (see details in http://seabass.gsfc.nasa.gov/).

In fact, a variety of researchers from the international ocean community have contributed to SeaBASS data using instrumentation with rigorous measurements, data processing protocols and community-defined deployment (Mueller and Fargion, 2002). SeaBASS in situ data are continuously used in support of SeaWiFS and MODIS ocean color remote sensing products for validation and algorithm evaluation (Bailey and Werdell, 2006). Specifically, the SeaBASS data are appropriate for the new algorithm evaluations. The data in our study were collected from the Mediterranean Sea using a Wetlabs AC-S hyperspectral instrument (underwater spectral absorption and attenuation meter) for measuring the inherent optical properties (IOPs) of water. It offers an almost order of magnitude increase in the spectral resolution of the in situ beam attenuation and absorption coefficients. The instrument has a compact size and excellent stability and features a proven flow-through system. With outputs of 80+ wavelengths from 400-730 nm, the 4 nm resolution enables deconvolution analysis and spectral 'fingerprinting'.

Feature selection and preprocessing

The regression is performed with the hyperspectral data as input vectors and the POC data as the target value. The complete dataset collected from the SeaBASS website (https://seabass.gsfc.nasa.gov/search#val) consists of 15499 data points.

One data point is defined by 66 hyperspectral bands, and one (POC) is defined as the target value. For the best interpretation and a good analysis, we applied feature band selection, resulting in a wavelength range between 400.1 m^{-1} and 700.5 m^{-1}.

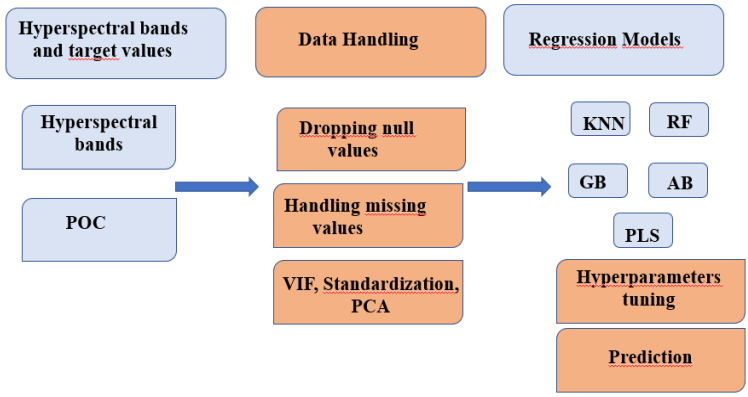

Figure 1: Schematic representation of the regression framework

Three preprocessing methods were applied to enhance the regression process with high-dimensional data.

The first method was the variance inflation factor (VIF), which is a measure of the amount of multicollinearity in a set of multiple regression variables. VIF is equal to the ratio of the overall model variance to the variance of a model. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

As a second preprocessing method, we applied standardization. This method scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one. The standard scaling formula is as follows: $Z = (x-u)/s$, where u is the mean of the training samples and **S** is the standard deviation of the training samples.

Principal component analysis (PCA) represents the third method. PCA is a commonly used multivariate statistical method (Wang, 2019) that generates a set of principal components that are linear transformations of the original variables. These new principal components are orthogonal to each other and sorted according to the explained variance (WENG, 2020).

In the regression frameworks, the complete dataset is split randomly into a training subset and a test subset. The training subset includes 80% full data points, and the test subset consists of 20% data points. A schematic representation of this study framework is presented in figure 1.

Regression models

To estimate POC, we selected appropriate regression models and included them in the framework. These are random forest (RF), gradient boosting (GB), k-nearest neighbors (KNN), Ada Boost regression (AB), and PLS Regression (PLS). The regression models

are trained on the training subset by linking the hyperspectral data to the POC target values during the training phase. All the machine learning models are implemented in the Python package (**3.9.7)** Scikit-learn. Regressions are performed by supervised and unsupervised learners.

RESULTS AND DISCUSSION

In this section, we present the performance of the regression framework to estimate the POC water quality parameters.

An example of the correlation between bands of our original data is shown in the heatmap (Fig. 2). This figure shows that the bands are heavily intercorrelated, which may generate a multicollinearity problem in the regression process.

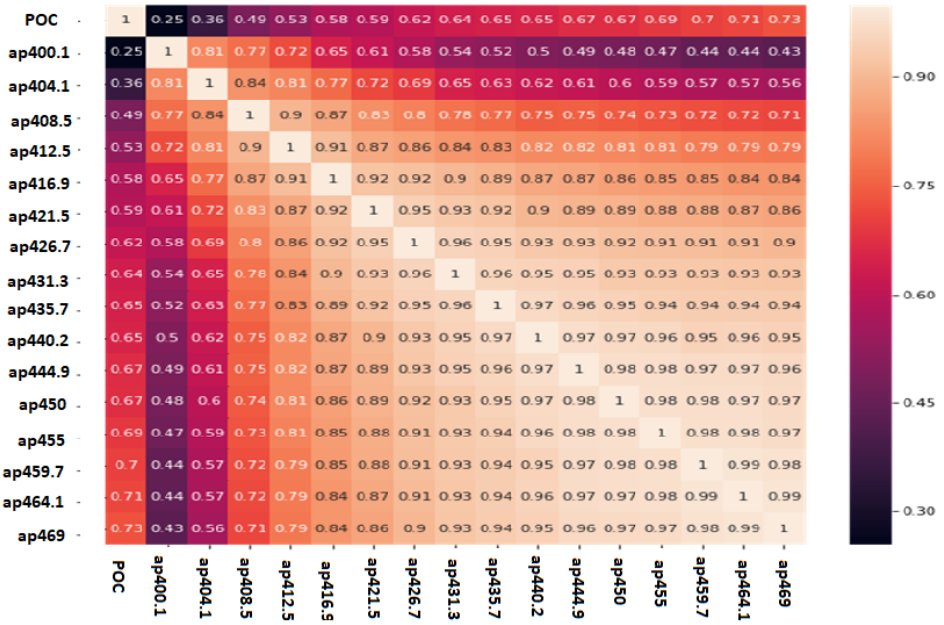

Figure 2: Heatmap of hyperspectral in situ data with POC. Each cell is colorized based on the coefficient of determination.

To explore trends in our data, we used a scatter matrix (pairs plot), which compactly plots the numeric variables in a dataset against each other; the results are presented in Fig. 3. Only the POC target and features with the greatest variability are shown.

To determine the adequacy of the fitted model, we used the coefficient of determination R^2, which is a popular tool. In statistical analysis, it represents the measure that assesses how well a model explains and predicts future outcomes.

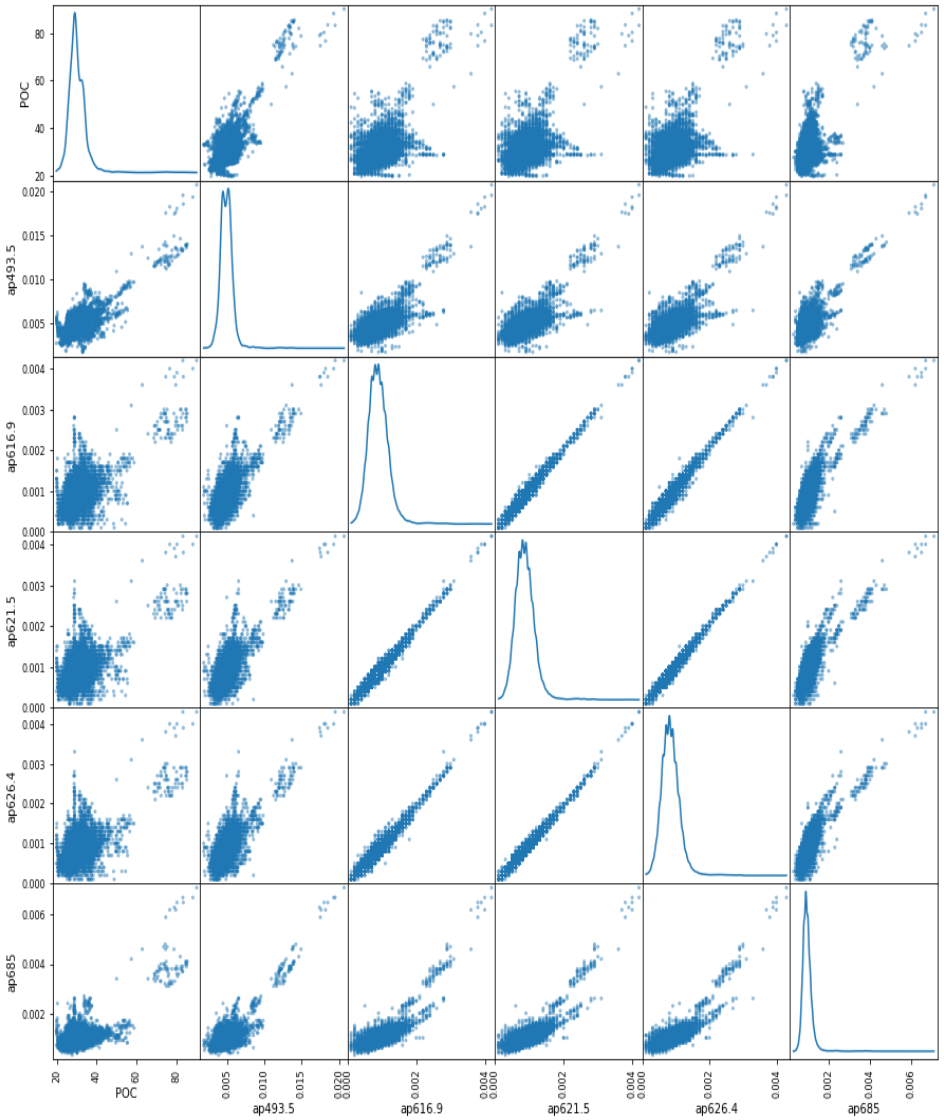

Figure 3: Scatter matrix of hyperspectral in situ data and POC showing the correlation of variables.

After the application of VIF, we ended up with 20 bands chosen regarding their lowest VIF values, and the results are presented in Table 1.

The variance inflation factor quantifies the increase in the predicted regression coefficient when predictors are linked (Wang et al, 2022). A VIF value greater than 10 can be used as a strong indicator of multicollinearity (Akinwande et al, 2015; Wooldridge, 2015).

The subset of predictors from the feature selection based on the VIF value consisted of twenty predictor variables. This subset of predictor variables can be seen in Table 1 and was further processed by using standardization and PCA before building the machine learning models.

Table 1: Variance inflation factor of twenty bands sorted in ascending order

N°	Features	VIF_values	N°	Features	VIF_values
1	ap700.5	14.74341	11	ap644.5	264.647
2	ap400.1	19.4459	12	ap635.2	296.8808
3	ap404.1	37.79633	13	ap649.1	300.3684
4	ap696.6	50.88418	14	ap688.7	304.9775
5	ap408.5	75.8985	15	ap626.4	317.6361
6	ap412.5	123.8203	16	ap630.7	318.0892
7	ap692.4	146.7568	17	ap426.7	352.5488
8	ap416.9	187.2344	18	ap621.5	359.2506
9	ap421.5	225.4132	19	ap653.9	366.1557
10	ap639.8	262.0507	20	ap616.9	386.3535

From Table 1, we note that all variables have a VIF value greater than 10. To avoid the problem of multicollinearity, which can be generated by such heavily correlated explanatory variables, we applied ACP to reduce the dimensionality of our data. The number of appropriate component numbers can be seen in Fig. 4.

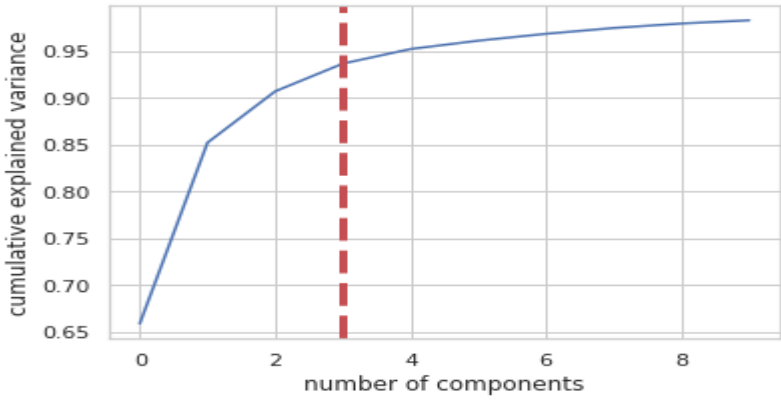

Figure 4: Explained variance and principal components

In this study, we used a scree plot to select the first three principal components instead of the full 20 features of the hyperspectral input data for the regression. These three components were chosen since they covered approximately 94% of the variance (Fig. 4). The cumulative and explained variance ratios are shown in Table 2.

Table 2: Cumulative and explained variance ratio

Components	Cumulative Variance Ratio	Explained Variance Ratio
PCA1	0.66	0.66
PCA2	0.85	0.19
PCA3	0.91	0.06

Since hyperparameters have a great effect on the performance of prediction models, the process of finding the optimal hyperparameters is inevitable in terms of the accuracy or time of the model. In many cases, the trial-and-error method is used by engineers to manually tune hyperparameters. Even experienced researchers will try to find the optimal combination of hyperparameters, and finding the best hyperparameters takes a long time and is difficult. To solve this problem, we used a grid search (GS) method from the scikit-learn library. This method defines the search space as a regular grid and evaluates every position in the grid. With a small step size and a large search range, GS has a high probability of discovering the global optimum. However, this method can be very computationally intensive and time consuming, especially when the number of hyperparameters to be tuned is relatively large. The optimum hyperparameters obtained with GS for the five models are shown in Table 3.

Table 3: Hyperparameter tuning of the regression models

Algorithm	Hyperparameters	Hyperparameters values	Optimum values
GB	max_depth	[2,3,7,11,15]	3
	min_samples_leaf	[2,3,4,5,6,7]	2
	min_samples_split	[2,3,4,22,23,24]	22
	n_estimators	[300,500,700]	300
AB	learning_rate	[0.01,0.05,0.1,0.3,1]	0.05
	loss	['linear', 'square', 'exponential']	Exponential
	n_estimators	[50, 100]	100
KNN	n_neighbors	[2,3,4,5,6]	6
	weights	['uniform','distance']	uniform
RF	max_depth	[5, 10, 15, 20, 25,30]	10
	max_features	['auto', 'sqrt']	auto
	min_samples_leaf	[1, 2, 5, 10]	1
	min_samples_split	[2, 5, 10, 15, 100]	10
	n_estimators	[100, 200,300,400,500]	200
PLS	max_iter	[100, 200,300,400,500]	100
	n_components	[2, 3, 4,6,8,10,12,14]	3
	tol	[1e-02,1e-04,1e-06,1e-08,1e-10]	0.01

Using five regression models after processing and dimensionality reduction, the results of regression with the coefficient of determination R^2 for GB, AB, KNN, RF, and PLS were 72.33% 61.5%, 73.84, 74.70%, and 50.12%, respectively.

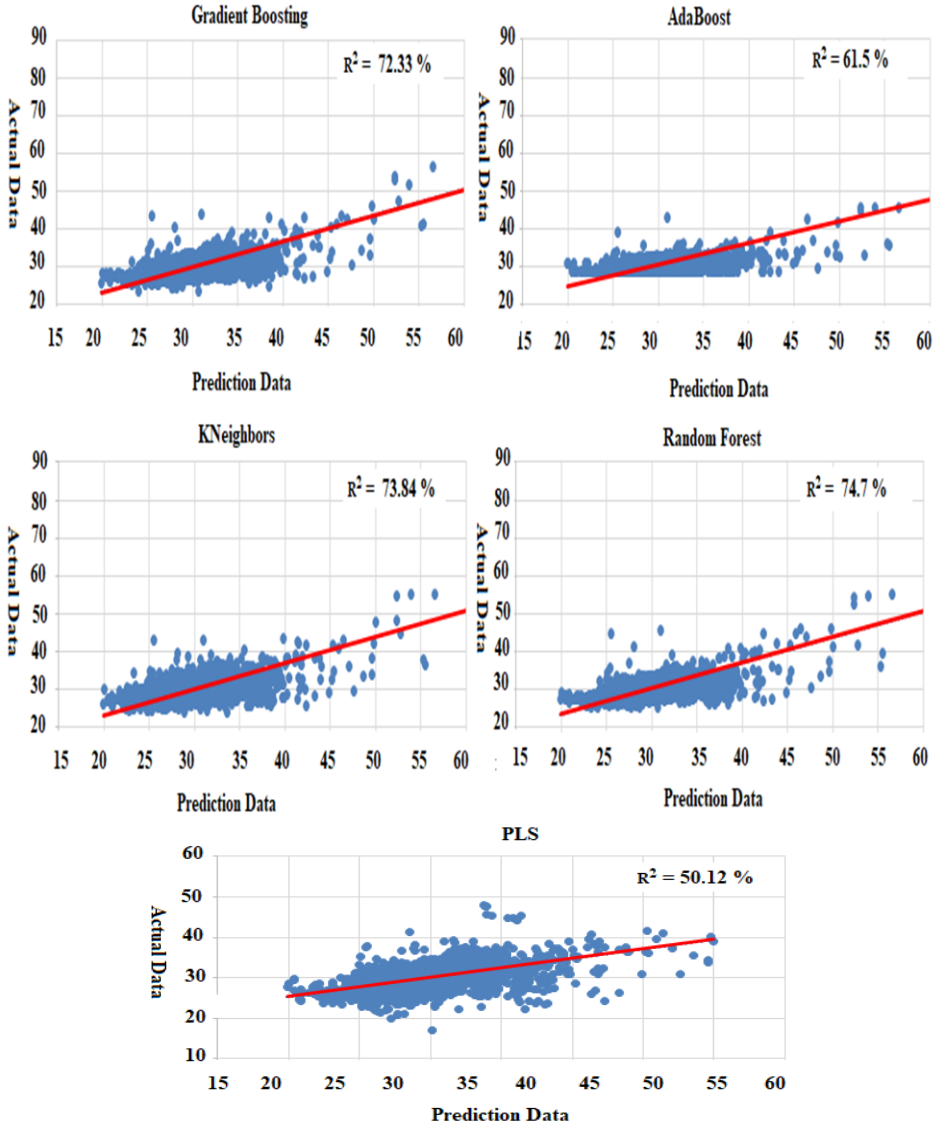

Figure 5: Prediction plots of the five regressors generated using Python software 3.9.7.

Among all regressors, PLS and AB had the worst results, which were in the range of 50 to 62%. For the remaining tree models, the regression values were very close, with the random forest regressor having a slightly higher value (74.7%), as mentioned above.

Machine learning algorithms can process and analyze high-dimensional information features and statistically analyze the importance of multidimensional features, which is a decisive step in understanding the hyperspectral characteristics of IOP parameters related to the POC. The benefit of machine learning is to perform regressions without possessing further knowledge of the investigated water quality and water body parameters. Furthermore, there is no need to engineer new features based on domain knowledge because this approach is purely data-driven.

The main objective of this paper is to investigate the potential of estimating POC as a water quality parameter with five regression models and in situ measured input data. It is important to note that there have been a few studies on the application of machine learning and hyperspectral data in estimating water quality parameters.

By applying standard scaling and PCA-based dimensionality reduction, we obtained regression results relying on the first three principal components with five regression models. The goodness of fit of the regression model estimation is indicated by the good R2 values; this performance may be in part due to the random split between training and test subsets (Keller, 2018). In this study, we obtained R^2 values of 72.33%, 61.5%, 73.84%, 74.70%, and 50.12% for GB, AB, KNN, RF, and PLS, respectively.

For comparison, Keller (2018), working on CDOM, obtained R^2 values of 91.2%, 91.9%, 85.3%, 91.4%, and 83.2% for GB, AB, KNN, RF, and PLS, respectively, with only PCA processing and 80.0%, 79.9%, 83.0%, 82.4%, 84.9% for GB, AB, KNN, RF, and PLS, respectively, using only min_max scaling.

On the other hand, the turbidity results obtained in the same study were 85.5%, 85.2%, 72.8%, 84.1%, and 70.9% with only PCA as the processing method and 69.3%, 66.4%, 70.7%, 67.7%, 73.3% for GB, AB, KNN, RF, and PLS, respectively, using only min_max as the scaling method. These results are better than those obtained from our study, except for the estimation of turbidity with scaling, in which our results are much better.

As reported in the study of Trung (2019), high POC values will increase absorption in the blue–green part of the spectrum at 490 nm. In this study, we had to address the problem of multicollinearity by reducing an important number of variables, especially those in the green spectrum, which are very informative in POC estimation; then, without these bands, the regression models may be less performant. In fact, the multicollinearity between the covariates is the most important obstacle that we must overcome (Ntotsis and Karagrigoriou, 2021). Due to the interdependence of the variables and standard errors, estimations are unstable if the regression coefficient is important, which decreases their precision and makes them unreliable (Alin, 2010). As two or more variables have linear relationships, it makes variable marginal impact hard to measure. The model will have poor generalization and overfit the data. Then, it performs poorly on new data (Chan et al, 2022).

Additionally, the variability in the particulate absorption coefficient is driven by the variability in particulate organic and inorganic components (biogenic detritus, atmospheric dust, biomass, bacterial phytoplankton, etc.) contained in POC (Bricaud et al., 1998; Stramski et al., 2004b; Rasse et al., 2017). As the concentration of POC is a function of absorption, the performance of our result may be influenced by the ratio of organic to inorganic components.

In summary, the regression results indicate the adequate applicability of machine learning as a regression framework when estimating the POC water quality parameter based on spectral data. These performances are confirmed by the good results obtained from three regression models among the five used in this study. As seen before, the GB, KNN, RF models had a regression coefficient greater than 70%, the worst was PLS with an R^2 of approximately 50%, and an average result was obtained by AB with an R^2 equal to 61.5%. In fact, an R^2 higher than 0.6 is considered a worthwhile prediction model (Chaplot et al. 2021).

CONCLUSIONS

In this study, we evaluated the potential of regression models to estimate POC water quality indicators with hyperspectral data. Five regression frameworks were applied to the measured in situ data, with scaling and PCA reducing dimensionality. After hyperparameter tuning of the five models, the tree (GB, KNN, RF) had a good regression coefficient, which was in the range of 72.33 to 74.7%. This result indicates that the machine learning algorithms, especially the ensemble model (RF), which had the best result in our study (74.7%), can be used to predict water quality parameters with good performance. In the future, we will explore more sophisticated models, such as neural networks, to gain accuracy for the prediction of water quality indicators.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

The authors thank all the researchers for their contribution to the SeaBASS datasets. The work would not be possible without the efforts of many scientists and organizations who made these datasets publicly available. We thank the anonymous reviewers for their constructive comments and suggestions that improved the manuscript significantly.

REFERENCES

AKINWANDE M.O., DIKKO H.G., SAMSON A. (2015). Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. Open Journal of Statistics, Vol. 5, pp. 754–767.

ALIN A. MULTICOLLINEARITY. (2010). Wiley Interdiscip. Rev. Comput. Stat, Vol. 2, pp. 370–374.

ALLISON D.B., STRAMSKI D., MITCHELL B.G. (2010). Empirical ocean color algorithms for estimating particulate organic carbon in the Southern Ocean, Journal of Geophysical Research Oceans, Vol. 115, Issue C10, pp. 1-16. https://doi.org/10.1029/2009JC006040

BAILEY S.W., WERDELL P.J. (2006). A multisensor approach for the on-orbit validation of ocean color satellite data products, Remote Sensing of Environment, Vol. 102, Issues 1-2, pp. 12-23. https://doi.org/10.1016/j.rse.2006.01.015.

BOPP L., MONFRAY P., AUMONT O., DUFRESNE J.L., TREUT H.L., MADEC G., TERRAY L., ORR J.C. (2021). Potential impact of climate change on marine export production, Global Biogeochemical Cycles, Vol. 15, No 1, pp. 81-99. https://doi.org/10.1029/1999GB001256

BRICAUD A., CLAUSTRE H., RAS J., OUBELKHEIR K., (2004). Natural variability of phytoplanktonic absorption in oceanic waters: Influence of the size structure of algal populations, Journal of Geophysical Research Oceans, Vol. 109, Issue C11, pp. 1–12. https://doi.org/10.1029/2004JC002419

BRICAUD A., MOREL A., BABIN M., ALLALI K., CLAUSTRE H., (1998). Variations of light absorption by suspended particles with chlorophyll a concentration in oceanic (case 1) waters: analysis and implications for bio-optical models, Journal of Geophysical Research. Oceans, Vol. 103, Issue C13, pp. 31033-31044. https://doi.org/10.1029/98JC02712.

CHAN J.Y.L., LEOW S.M.H., BEA K.T., CHENG W.K., PHOONG S.W., HONG Z.W., CHEN Y.L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review Mathematics, Vol. 10, pp. 1283. https://doi.org/10.3390/math10081283

CHAPLOT B., PETERS M., BIRBAL P., PU J.H., SHAFIE A. (2021). Use of gene-expression programming to estimate manning's roughness coefficient for a low flow stream, Larhyss Journal, No 48, pp. 135-150

IFARRAGUERRI A., CHANG C.I. (2000). Unsupervised hyperspectral image analysis with projection pursuit, In IEEE Transactions on Geoscience and Remote Sensing, Vol. 38, No 6, pp. 2529–2538. DOI: 10.1109/36.885200.

KELLER S., MAIER P.M., RIESE F.M., NORRA S., HOLBACH A., BOERSIG N., WILHELMS A., MOLDAENKE C., ZAAKE A., HINZ S. (2018). Hyperspectral data and machine learning for estimating CDOM, Chlorophyll a, diatoms, green algae and turbidity, International Journal of Environmental Research and Public Health, Vol. 15, Issue 9, 1881.

KIMON N., KARAGRIGORIOU A., ARTEMIOU A. (2021). Interdependency Pattern Recognition in Econometrics: A Penalized Regularization Antidote. Econometrics Vol. 9, pp. 44. https://doi.org/10.3390/econometrics9040044 https://doi.org/10.3390/ijerph15091881

MUELLER J.L., FARGION G.S. (2002). Ocean Optics Protocols for SeaWiFS Validation, Revision 3, NASA Technical Memorandum, 2002–210004, NASA Goddard Space flight center, Greenbelt, Maryland, 308 p.

RASSE R., DALL'OLMO G., GRAFF J., WESTBERRY T. K., VAN DONGEN-VOGELS V., BEHRENFELD M.J. (2017). Evaluating optical proxies of particulate organic carbon across the surface Atlantic Ocean, Frontiers in Marine Science, Vol. 4, pp. 1-18. https://doi.org/10.3389/fmars.2017.00367.

STRAMSKA M., STRAMSKI D. (2005). Variability of particulate organic carbon concentration in the north polar Atlantic on SeaWiFS ocean color observation, Journal of Geophysical Research Oceans, Vol. 110, Issue C10, C10018. https://doi.org/10.1029/2004JC002762.

STRAMSKI D., REYNOLDS R.A., BABIN M., KACZMAREK S., LEWIS M.R., RÖTTGERS R., SCIANDRA A., STRAMSKA M., TWARDOWSKI M.S., FRANZ B.A., CLAUSTRE H. (2008). Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans, Biogeosciences, Vol. 5, pp. 171-201. https://doi.org/10.5194/bg-5-171-2008.

WANG W., JIN S., GUO Y.L. (2019). Exploration of a Method of Distinguishing Different Nongxiang Tieguanyin Tea Grades Based on Aroma Determined by GC–MS Combined with Chemometrics. Molecules, Vol. 24, 1707. https://doi.org/10.3390/molecules24091707

WANG K.J. (2022). How Is the Lung Cancer Incidence Rate Associated with Environmental Risks? Machine-Learning-Based Modeling and Benchmarking, International Journal of Environmental Research and Public Health, Vol. 19, 8445. https://doi.org/10.3390/ijerph19148445.

WENG S., GUO B., TANG P., YIN X., PAN F., ZHAO J., HUANG L., ZHANG D. (2020). Rapid detection of adulteration of minced beef using Vis/NIR reflectance spectroscopy with multivariate methods, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, Vol. 230, 118005.

WERDELL P.J., BAILEY S.W., FARGION G., PIETRAS C., KNOBELSPIESSE K., FELDMAN G., MCCLAIN C. (2003). Unique data repository facilitates ocean color satellite validation, EOS Science News, Vol. 84, Issue 38, pp. 377-387. https://doi.org/10.1029/2003EO380001.

WERDELL P.J., BAILEY S.W. (2005). An improved bio-optical dataset for ocean color algorithm development and satellite data product validation, Remote Sensing of Environment, Vol. 98, Issue 1, pp. 122-140. https://doi.org/10.1016/j.rse.2005.07.001.

WOOLDRIDGE J.M. (2015). Introductory Econometrics, A Modern Approach; Cengage Learning: Boston, MA, USA.

XIE F., TAO Z., ZHOU X., LV T., WANG J. (2019). Spatial and temporal variations of particulate organic carbon sinking flux in global ocean from 2003 to 2018, Remote Sensing, Vol. 11, Issue 24, 2941. https://doi.org/10.3390/rs11242941.