Bluedata - Testing and Development of HAB Algorithms in the North East Atlantic

REPORT 1

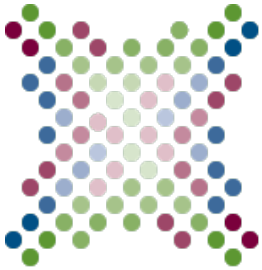
ACTIVITIES REPORT:

REF.: M3.1.A (BGCT)/F/001/BGCT/2018

ISSAH NAZIF SULEIMA

Terceira, Faial, Açores

2020/2021



FRCT

FUNDO REGIONAL PARA A CIÊNCIA E TECNOLOGIA

UAc
UNIVERSIDADE
DOS AÇORES

.

I. Introduction

The continuous infestation of algae in coastal areas around the world has become a concern to both stakeholders and researchers. This is particularly true when Harmful Algal Blooms, or HABs, occur, that is, when colonies of algae grow out of control and produce toxic or harmful effects on people, fish, shellfish, marine mammals, and birds. The human illnesses caused by HABs, though rare, can also be debilitating or even fatal. As a result, scientists continuously work towards building a much more improved monitoring techniques to measure the growth and the distribution of algae as a way of detection to give an early alert or warning to stakeholders to carry out mitigative measurement. Harmful algal blooms are known to cause not only a threat to the diversity of ecosystem but also to cause severe negative economic impacts around the world (Bauman 2010) which is a more reason why the monitoring and prediction of HABs is of great importance.

Monitoring and detection of HABs over the years has seen many interests among researchers with the very aim of improving the already available techniques in HABs detections as well as, ocean event detection. Remote sensing techniques among other approaches have become readily available for the study of large-scale biological processes and endangered environments including coastal areas. Research reveals that RS approach to algal detection provides a wide coverage zone with high temporal resolution which in turn are efficient for long term monitoring.

II. Background

Basically, the early days of HAB detection using remote sensing used spatially isolated and single satellite data samples. Nevertheless, there are various other approaches to the detection of HABs which takes into consideration many satellite sensors and bands.

Chl-a concentration in many cases is considered as a proxy variable for phytoplankton concentration in ocean and lake water because of the relationship between the concentration of phytoplankton and primary productivity (W. Song, 2015). Many researchers summarize the use of Chl-a as primary indicator of photoautotrophic biomass based on the relationship it has with primary productivity (Steele JH, 1962). Literatures have shown that, Chl-a concentration in water body is accepted as an important index for the detection and monitoring of pollution (i.e., HABs) in inland waters.

The conventional method of HAB detection from literature are mainly the optical and the physical approach. These techniques are used based on a binary tree question of Yes or No (Does the bloom change water colour or not). An optical approach is then deployed if the answer is yes while the

physical forcing approach is used if the answer is a No. The optical techniques (Empirical, Analytical and Semi-Analytical algorithms) can distinguish some blooms types in case 1 waters (open ocean, dominated by the optics of phytoplankton and its detrital products) although it appears to be difficult in coastal areas due to sediments and CDOM (i.e. case 2 waters). The algorithms for the HABs detections using this technique uses absorption, backscattering, relative patterns and spectral shape. The physical forcing technique defines a relationship between blooms and its physical features. It uses SST as a primary attribute such as fronts, upwelling and wind events. It should be said however that it is difficult to view the extent of blooms using the physical forcing technique as no direct bloom detection is generally the case.

Chl-a mostly can be estimated using an optical technique of water leaving reflectance which establishes a relation such as "band ratio". The approach relies on the basic principle of selecting two spectral bands that are representative of absorption/scattering features of Chl-a (Gin et l, 2002). [1] This algorithm has the strength to allow the compensation for variation from atmospheric influences (Jensen, 2005), [2] the scattering and absorption features of Chl-a with more than single band.

There are few literatures on using Machine learning algorithms to monitor or map HABs. Some of the very few use GIS based Machine learning techniques. Weilong Song used ocean color imagery from SeaWiFS to provide an early warning system for HABs in the eastern Gulf of Mexico. Gokaraju et al used spatiotemporal analysis to predict HABs. Machine learning algorithms like the Support Vector Machine Model has also been used to predict HABs (Song et al7). Support Vector Machine model has been used to predict ocean events but could not validate the model (Bernstein M, 2013).

III. Study site

The basic starting point of HABs detection is spotting out a location for the monitoring and detecting of HABs and the extent, which mostly are associated with either an increase or decrease in the brightness of water bodies that occurs in offshore waters. After having observed the existence of an unusual colour changed or bloom, the researcher will then proceed to extract the biomass on the water surface which in most cases is defined by the phytoplankton biomass (described by the concentraion of Chl-a and other accessory pigments).

Based on a well-defined algorithm (which in our case GIS-based) to be used in the detection (Empirical, Analytical, Statistical, etc.) and the availability of data, the researcher decides on the which satellite to use for the data collection. Satellite sensors are known to be the sole providers of spatial coverage for a large area monitoring however, airborne remote sensing is used despite its higher costs (Kutser 2009) in small to medium-sized lakes. The most commonly used satellites are: Medium Resolution Imaging Spectrometer

(MERIS); Moderate Resolution Imaging Spectroradiometer (MODIS); Sea-viewing Wide Field-of-View Sensor (SeaWİFS); Advanced Very Hight Resolution Radiometer (AVHRR); Advanced Along-Track Scanning Radiometer (AATSR), etc. Each of them have different specifications and features (resolution, availability of data, spectral bands among many).

Much research has been carried out on the detection, mapping or monitoring of HABs around the world using various methods. Most of these algorithms make its classification by discriminating the predominant components (which in most case is Chl-a concentration) in the research site. Some methods used for the mapping, detection and forecasting of HAB events have included, reflectance band-ratio, reflectance classification using chl-a anomaly, satellite product based using threshold and spectral band differences. The most successful methods have sort to use chlorophyll concentration as it is a well-known fact that phytoplankton increases the backscattered light within pigment absorption spectral frequency (Blondeau-Pastissier et al, 2014).
Most research estimate the Chl-a concentration using statistical approaches like the "band-ratio algorithm" and analytical methods (Yaohuan, 2010; Luoheng, 2015).

IV. Data collection

Two data types are usually used in the monitoring of HABs detection, satellite data and in-situ data. Satellite datasets are generally collected in the form of imagery data which are usually available on any of the satellites mentioned earlier. Depending on the sensor or satellite type and data, the datasets may require some preprocessing steps to be carried out on them before any further statistical analyses can be made. The most common data preprocessing steps for multispectral imagery are atmospheric correction and orthocorrection. While the need for orthocorrection depends on the processing level of the datasets (this may include sensor correction and terrain correction), atmospheric correction requires the use of an external digital elevation model (DEM). Terrain corrections helps to correct problems associated with datasets or imagery acquired with off-nadir (oblique) view angle. Atmospheric corrections correct the atmospheric effect (to derive Top-Of- Canopy TOC and Bottom-Of- Atmosphere BOA). The images from these satellites are collected under a considered weather condition (the sky and wind). The satellite data from these satellite sources are generally of different pigments (Chl-a, fluorescence line height, Turbidity's, POC, IPAR, PAR among many). These data are collected within a specified window of time mostly

due to the availability of data. Some feature selections are carried on these satellites to deal with different resolutions to attain a much more reliable model inputs.

In-situ data are field, or study site collected measurements which are collected purposely for the verification of actual field data collected. This could be a field sample collection or sample data collected by an instrument in direct contact with the medium.

V. Summary of the GIS Algorithm application

The basic approach used here is the classification approach which most researchers use the GIS approach or Statistical approach (Pixel-based vs object based classification and the supervised classification of satellite imagery) Empirical, Analytical or even the semi-analytic approach.

Having a well-defined algorithm, the researcher will then use different chl-a estimation indices such as; Slope Algorithm index (Sred-NIRAI),normal difference Chlorophyll index (NDCI), normal difference vegetation index, floating algal index, normal difference water index among others which generate different thresholds for the estimation of Chl-a. Each of these indices are known to have to performed better in classifying specific pigments or more. Oyama et al (year?) using NDVI (best index to classify water), NDWI (best to distinguish cyanobacteria and marcrophytes) and FAI, calculated for each index a different threshold using a spectral decomposition algorithm. The algorithms are based on a unique combination of spectral bands (in blue, green, red and NIR) to estimate Chl-a. Some algorithms has restrictive applicability as they are built on some assumptions (Oyama et al, 2010).

The ability of the selected indexes (NDVI, FAI, NDCL, EVI and more) to estimate Chl-a by a constant tunning of the bands through the satellite interface to give an ideal threshold is not straightforward as it is affected by temperature, wind, turbidity among other variable factors. The algorithms classify the images as "HAB" or "not HAB" based on the threshold calculated from the index.

After having built the model, this is then tested using the validation sample data or in-situ data to assess its performance. The accuracy of the model along with other parameters like the sensitivity, specificity and the RME are also determined to assess the efficiency of the model.

.

VI. Machine Learning Algorithm

HAB detection or monitoring is a general classification method in which every algorithm whether statistical or not seeks to define the data as positive HAB or negative HAB labelled as "1" and "0", respectively. This type of classification is generally a supervised learning scheme although there could be a complex application of non-supervised approach to cluster the data which is a combination of several neural networks. The most used supervised Machine learning classification algorithms are:

- Logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Naive Bayes. Each of these algorithms comes with their advantages and limitations which is very important to consider when choosing a model for a particular dataset. For example, RF model will perform better on datasets with an imbalance entry as it has the bootstrap sampling or bagging feature.

Data collected from the satellite are preprocessed using any of the available programming languages (mostly python, R and Matlab). Data exploration are the very first step to better understand the visual relationships between the variables and with that, it is easier to understand which kernel is best to be used to tune the model statistically. It is equally a good research approach to equally explore the relationship between the predictive variables to have a fair idea of which variable will have an important effect in the model prediction. In the case where Logistic regression or Support Vector model is to be used, it is important to check for the p-value, RME and the Variance Inflation Factor (VIF) to understand which variables will have or had the least contribution to the model prediction.

There are generally two approaches to this method that is, either using the supervised approach (Classification) or the Non-supervised approach (Clustering). In the supervised approach, data collected are first labelled as +HABs or –HABs then after, split into training and test datasets. The training datasets are used to build the model and the test datasets or sometimes called the validation datasets are used to validate or test the performance of the model by using it to make classification (in our case, +HAB or -HAB). The model or algorithms using the supervised algorithm purposely feed on the feature of the labelled training datasets to train itself under some parametric settings and optimal settings. The build model can be optimized using some statistical features of

the models to choose the best parametric values that gives an optimal prediction. This step is important as it seeks to minimize the problem of overestimation and underestimation. The model performance are then assessed using a confusion matrix which provides the model Accuracy, Sensitivity, Specificity, Positive predictive Value and Negative predictive value.

The non-supervised approach is somewhat different from the supervised approach as it does not require a pre-labelled data. The datasets are feed directly into the model for the model to use the characteristic features of the datasets to group the datasets into clusters. The model does the clustering by grouping all datasets that has similar features into one group (say +HABs) and putting the others into another group (say -HABs) as in our case. The validation and model optimization is equally carried out in similar way as in the sister approach as well as the model performance.

Decisions are made based on the outcome of the Confusion matrix parameters and the descriptive statistics determined during the data exploration and modelling. The general model performance is evaluated by comparing some numerical performance indices such as Accuracy, Sensitivity, Specificity, Precision, graphical evaluations using ROC and AUC, RME among many other indices. The validation of the model is generally considered to be an empirical validation of the ML model which serves as a conclusive summary of the potentiality of the model built. Statistical inferences can then be made with the results collected to serve as a power of support to any future test of any given hypothesis.

VII. Observed areas that Machine Learning could be applied to (Future)

The below suggestions are purely based the first months readings of various research works in the area of HAB and ocean event predictions.

PREPOSITION 1.

Most of the already available algorithm across the world lack a more generalized model for the monitoring and mapping of Ocean Events and HABs. Basically, the algorithms available are built per a specific algorithmic index which only applies to a particular region and is not applicable to

other regions because of the effect of atmospheric corrections and terrain corrections as well as, some factors which account for a change in the threshold.

The formulation of a generalized model for Ocean event detection or HABs detection could be achieved with a study of the distribution of HABs nationally. This study will allow the understanding of the distribution of the algorithmic index or thresholds across the nation. The said distribution will be a collective assessment of HABs or Ocean events nationally and using the parametric results as a new data, a nationwide algorithm could be built to predict the existence of Ocean Events and HABs.

VIII. Machine Learning involvement

Using a training data from different regions the distributive parametric index or threshold could be used to build a more Generalized Model which can be validated using field true data or in-situ data.

PREPOSITION 2.

Using machine learning models to estimate the Chl-a with water reflectance data.

PREPOSITION 3. (an improved study of preposition 2)

Machine learning approach for the detection of HAB using water reflectance data.

PREPOSITION 4.

Machine learning approach to predict ocean events using Chl-a and Fluorescent line height data as predictive factors.

.

References

1. Carvalho, G., Minnett, P., Fleming, L., Banzon, V. and Baringer, W., 2010. Satellite remote sensing of harmful algal blooms: A new multi-algorithm method for detecting the Florida Red Tide (Karenia brevis). *Harmful Algae*, 9(5), pp.440-448.

2. DUPOUY, C., PETIT, M. and DANDONNEAU, Y., 1988. Satellite detected cyanobacteria bloom in the southwestern tropical Pacific Implication for oceanic nitrogen fixation. *International Journal of Remote Sensing*, 9(3), pp.389-396.

3. Simon, A., & Shanmugam, P. (2012). An Algorithm for Classification of Algal Blooms Using MODIS-Aqua Data in Oceanic Waters around India. *Advances In Remote Sensing*, *01*(02), 35-51. doi: 10.4236/ars.2012.12004

4. DYBAS, C. (2003). Harmful Algal Blooms: Biosensors Provide New Ways of Detecting and Monitoring Growing Threat in Coastal Waters. *Bioscience*, *53*(10), 918. doi: 10.1641/0006-3568(2003)053[0918:habbpn]2.0.co;2

5. Gokul, E., Raitsos, D., Gittings, J., Alkawri, A., & Hoteit, I. (2019). Remotely sensing harmful algal blooms in the Red Sea. *PLOS ONE*, *14*(4), e0215463. doi: 10.1371/journal.pone.0215463

6. Gokul, E., & Shanmugam, P. (2016). An optical system for detecting and describing major algal blooms in coastal and oceanic waters around India. *Journal Of Geophysical Research: Oceans*, *121*(6), 4097-4127. doi: 10.1002/2015jc011604

7. Sathyendranath, S., Cota, G., Stuart, V., Maass, H., & Platt, T. (2001). Remote sensing of phytoplankton pigments: A comparison of empirical and theoretical approaches. *International Journal Of Remote Sensing*, *22*(2-3), 249-273. doi: 10.1080/014311601449925

8. Blondeau-Patissier, D., Gower, J., Dekker, A., Phinn, S., & Brando, V. (2014). A review of ocean color remote sensing methods and statistical techniques for the detection, mapping and analysis of phytoplankton blooms in coastal and open oceans. *Progress In Oceanography*, *123*, 123-144. doi: 10.1016/j.pocean.2013.12.008

9. Havens, K. (2018). The Future of Harmful Algal Blooms in Florida Inland and Coastal Waters. *EDIS*, *2018*(1), 4. doi: 10.32473/edis-sg153-2018

10. Oyama, Y., Matsushita, B., Fukushima, T., Chen, J., Nagai, T., & Imai, A. (2010). Testing the spectral decomposition algorithm (SDA) for different phytoplankton species by a simulation based

on tank experiments. *International Journal Of Remote Sensing*, *31*(6), 1605-1623. doi: 10.1080/01431160903475365

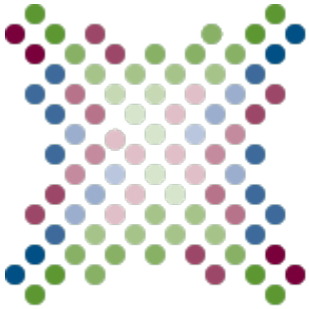
11. Tomlinson, M., Stumpf, R., Ransibrahmanakul, V., Truby, E., Kirkpatrick, G., & Pederson, B. et al. (2004). Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico. *Remote Sensing Of Environment*, *91*(3-4), 293-303. doi: 10.1016/j.rse.2004.02.014

12. Gokaraju, B., Durbha, S., King, R., & Younan, N. (2011). A Machine Learning Based Spatio-Temporal Data Mining Approach for Detection of Harmful Algal Blooms in the Gulf of Mexico. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*, *4*(3), 710-720. doi: 10.1109/jstars.2010.2103927

13. Pitcher, G., & Louw, D. (2020). Harmful algal blooms of the Benguela eastern boundary upwelling system. *Harmful Algae*, 101898. doi: 10.1016/j.hal.2020.101898

14. Song, W., Dolan, J., Cline, D., & Xiong, G. (2015). Learning-Based Algal Bloom Event Recognition for Oceanographic Decision Support System Using Remote Sensing Data. *Remote Sensing*, *7*(10), 13564-13585. doi: 10.3390/rs71013564

15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. doi: 10.1007/bf00994018



FRCT

FUNDO REGIONAL PARA A CIÊNCIA E TECNOLOGIA