Clustering Sargassum Mats from Earth Observation Data

Estèle Glize, Marie-José Huguet, Marc Lucas, Marion Sutton, Gilles Trédan

Clustering Sargassum Mats from Earth Observation Data*

Estèle Glize[1], Marie-José Huguet[1], Marc Lucas[2], Marion Sutton[2], and Gilles Trédan[1]

[1] LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
{glize,huguet,tredan}@laas.fr
[2] CLS, France
{mlucas,msutton}@groupcls.com

Abstract. Sargassum seaweed forms large floating mats drifting on the oceans. These mats are increasingly beaching on Caribbean islands, threatening the local wildlife and economies. This paper focuses on their tracking from space, in order to monitor these mats. More specifically, we focus on clustering sargassum mats on satellite images. This constitutes an important building block of the sargassum monitoring system, which then predicts the drift of those clusters to anticipate beachings and warn the local authorities. The difficulty of the clustering operation comes from the noisy nature of input data: image artefacts, partial cloud occlusion, mats discontinuities due to sea conditions, and the difficulty of acquiring ground truth data. This paper details our approach to overcome those challenges. We propose a method (hereafter named Sargassum Mats Detection Method - `SMDM`) that improves the mats identification by combining a first artefact detection step, a tailored clustering algorithm and a region growing algorithm.

Keywords: Unsupervised Machine Learning · Clustering · Sargassum Detection · Earth Observation Data.

1 Introduction and Industrial Context

Sargassum seaweed (sargassum fluitans and sargassum natans) has always been present in the Atlantic Ocean. These floating algae have the particularity of forming large mats, sometimes reaching over 10 kilometers in length. These mats drift around the ocean, driven by currents and winds. They are home to a unique ecosystem and are a breeding ground for many marine species. However, when these large mats make landfall, the algae vegetation decomposes, resulting in the atrophication of the nearshore waters and the release of harmful sulfur dioxide gases. The sargassum seaweed invasion came to the fore from 2011 onwards when large amounts of mats started landing on the coastlines of the Caribbean islands and the central American countries. The increasingly heavy impact on the livelihood of the Caribbean societies spurred the scientific community into investigating the phenomena.

* This work was funded by the CNES (Centre National des Etudes Spatiales)

Challenges in data exploitation. In 2015, CLS, with the financial support of ESA, started an operational system to provide a variety of users with high quality mat location data using Earth Observation. The project worked on the computation of an improved algae index for each pixel of observation data and on enabling a forecast capability. By using wind and current data, mats movements are estimated to identify the most likely beaching zones. To this end, a drift model is fed with the current location of the mats. Although sargassum mats are easy to spot using the human eye, automating the mats detection is a challenge due to the noisy nature of the observation data. Furthermore, the behaviour of the mats is heavily impacted by environmental conditions as sections of mats might be temporarily submerged in certain wind and wave conditions or not as visible due to the satellite sensor angle.

Contributions. This paper focuses on unsupervised machine learning approach. We propose a method, named Sargassum Mats Detection Method (`SMDM`), to cluster the distribution of pixels of individual satellite images into a set of sargassum mats.

This paper is organized as follows. Section 2 details the studied problem and reviews some of the state of the art methods to achieve clustering. Our approach, `SMDM`, is presented in Section 3 and its performances are evaluated in Section 4. Then, we report conclusions of this study and list some perspectives in Section 5.

2 Problem statement and unsupervised learning

The CLS approach relies on the analysis of satellite images to detect the presence of sargassum mats and to simulate their drift. It is based on 3 steps. First, a data processing step that computes the algae index. Second, a clustering step that detects mats. Finally, a drift simulation step simulates the drifting of detected mats to forecast their stranding.

Data processing. In this paper, we focus on the sargassum mats detection step. Before detailing this step, we briefly present the data that will be used as input. Sargassum presence is detected by the increase of the reflectance spectrum between the red and near infra-red (`NIR`) wavelengths. The `NDVI` (Normalized Difference Vegetation Index) has been widely used in the past to map vegetation over land surfaces. Over the ocean, it has been successfully used with MODIS 250-m resolution data to study a severe floating green algae bloom event in the Yellow Sea [6]. Well-known sargassum indices were proposed in the literature, for example the Maximum Chlorophyll Index [4], the Floating Algae Index (`FAI`) [5], the Alternative Floating Algae Index [9], follows the same mathematical statement : `AFAI`$= \rho_{NIR} - \rho'_{NIR}$ where ρ_{NIR} denotes a reflectance, partially (or not) corrected for atmospheric effects in the `NIR` band, and ρ'_{NIR} is the equivalent `NIR` reflectance that would be measured at the same point in absence of sargassum. ρ'_{NIR} is approximated by a linear interpolation between the two reflectances measured at nearby wavelengths in the red and short-infrared bands. We use

here a normalized version of the FAI, in which the normalization by the sum of reflectances is introduced to mitigate the variability of the FAI due to atmospheric conditions and observation geometry, as done for the NDVI over land surfaces: NFAI= $(\rho_{NIR} - \rho'_{NIR})/(\rho_{NIR} + \rho'_{NIR})$. Derivation of floating vegetation presence using NFAI seems thus straightforward. However, a lot of wrong detections are found mostly in the vicinity of clouds. It is thus necessary to implement an automatic editing procedure to remove them. The drawback of such automatic procedure is to also remove true detections. At the end of the data processing step, to each pixel is associated an indicator:

- P to represent the potential presence of algae: P pixel represents the non edited values of NFAI
- A to guarantee the presence of algae: A pixel is a pixel with the edited values of NFAI;
- S when the pixel corresponds to the sea;
- C indicates the presence of clouds or unspecified data.

Throughout this paper, we will illustrate the different steps with two satellite images (part of the Caribbean Sea) taken from MODIS 250-m sensor: Map1 (long: [-60.8,-58.8], lat:[15.2,16.7], day: 10/07/19) and Map2 (long: [-63.2,-61.2], lat: [13.9,15.4], day: 10/07/19) These images are grids of [801$\dot{6}$01] pixels. The Figure (1) gives an illustration of Map1 with indicator on each pixel.

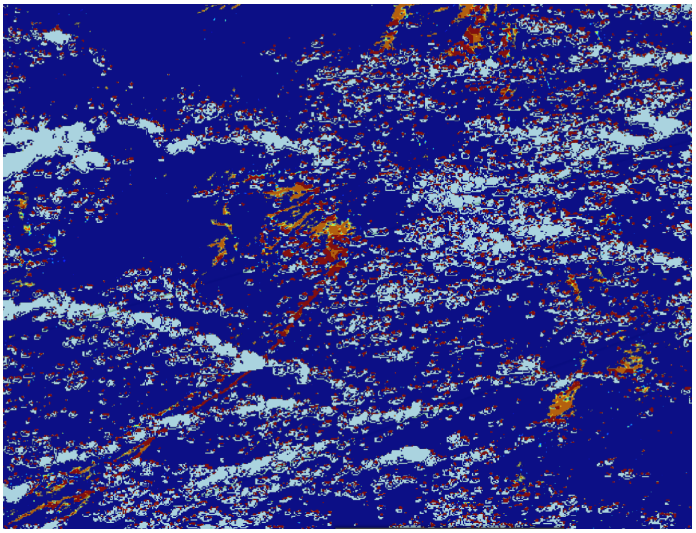


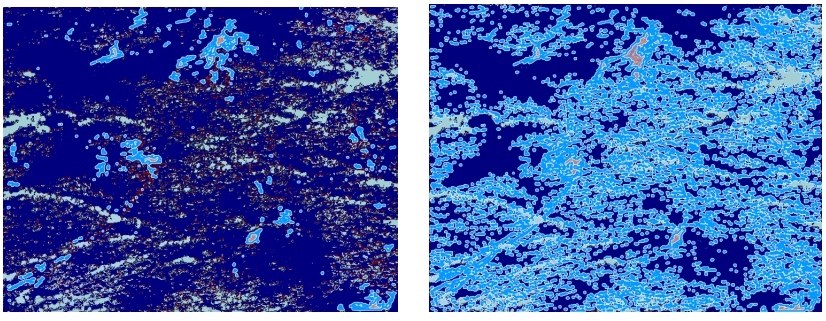
Fig. 1: Map1-Input of SMDM. Pixels A in Yellow; P in Red; S in Blue; C in White.

Based on the pixel grid with these indicators, the aim is to find a method to efficiently identify sargassum mats to compute their drift and forecast their strandings. The characteristic shape of these mats is a long and elongated shape,

and they are possibly made up in parallel stratum. Even if mats are quite easy to spot using the human eye (especially for oceanographic experts), an automatic detection method faces several obstacles:

1. Mats can be composed of only A pixels, only P pixels, or both.
2. The dataset of images presents artefacts due to multiple scattering in the near infrared occurring in the close vicinity of clouds. These artefacts induce falsely positive values (some of the P pixels). They are characterized by isolated and small groups of P pixels. Other artefacts such as those arising from the satellite angle, the proximity of land as well as the sun glint effect are beyond the scope of this paper.
3. Some C or S pixels may actually divide sargassum mats (with clouds or mats partially submerged). Reconstruction of whole mats by putting together discontinuous clusters should also avoid overestimating the size of mats.
4. Algae density is highly variable, both in different parts of a given image and from one image to another.

Algae contouring. A standard contouring method, the marching squares algorithm (`MS`), was implemented before this study. `MS` is a contouring method commonly used, for instance for generating isobars in weather maps. As a generic method, `MS` is easy to use and requires little parametrization. In `MS`, only A pixels are used to identify objects that will be associated with sargassum mats. However, A pixels are often disconnected from each other, and this approach produces many contours of small objects. The use of P pixels (or A and P pixels) faces the same limitation: the number of contouring objects is too high and mats appear as multiple independent clusters, which is then problematic for drift simulation. Some sargassum mats can be filtered (for example if they do not contain a sufficient number of A pixels), nevertheless, for the drift step, using such contouring algorithm implies to consider an important set of pixels in input. Figure (2) presents the limitation of `MS` on A pixels (loss of information) and on both A and P pixels (too many small clusters).



(a) With only A pixels. (b) With A and P pixels

Fig. 2: Illustration of `MS` on Map1. The convex hull of sargassum mats are in light blue.

Learning methods. Image segmentation methods identify clusters of pixels with similar properties to detect regions or objects into an image. These methods are less sensitive to spatial discontinuity than thresholding methods, and they may overcome the issues of `MS`. Among image segmentation methods, supervised learning is growing in popularity, particularly for object detection. Convolutional Neural Networks are really effective for visual recognition issues which assign an object to an image [7]. For sargassum algae detection, [1] develop a neural network, called ERISNet, to determine if a pixel in a satellite image contains sargassum algae. They train their network on labelled data where a pixel is a 1km square of the Mexican coast. This work is an alternative to the `NFAI` index computation, but doesnt detect sargassum mats. Supervised learning methods require building a set of annotated data (ground truth) to train and validate the neural network. For sargassum mats, there is no ground truth associated with each image. This represents a huge amount of specific work, so we do not select this class of machine learning methods for our sargassum mat detection.

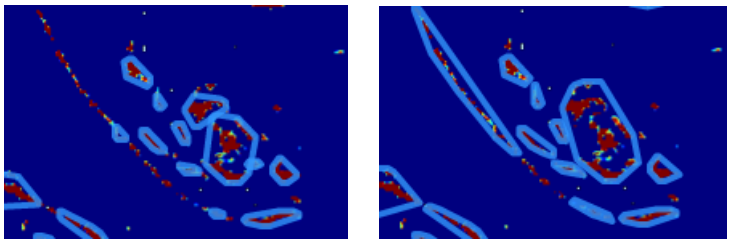
Unsupervised learning methods can group data with similar characteristics from unlabelled data. The interesting features for grouping data are decided upstream by the designer of the algorithm, then the clustering is automatic without outside intervention. For instance, [2] use a k-means method fixing k to 5 to detect sargassum mats. Despite the computational efficiency of the algorithm, fixing the number of sargassum mats upstream in an image doesnt seem appropriate to our problem as the number of mats is unpredictable. Furthermore, this algorithm is noise-sensitive and will affect each P pixel to a mat, even if they are isolated. `DBSCAN` [3] is another unsupervised learning method which considers that clusters are connected dense regions in the data space. The algorithm extracts one unvisited random pixel p in the image, and marks it as visited. If p has at least m neighbors within a distance ϵ, these neighbors are in the same cluster as p and they are recursively visited to check their own ϵ-neighborhood. Otherwise, p is considered noise (without a cluster). This process continues until there is no unvisited pixel in the image. Therefore, `DBSCAN` requires the setting of two parameters: ϵ and m. In our case study, the method has several advantages: the computational efficiency regardless of the size and the shape of the clusters, the identification of noise, the flexibility to the varying number of clusters and the relatively insensitivity to initialization. Because of all the advantages of `DBSCAN` over other methods, we will use `DBSCAN` in `SMDM` presented in the next Section.

3 Proposed approach

This section presents the processing chain of `SMDM` to detect sargassum mats on the high seas. The input to this pipeline is the pixel grid, presented in Section 2, containing the indicators P (potential presence of algae), A (algae), S (sea) and C (cloud or unspecified data). `SMDM` consists of three main steps described hereafter.

Step 1: Remove artefacts. We consider small clusters that do not contain A pixels as artefacts. This step starts with a clustering on the pixels with potential presence of algae (P pixels), and then processes each individual cluster to identify and remove those originating from artefacts. The clustering is performed by `DBSCAN` on the P pixels with the parameters set to focus the detection on small sized clusters: $\epsilon = 2$ et $m = 5$. This enables the detection of clusters of at least 5 neighboring P pixels. A cluster is then considered as an artefact if it doesn't contain A pixel (guaranteed presence of algae) and if it is sufficiently small (in practice, at most 30 pixels). At the end of this first step, all the presumed artefacts (false positive P pixels) are deleted.

Step 2: Identify Sargassum Clusters. In this step, we identify sargassum clusters in two stages: clusters detection and clusters extension.
Detection. The clustering algorithm `DBSCAN` is used again on the P pixels with slightly relaxed parameters. We set $\epsilon = 2.8$ and $m = 12$ to identify larger clusters than in Step 1. All the identified clusters are given at the extension stage.
Extension. As `DBSCAN` relies on circle-based exploration, elongated clusters (which are quite common) may not be fully detected. To compensate for this effect, we extend the clusters identified in the previous detection phase. Sargassum clusters affected by this step are the continuous and thin ones that are partially detected. For instance, Figure (3) illustrates this problem: while discontinuous mats of circular shape are correctly detected after the identification step (left) and unaffected by the extension (right), elongated and continuous mats are only partly identified by `DBSCAN`. This partial detection is therefore extended. This stage uses the region growing method [8]. More precisely, for each cluster $\mathcal{G}$, we inspect each pixel $p \in \mathcal{G}$. If p is adjacent to another P or A pixel p' that is not in $\mathcal{G}$, we add p' to $\mathcal{G}$. We recursively inspect adjacent pixels of p' until we find no more adjacent P or A pixel. If the growing of a cluster $\mathcal{G}$ leads to add pixels of another identified cluster $\mathcal{G}$, both clusters are merged: $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{G}$.



(a) After detection. (b) After detection and extension.
Fig. 3: Illustration of Step 2. The convex hull of sargassum clusters are in light blue.

Step 3: Aggregate Clusters into Mats. We extend the clusters from the previous step to find final sargassum mats. Only sufficiently stretched and large clusters are extended. Both criteria were derived from the experts perspective on sargassum mats: large and thin mats. The elongation of a cluster is computed with a

Principal Component Analysis on the pixels of this cluster. It gives the direction in which the cluster is the most elongated, as well as the elongation rate on this axis between 0.5 and 1.0. A rate of 0.5 represents a rounded cluster whereas a rate of 1.0 represents a thin and elongated cluster. To extend a cluster, its elongation must be greater than 0.8, and it must consist of at least 50 pixels.

The extension of a cluster is made using a slightly modified region growing method to assemble discontinuous clusters. Therefore, the clusters affected by this step are those partially separated by a few pixels of sea or clouds (C or S pixels). More precisely, for a cluster $\mathcal{G}$, we consider each pixel $p \in \mathcal{G}$, and we inspect each pixel p at 2 pixels from distance to p. If p is a P, A or C pixel and is at less than 5 pixels from the direction of elongation of $\mathcal{G}$, we add p to $\mathcal{CL}$. Then, we recursively inspect adjacent pixels of p' in the direction of elongation of $\mathcal{G}$ until we find only S pixels. If the growing of a cluster $\mathcal{G}$ leads to add pixels of another identified cluster $\mathcal{G}$, both clusters are merged: $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{G}$.

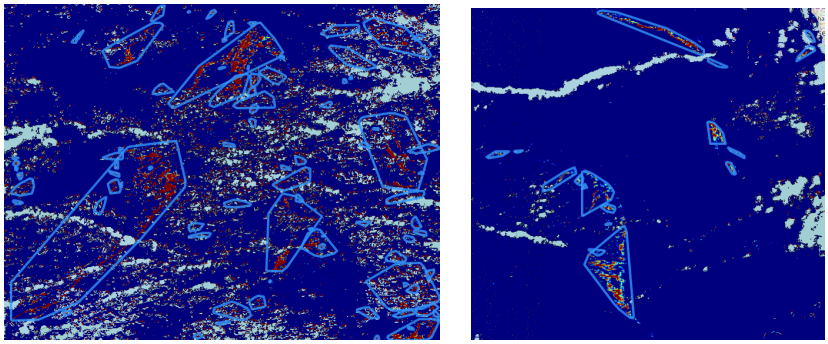
Algorithm Output and Post-processing. A post-processing stage allows to attribute a confidence level to sargassum mats. We have defined three confidence levels sorted by decreasing likelihood. A pixel in a mat of a level is necessarily in a mat of an upper level.

- Level 1 regroups the sargassum clusters found at the end of Step 2 containing at least one A pixel.
- Level 2 considers all sargassum mats found at the end of Step 3, but it removes all pixels located less than 10 pixels from a C pixel and which are not in a mat of the confidence level 1. In this level, a mat with no A pixel can appear, but it removes the uncertainty of algae near clouds.
- Level 3 regroups all sargassum mats found at the end of Step 3. In the confidence level 3, a mat with no A pixel can appear, and mats are more likely to cross clouds or sea.

4 Evaluation

In this section, we present our approaches to validate our solution. The major difficulty of this validation is the absence of ground truth data against which we could quantitatively compare. We circumvent this problem by relying on two complementary approaches. First, we rely on the qualitative feedback from oceanography experts that were involved throughout the process. Second, we compare against `MS`, which constitutes the baseline used in production over some metrics capturing the size and the elongation of identified clusters.

Expert Input. Since there exists no ground truth regarding sargassum mats, validation of the previous approaches has been conducted using expert input. More precisely, oceanographers leveraging both their sargassum and earth observation knowledge are presented with a visualisation of the detected mats. Figure (4) presents the results of our solution as presented to the experts. While some artefacts remain (e.g. East of both maps), the detection avoided the majority (e.g.



(a) Map1 results. (b) Map2 results.

Fig. 4: Example output on Maps 1 & 2. Detected mats are circled in light blue.

Center of Map1). Moreover, SMDM manages to identify clear elongated mats (e.g. North of Map2), even in challenging conditions like partial cloud occlusions (e.g. South-West are of Map1).

Comparing against Marching Squares. Prior to the deployment of our solution, sargassums mats were clustered using the MS algorithm. Because of its historical position and the absence of other competitor, MS is considered as the baseline hereafter. Since MS does not produce confidence levels, to ensure a fair comparison we set the "High confidence Level" of MS as the result of MS on A pixels only, and the "Low confidence Level" as the result of MS on P pixels. Similarly, we focus on the two highest confidence levels of SMDM.

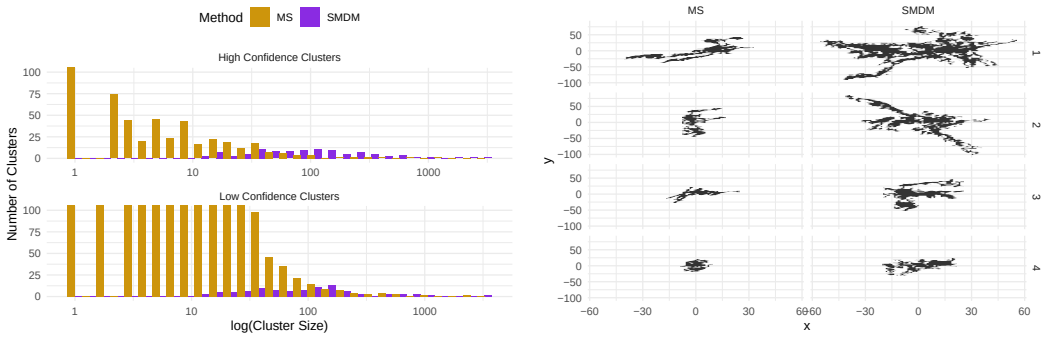
Table 1 compares both approaches on Maps 1 and 2. Recall that Map 1 is paradigmatic of an occluded region containing many clouds and many artefacts, whereas Map 2 represents a good situation with few clouds and clear mats. First, SMDM is slightly slower than MS on A pixels. This originates from the processing time for steps 1 and 3 in SMDM. On the bright side, SMDM generates all confidence levels within this time, whereas MS, when run on P pixels takes considerably longer. This gap in MS performance is explained by the number of clusters it identifies. For instance, on low confidence level in Map 1, MS identifies around 80 times more clusters than SMDM. The identified clusters are all very small. In contrast, SMDM identifies a small number of large clusters, even at the high confidence level.

Figure (5) sketches an exploratory analysis of the identified clusters. The left part represents the distribution of cluster sizes identified on Maps 1 and 2 for both methods. It confirms the previous results: the vast majority of the many clusters identified by MS are below 100 pixels, whereas the SMDM clusters are nearly always above 100 pixels. A careful comparison of SMDM in Low and High confidence levels allow to observe that SMDM traded some 500 High confidence clusters for fewer large (> 2000 pixels) Low confidence clusters.

Overall, those results shows SMDM meets its objectives: identify less clusters of bigger size. While interacting with the experts and exploring the results of

Confidence	High				Low			
Map	1	1	2	2	1	1	2	2
Method	MS	SMDM	MS	SMDM	MS	SMDM	MS	SMDM
Time (s)	22	33	6	8	266	-	64	-
Pixels in clusters	7536	16522	673	6022	66413	17882	11986	3849
Number of clusters	308	88	166	16	6155	79	1127	16
Elongation (mean)	0.85	0.80	0.88	0.90	0.81	0.80	0.83	0.88
Elongation (std.)	0.12	0.11	0.14	0.07	0.13	0.12	0.14	0.11
Size of clusters (mean)	24	188	4	376	10	226	10	241
Size of clusters (std.)	88	379	4	572	42	525	71	314

Table 1: Comparison of the proposed approach against `MS` on Maps 1 & 2.



(a) Size

(b) Layout of the 5 largest detected mats

Fig. 5: Statistics of the identified clusters, depending on Confidence Level, on Maps 1 & 2. The number of clusters is limited to 100 for readability. Maximum number of clusters for MS is over 900 in the Low Confidence level. The layout of the largest clusters is for high confidence clusters on Map 1.

both methods, we used 6 additional maps that are not presented here as they confirm the above conclusions without bringing further insights.

5 Conclusion

To sum up, this paper presents `SMDM`, an unsupervised clustering approach for Sargassum mats detection. It consists in a sequence of steps to i/ detect and remove artefacts from the data ii/ cluster the cleaned data conservatively and iii/ grow the clusters into mats by exploiting the specificities of their geometry. This approach is then compared against current SoA method, and its benefits are illustrated on two test maps. Dealing with the absence of ground truth data has been the central challenge of Sargassum detection, both for the design step (restricting the approach to unsupervised methods) and for the validation step. Hence a first promising research direction is to explore the temporal dimension of the problem. Since two consecutive images should contain reasonably similar geographic distributions of mats, exploiting these similarities could make up for a ground truth proxy. However, tracking a sargassum mat across multiple shots

is not trivial due to drifts, cloud movement and varying satellite angles. Driven by the growing number of freely accessible Earth Observation images, a second research direction consists in applying `SMDM` to other oceanographic features, in particular river plumes and chlorophyll filaments.

Acknowledgements The authors would like to thank Jacques Stum for kindly explaining the data processing step and the `NFAI` index computation.

References

1. Arellano-Verdejo, J., Lazcano-Hernandez, H.E., Cabanillas-Terán, N.: ERISNet: deep neural network for Sargassum detection along the coastline of the Mexican Caribbean. PeerJ **7**, e6842 (2019)
2. Bernard, D., Biabiany, E., Sekkat, N., Chery, R., Cécé, R.: Massive stranding of pelagic sargassum seaweeds onthe french Antilles coasts : Analysis of observed situations with Operational Mercator global oceananalysis and forecast system. 24ieme Congrès Francais de Mécanique (2019)
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
4. Gower, J., King, S., Borstad, G., Brown, L.: Detection of intense plankton blooms using the 709 nm band of the MERIS imaging spectrometer. International Journal of Remote Sensing - INT J REMOTE SENS **26**, 2005–2012 (05 2005)
5. Hu, C.: A novel ocean color index to detect floating algae in the global oceans. Remote Sensing of Environment **113**(10), 2118 – 2129 (2009)
6. Hu, C., He, M.X.: Origin and Offshore Extent of Floating Algae in Olympic Sailing Area. Eos, Transactions American Geophysical Union **89**(33), 302–303 (2008)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
8. Mehnert, A., Jackway, P.: An improved seeded region growing algorithm. Pattern Recognition Letters **18**(10), 1065–1071 (1997)
9. Wang, M., Hu, C.: Mapping and quantifying Sargassum distribution and coverage in the Central West Atlantic using MODIS observations. Remote Sensing of Environment **183**, 350 – 367 (2016)