

Mapping marine phytoplankton assemblages from a hyperspectral and Artificial Intelligence perspective

E. Torrecilla, J. Piera, S. Pons, I. F. Aymerich
Marine Technology Unit, UTM
Consejo Superior de Investigaciones Científicas, CSIC
Barcelona, Spain E-08003
{torrecilla, jpiera, spon, ismaelf}@utm.csic.es

A. Vilamala, J. Ll. Arcos, E. Plaza
Artificial Intelligence Research Institute, IIIA
Consejo Superior de Investigaciones Científicas, CSIC
{vilamala, arcos, enric}@iia.csic.es

Abstract—The aim of this contribution is to demonstrate the feasibility of different processing techniques to identify phytoplankton assemblages when applied to oceanographic hyperspectral data sets (i.e. above surface measurements and vertical profiles). In order to address this issue and validate the proposed techniques, a simulated framework has been used based on the oceanic radiative transfer model Hydrolight-Ecolight 5.0. The potential offered by an unsupervised hierarchical cluster analysis technique and two Artificial Intelligence algorithms (i.e. Particle Swarm Optimization and Case-Based Reasoning) have been explored. Our results confirm their suitability to map phytoplankton's distribution from hyperspectral information given a variety of hypothetical oceanic environments.

I. INTRODUCTION

In oceanography, the increasing availability and use of optical measurements has provided the potential for better assessing information regarding optically significant ocean water constituents, especially the phytoplankton component [1]. Optical information has been used extensively in oceanography to retrieve ocean surface estimates of chlorophyll *a* concentrations, which has been commonly used by marine biologists as an index of the algal biomass. However, beyond the chlorophyll *a* concentration which is common to all taxonomic groups of phytoplankton, new optical instrumentation deployed on the oceanographic observing platforms offers now the potential to also map the phytoplankton community composition in the ocean [2]. In particular, recent advances in monitoring surface and underwater optical properties have been related to the progressive shift from using multispectral to hyperspectral resolution acquisition systems (i.e. spectral resolution down to few nanometers instead of several discrete spectral bands). The ability to identify different oceanic phytoplankton assemblages stems from the fact that variability on optical properties from each algal group can be better assessed by the use of higher spectral resolution sensors. The higher spectral resolution enables the extraction of more detailed spectral information which is strongly related to absorption features of water constituents, specifically pigments present in the samples. Among the marine applications that benefit from hyperspectral observations are monitoring of recurrent events such as harmful algal blooms (HABs) or detection of phytoplankton patches known as 'thin layers', which range

in thickness from 10 cm to 5 m and extend horizontally for several kilometers in the ocean [3], [4].

The radiometric approach either with single-, multi- or hyperspectral resolution has become essential for monitoring water and phytoplankton dynamics. It represents a non-invasive and less time consuming option in comparison to traditional observing procedures used in marine biology. Typically, the conventional analysis of a single water sample through the use of microscopic cell counts requires on the order of several hours to be completed. It contrasts with spectroradiometer's performance given that these sensors are able to acquire an entire spectrum over some finite spectral region in less than few microseconds. Another advantage of spectral-based instrumentation is that it can be used to characterize phytoplankton distributions from large scale patterns if sensors are placed on remote-sensing observing platforms (e.g. satellite, aircraft, etc.) to small scale structures if in situ platforms such as high spatial resolution profilers, autonomous underwater vehicles or gliders are considered.

To improve our understanding of phytoplankton dynamics in the ocean, the Marine Technology Unit (UTM) in Barcelona, is developing an intelligent oceanographic probe with high resolution autonomous sampling and collecting capabilities within the framework of the multidisciplinary project ANERIS [5]. The ANERIS profiling system is mainly being designed: (1) to gather fine-scale profiles of biological structure by using hyperspectral sensors during the free-fall descent, (2) to detect phytoplankton structures of interest by processing the whole set of hyperspectral distributions along depth and (3) to autonomously collect discrete water samples within significant depths in terms of phytoplankton composition during the ascent by incorporating a number of automatically-triggered bottles into its design (Figure 1). This multi-instrumented high resolution system will be able to measure optical properties in the upper meters of the water column, while simultaneously collect water samples for laboratory analysis. This mode of operation is intended to be useful for studying coastal environments, covering a wide range of temporal and spatial scales where different events may happen [6] and even validating remotely-sensed hyperspectral ocean data collected by airborne imagers.

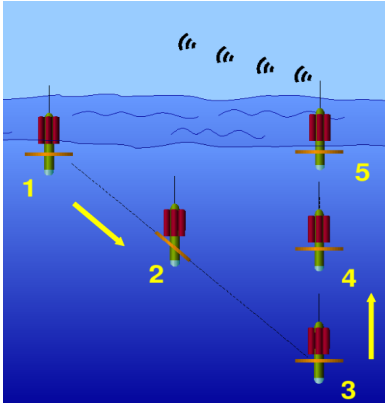


Fig. 1. Schematic view of the ANERIS profiling system.

The amount of oceanographic hyperspectral data sets available and collected by different types of platforms will increase in the near future. In that sense, appropriate processing strategies must be undertaken in order to perform its further analysis. The aim of this contribution is to demonstrate the feasibility of different spectral processing techniques to identify different phytoplankton assemblages or patches when applied to oceanographic hyperspectral measurements, like those that will be obtained in the field by the above described ANERIS probe. First, the potential offered by an unsupervised hierarchical cluster analysis (HCA) technique to discriminate phytoplankton assemblages from hyperspectral measurements given different hypothetical oceanic environments is shown. HCA technique is explored in combination with the commonly used derivative spectroscopy in order to enhance subtle features in shape of spectral data before performing the cluster analysis. Furthermore, successful application of two Artificial Intelligence (AI) techniques is also described. A multiple Particle Swarm Optimization (mPSO) algorithm and a Case-Based Reasoning (CBR) approach provide a robust identification of thin layers of phytoplankton. mPSO algorithm is applied to automatically detect optically significant peaks in the spectral information even when the input signal suffers from high noise levels. CBR permits to characterize thin layer's dominant algae group as well as its thickness and level of concentration.

This study is organized as follows: First, a simulation framework is established to generate different underwater scenarios and several processing techniques are reviewed in section II. Next, some results and discussion are shown in section III. Finally, section IV includes a summary and conclusions of the developed research.

II. DATA AND METHODS

A. Radiative Transfer Simulations

In order to validate the proposed techniques before applying them to data from the field, a simulated framework was used based on the oceanic radiative transfer model Hydrolight-Ecolight 5.0 (HE 5.0) [7]. The simulated-based methodology of this research is presented schematically in Fig. 2. Different radiance distributions through the water column and derived

quantities were computed to generate different surface and underwater optical scenarios. Numerical simulations in controlled environments were performed taking into account the constituents present in the water column (e.g. pure water, phytoplankton, etc.) at different concentrations and their inherent optical properties (IOPs) such as absorption and scattering spectral coefficients [8]. Other environmental parameters (i.e. ocean surface wind speed, sun zenith angle, cloud coverage) were also defined. Upwelling and downwelling irradiances (E_u , E_d), which are apparent optical properties (AOPs), were therefore obtained and processed with the aim of automatically identifying phytoplankton distributions considered for each generated scenario.

In particular, three different scenarios were simulated according to the following criteria:

Scenario 1

The first hypothetical underwater optical scenario was modeled presenting two near-surface structures of phytoplankton at different depths (i.e. 6 and 12 m depth, respectively). Such thin layers were modeled to be dominated by single phytoplankton groups, i.e. diatoms or cyanobacteria. Concentration profiles for algae groups causing each layer are depicted in Fig. 3 (top panel), with a maximum value of chlorophyll of 20 mg/m^3 . The remaining input modeling parameters were set as follows: horizontally homogeneous ocean, spectral range from 350 to 750 nm, sun at 30° from the zenith, 5 m/s wind speed and inelastic internal sources of light included (i.e. fluorescence and Raman scattering). A total of 300 spectral reflectance distributions, $R(\lambda)$, were obtained and further processed from estimates of $E_d(\lambda)$ and $E_u(\lambda)$, which were provided by the model at depths from 0 to 15 m, every 5 cm (Fig. 3, bottom panel). $R(\lambda)$ is a relative measure resulting from the ratio of $E_d(\lambda)$ and $E_u(\lambda)$, that is considered more robust in front possible external variations during experimental measurements in the field.

Scenario 2

Different open water masses distributed horizontally in space and each dominated by a single phytoplankton group at a very low concentration were simulated. The hyperspectral data set, covering a wide range of environmental conditions, was created using absorption and scattering properties from six different phytoplankton groups (e.g. diatoms, cyanophyceae, etc.). Figure 4 (top panel) shows their distinct specific absorption spectra, which were used to generate a total of 24 water column scenarios by combining different levels of concentration (i.e. 0.03, 0.05, 0.07 and 0.09 mg/m^3). The following set of assumptions were also considered: water column vertically homogeneous and infinitely deep, sun at 30° from the zenith, 5 m/s wind speed, inelastic scattering included. After running the model, above surface hyperspectral remote-sensing reflectances ($R_{rs}(\lambda)$) from 350 to 750 nm, 1 nm spectral resolution, were obtained (Fig. 4, bottom panel).

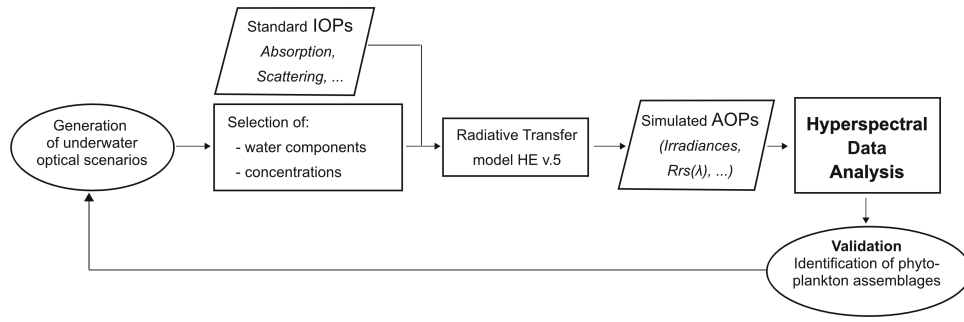


Fig. 2. A diagram showing the oceanic radiative transfer modeling and hyperspectral data analysis methodology.

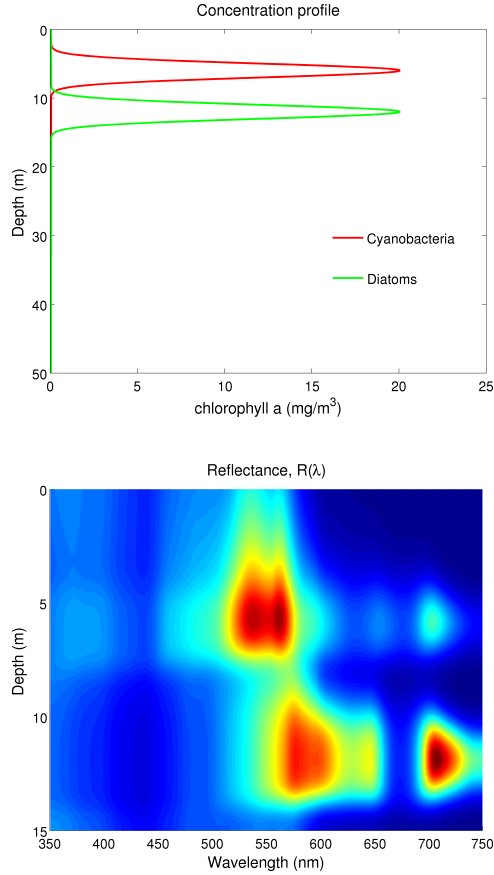


Fig. 3. (top) Vertical distribution along the water column of each considered species of phytoplankton: cyanobacteria and diatoms. (bottom) Top view of simulated hyperspectral reflectance spectra, versus depth and wavelength (0.05m spatial resolution and 1 nm spectral resolution), corresponding to the water column scenario shown on top panel.

Scenario 3

This scenario consisted of 300 simulations (profiles) generated assuming the following common parameters: a dominating algae group along the water column (e.g. type A, B, C or D), horizontally homogeneous ocean, sun at 30° from the zenith, 5 m/s wind speed, inelastic scattering included and spectral range from 400 to 650 nm. However, each simulation was characterized by a unique concentration profile for the phytoplankton component, which presented two peaks with different maximum depth (i.e. from 4 to 12 m), thickness (i.e.

from 0.4 to 1.4 m) and maximum value (i.e. from 4 to 20 mg/m³). By modifying the concentration profile settings (i.e. combining different depth, thickness and maximum values for both peaks), different stratified scenarios were thus created and analyzed from surface to 15 m depth, with an spatial resolution of 5 cm. In particular, this batch of profiles was processed by AI techniques described below, considering the optical property of attenuation coefficient ($K_d(\lambda)$). K_d spectral distributions at each depth were estimated from the spectral downwelling irradiances ($E_d(\lambda)$) provided by the model. In addition, several noise levels (up to 50%) were applied to $K_d(\lambda)$ raw data for testing AI approach's robustness.

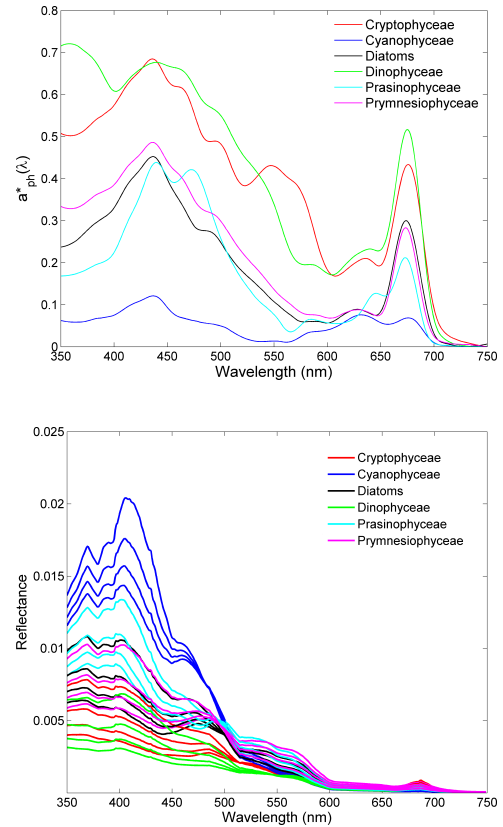


Fig. 4. (top) Specific absorption spectra of different phytoplankton groups. (bottom) Hyperspectral remote-sensing reflectance spectra, corresponding to different open water scenarios dominated by a single phytoplankton species (identified by different colors) at different concentrations.

B. Cluster analysis

A hierarchical cluster analysis (HCA) was carried out for classifying the simulated scenarios into different phytoplankton pigment assemblages based on each hyperspectral data set. HCA methodology is a common unsupervised classification algorithm, the details of which are described in [9]. Each cluster tree is obtained using a linkage algorithm based on a previously calculated pairwise distance between all objects included in the input data set. In this study the similarity between each pair of objects (i.e. spectra) was always calculated using the cosine distance i.e. one minus the cosine of the included angle between the two objects. And as a linkage algorithm the shortest distance, also called nearest neighbor, was computed and used to measure the distance between two clusters of objects in the tree. The traditional representation of this hierarchical tree is a dendrogram, with individual objects at one end and a single cluster containing all objects at the other.

In our case, pairs of spectra showing a small cosine distance between them, i.e. similar spectral features given their similar phytoplankton composition, will provide small linkage distance and therefore will come to lie closer to each other in the cluster tree. In addition, derivative spectroscopy was applied to hyperspectral data in some cases with the aim of enhancing spectral features of interest in each case and improving performance of our approach.

C. AI techniques

Particle Swarm Optimization

Particle Swarm Optimization is an algorithm that makes the most of the social interaction metaphor [10]. In particular, a multiple PSO (mPSO) is an optimization technique where a collection of particles explore a specific search space combining information of their individual optima and their neighborhood optima. Based on a collaborative intelligence behavior, particles achieve a local optimum in the solution space. mPSO technique iteratively defines several sets of particles forming different neighborhoods. Then, at every step of the algorithm, each particle standing at a certain position moves to a new position following a direction and speed according to its own experience and the experience gained by its neighbors. At the end of this iterative process, every set of particles is able to find a maximum in its neighborhood. Hence, we can obtain good quality results by combining the partial solutions.

In the present study, particles are launched over a 3D space which consists of the attenuation coefficient (K_d) at different depths and spectral wavelengths, corresponding to an specific underwater scenario. mPSO is therefore applied in order to estimate the depth where some phytoplankton thin layers are located, by finding the most significant peaks in the spectral attenuation coefficients along depth.

Case-Based Reasoning

Case-Based Reasoning (CBR) is an AI approach that takes advantage of the experience gained when solving previous

problems in a certain domain by applying such experience to solve the problem at hand [11]. CBR defines past problem solving situations as cases, where each case is made of a (problem, solution) pair. The *problem* describes the situation to be tackled and the *solution* describes what is correct to do in that situation. A collection of cases is called the case base. When a new problem arrives at a CBR system, the *retrieve* step is performed. Its purpose is finding cases with similar problems to the current problem. Then, in the *reuse* step, the solution(s) of retrieved case(s) are used to determine the solution for the current problem.

Every time we are facing a new *problem*, we specify it according to a tuple of properties and we search in the case base for cases with similar tuples using the Euclidean distance. For our study, phytoplankton thin layers were distinguished based on the hyperspectral information by detecting concentration peaks along depth and singular features along spectral wavelength. Thus, the *problem* of a case was defined using a tuple of 3 properties. First property was represented by the first four coefficients from the Discrete Fourier Transform (DFT) applied to each spectrum at each depth [12]. DFT provided us a contour characterization of each spectrum, which can be related to absorption features and pigments of phytoplankton groups present in each scenario. The second property corresponded to the value that K_d takes at a relevant wavelength, which is informative of the concentration of phytoplankton of the detected thin layer. Finally, the third property was assigned by the w parameter (i.e. the Full Width at Half Maximum of a normal distribution), that takes into account the thickness of the thin layer. Next, to create the *solution*, we needed to determine the algae group, its real concentration and its real thickness. The algae group was determined by a majority vote among the retrieved cases. Because conversion from K_d and w to real level of concentration and thickness values depended on the algae group, we used this retrieved algae type information together with a linear regression over estimated values of concentration and thickness versus real values from the previous cases. This combined approach provided us a multiplicative factor, which was used to predict the real values.

III. RESULTS AND DISCUSSION

A. Identification of phytoplankton assemblages by cluster analysis

Hyperspectral data generated from scenario # 1 represents the case of having a stratified water column with two thin layers of phytoplankton at different depths, each dominated by a single algae group: diatoms and cyanobacteria (Fig. 3, top panel). Figure 5 (top panel) displays the dendrogram obtained after applying HCA techniques based on 300 modeled reflectance spectra ($R(\lambda)$) at different depths every 5 cm (Fig. 3, bottom panel). X-axis indicates the linkage distance computed between spectra corresponding to different depths. Results from cluster analysis show a specific partition depending on the spectral variability due to the not homogeneous simulated profile. Moreover, detecting natural groupings in a dendrogram and selecting an optimal number of clusters is performed by

representing a diagram of the evolution of linkage distances along the dendrogram (Fig. 5, bottom panel). It can be noted that in our case an optimal number of 4 clusters is assessed based on the point in which the linkage distance between objects change abruptly, associated with a steep increase of the within cluster variance (see dotted line in Fig. 5, both panels). Thus, all objects (i.e. $R(\lambda)$ spectra) located below this arbitrary linkage distance where the hierarchical tree is cut off, leading to the formation of 4 clusters, are assigned to a single cluster (i.e. a specific color).

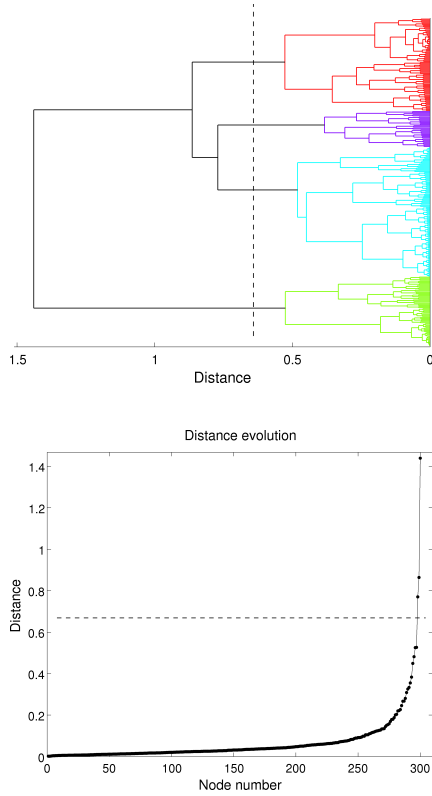


Fig. 5. (top) Hierarchical cluster tree based on the hyperspectral reflectance spectra, along the water column shown in Fig. 3. (bottom) Corresponding diagram of evolution of linkage distances of the cluster analysis. The dot line in both panels indicates the distance associated to the first steep increase of the within cluster variance, which yields an optimal number of clusters of 4 to partition the water column scenario.

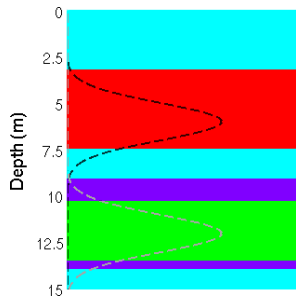


Fig. 6. Segmentation of the water column as a result of the cluster analysis based on the $R_{rs}(\lambda)$ spectra. Superimposed are the chlorophyll profiles of each phytoplankton species.

A detailed segmentation of the water column has been obtained by the cluster analysis based on $R(\lambda)$ spectra along depth. Figure 6 illustrates how each part of the water column scenario has been assigned to a color based on the cluster analysis described above. Both thin layers of cyanobacteria and diatoms, shown in red and green respectively, have been clearly distinguished one from the other and separated from the regions where no phytoplankton was present, shown in cyan. These results confirm the potential of considering observing platforms able to gather reflectance information at a very high spectral and spatial resolution to infer detailed information with regard to fine-scale phytoplankton structure in the water column.

A second data set was generated in order to keep testing performance of HCA techniques. In this case, the purpose was two fold: (1) to check if HCA techniques were suitable to identify different phytoplankton assemblages in open water scenarios, where the level of concentration of the phytoplankton component is very low; (2) to evaluate if such identification was possible globally from a remote-sensing approach, using above surface reflectance spectra ($R_{rs}(\lambda)$). Scenario # 2 consisted of 24 open ocean scenarios, vertically homogeneous and dominated by single phytoplankton groups at different very low concentration levels. $R_{rs}(\lambda)$ spectra showed a great variability in both magnitude and spectral shape, in correspondence with the variable phytoplankton composition and concentration rates (Fig. 4, bottom panel). Figure 7 shows the results of applying HCA analysis to hyperspectral $R_{rs}(\lambda)$ data set.

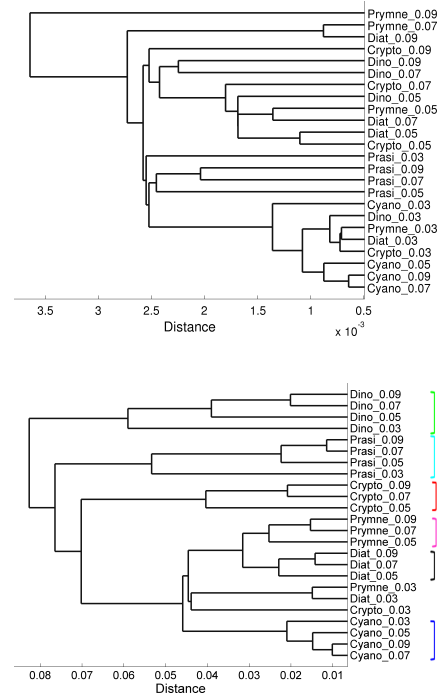


Fig. 7. Results of the cluster analysis applied to: (top) hyperspectral remote-sensing reflectance spectra shown in Fig. 4 and (bottom) its second derivative spectra.

Each classified $R_{rs}(\lambda)$ was identified with a specific label, consisting of the name of the dominating phytoplankton group and the concentration value (e.g. diatoms at a concentration of 0.07 mg/m^3 along the homogeneous water column is indicated with the label *Diat_0.07*). In order to obtain a better performance on the automatic classification of phytoplankton assemblages, computation of the second derivative spectra of each $R_{rs}(\lambda)$ was here necessary. Derivative spectroscopy permitted to enhance subtle fluctuations in the spectra and better separate closely related spectral features [9]. Results indicate that phytoplankton assemblages were mostly identified regardless the concentration rate considered when derivative spectra were considered (Fig. 7, bottom panel). It means that hyperspectral $R_{rs}(\lambda)$ corresponding to the same phytoplankton dominating group were connected by closer dendrites in the dendrogram. The only cases that were not grouped according to the phytoplankton dominance were those corresponding to the lowest concentration rate (i.e. 0.03 mg/m^3). When having such small concentrations, an accurate assessment of most suitable derivative settings or spectral range of analysis should be done to obtain a better classification.

B. Identification of phytoplankton assemblages by AI techniques

Particle Swarm Optimization

mPSO algorithm has been applied to data profiles from scenario # 3 in order to detect the presence of thin layers of phytoplankton in the water column structure. Four groups of ten particles each have been launched for every profile, accessing just 13% of the existing 7.5×10^4 measurements. Figure 8 (top panel) shows the results achieved in an illustrating scenario. This procedure has been executed 20 times obtaining an average failure result of 1.38% in detecting the thin layers. Such error ranges in average 2.8 cm in the space with a standard deviation of 3.0 for the profiles where no noise was applied. For those profiles perturbed by 50% of added noise, we get 1.82% of failures, an error of 4.2 cm and a standard deviation of 10.8.

Case-Based Reasoning

Phytoplankton-type characterization of thin layers defined in scenario # 3 have also been assessed by the CBR module. Figure 8 (bottom panel) depicts a 3D representation of the distance relations for each simulated case when the CBR technique is applied, where colors indicate the dominant phytoplankton group. Its accuracy has been tested by performing a set of experiments by means of a 10-fold cross validation. The results obtained show an accuracy over 97% with respect to the algae group identification when the noise added to data does not exceed 30%. For noise perturbation ranging between 30% and 50%, accuracy measure drops steadily reaching 95% in the worst case. Regarding the prediction of real concentration and thickness of each thin layer according to algae group, we observe an accuracy over 95% whether the dominating algae group when no noise is applied to data.

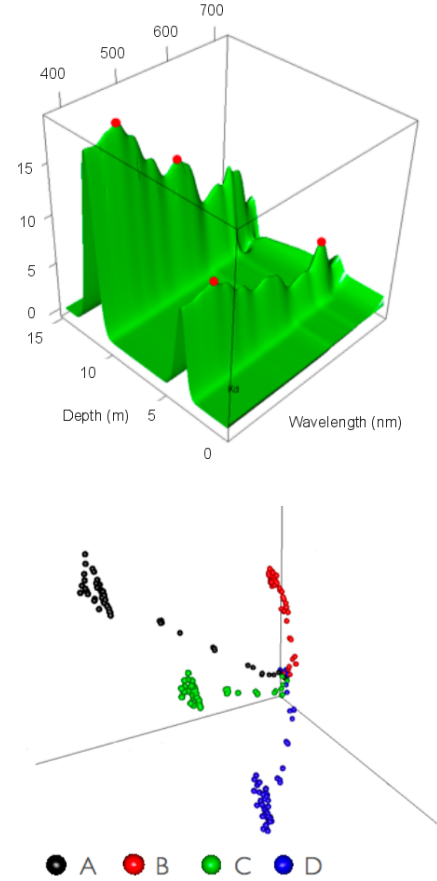


Fig. 8. (top) Detection of optically significant spectral peaks applying mPSO. (bottom) 3D projection of different water column scenarios characterized by four dominating phytoplankton groups at different concentrations, after applying the CBR module.

IV. SUMMARY AND CONCLUSIONS

In accordance with the described results, the main goal of this study is to point out the remarkable suitability of hyperspectral information for mapping phytoplankton assemblages in the ocean. Several processing techniques and Artificial Intelligence algorithms have been satisfactorily applied to oceanographic hyperspectral data sets and have permitted the classification of different environments in terms of phytoplankton biodiversity given different scenarios. In particular, the potential offered by an unsupervised hierarchical cluster analysis technique to discriminate phytoplankton assemblages from hyperspectral measurements given a variety of hypothetical oceanic environments has been shown. Moreover, successful application of two AI techniques has also been demonstrated. A multiple Particle Swarm Optimization algorithm and a Case-Based Reasoning approach have provided a robust identification of thin layers of phytoplankton. mPSO algorithm automatically detected optically significant peaks in the spectral information even when the input signal suffered from high noise levels and CBR permitted to characterize thin layer's dominant algae group as well as its thickness and level of concentration.

We have demonstrated that identification of different phytoplankton assemblages or thin layers when applied to simulated oceanographic hyperspectral data sets is feasible by using our processing methods. They have been tested by modelling different types of underwater optical scenarios including from open water masses (i.e. non-algal-bloom conditions) to stratified scenarios showing thin layers of phytoplankton at the upper part of the ocean water column (i.e. higher levels of biomass). Robustness of our approaches opens the possibility to successfully apply these methods to oceanographic hyperspectral data sets, like those obtained in the future by the above described ANERIS high resolution profiling system.

ACKNOWLEDGMENT

This study was supported by the intramural project ANERIS (PIF08-015) funded by the Spanish National Research Council (CSIC), Ministry of Science and Innovation.

REFERENCES

- [1] T. Dickey, M. Lewis, and G. Chang, "Optical oceanography: recent advances and future directions using global remote sensing and in situ observations," *Reviews of Geophysics*, 44, RG1001, pp. 1-39, 2006.
- [2] A. Nair, S. Sathyendranath, T. Platt, J. Morales, V. Stuart, M. Forget, E. Devred, and H. Bouman, "Remote sensing of phytoplankton functional types," *Remote Sensing of Environment*, 112, pp. 3366-3375, 2008.
- [3] O. Schofield, J. Grzyski, W. P. Bissett, G. J. Kirkpatrick, D. F. Millie, M. Moline, and C. S. Roesler, "Optical monitoring and forecasting systems for harmful algal blooms: Possibility or pipe dream?," *Journal of Phycology*, 35, pp. 1477-1496, 1999.
- [4] M. M. Dekshenieks, P. L. Donaghay, J. M. Sullivan, J. E. B. Rines, T. R. Osborn, and M. S. Twardowski, "Temporal and spatial occurrence of thin phytoplankton layers in relation to physical processes," *Marine Ecology Progress Series*, 223, pp. 61-71, 2001.
- [5] ANERIS intramural project, 2010. Retrieved on 10 March 2010. <http://www.utm.csic.es/aneris>
- [6] T. D. Dickey, and R.R. Bidigare, "Interdisciplinary oceanographic observations: the wave of the future," *Scientia Marina*, 69 (Suppl. 1), pp. 23-42, 2005.
- [7] C. D. Mobley, and L. K. Sundman, Hydrolight-Ecolight 5.0 user's guide, Technical report, Sequoia Scientific, Inc., Bellevue, WA, 2008.
- [8] M. Kim, and W. Philpot, Optical Phytoplankton Simulator, 2006. Retrieved on 12 June 2009. http://ceeserver.cce.cornell.edu/wdp2/OOPS/oops_manual.html
- [9] E. Torrecilla, J. Piera, and M. Vilaseca, Derivative Analysis of oceanographic hyperspectral data, in G. Jedlovec (Ed.), *Advances in Geoscience and Remote Sensing*, Vienna: In-Tech, chapter 29, pp. 597-619, 2009.
- [10] J. Kennedy, and R. C. Eberhart, "Particle swarm optimization," *Proceedings of the IEEE international conference on neural networks IV*, pp. 1942-1948, 1995.
- [11] J. Kolodner, *Case-based Reasoning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1993.
- [12] R. Agrawal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," *Proceedings of Fourth International Conference on Foundations of Data Organization and Algorithms*, 1993. Also, in *Lecture Notes in Computer Science* 730, Springer Verlag, pp. 69-84, 1993.