Statistical analysis of a database of absorption spectra of phytoplankton and pigment concentrations using self-organizing maps

Aymeric Chazottes, Annick Bricaud, Michel Crépon, and Sylvie Thiria

We present a statistical analysis of a large set of absorption spectra of phytoplankton, measured in natural samples collected from ocean water, in conjunction with detailed pigment concentrations. We processed the absorption spectra with a sophisticated neural network method suitable for classifying complex phenomena, the so-called self-organizing maps (SOM) proposed by Kohonen [Kohonen, Self Organizing Maps (Springer-Verlag, 1984)]. The aim was to compress the information embedded in the data set into a reduced number of classes characterizing the data set, which facilitates the analysis. By processing the absorption spectra, we were able to retrieve well-known relationships among pigment concentrations and to display them on maps to facilitate their interpretation. We then showed that the SOM enabled us to extract pertinent information about pigment concentrations normalized to chlorophyll a. We were able to propose new relationships between the fucoxanthin/Tchl-a ratio and the derivative of the absorption spectrum at 510 nm and between the Tchl-b/Tchl-a ratio and the derivative at 640 nm. Finally, we demonstrate the possibility of inverting the absorption spectrum to retrieve the pigment concentrations with better accuracy than a regression analysis using the Tchl-a concentration derived from the absorption at 440 nm. We also discuss the data coding used to build the self-organizing map. This methodology is very general and can be used to analyze a large class of complex data. © 2006 Optical Society of America

1. Introduction

In oceanic case 1 waters, the spectral reflectance of the ocean is largely determined by the absorption properties of phytoplankton cells present in the upper layer. Therefore under certain assumptions, the absorption coefficients of phytoplankton can be derived from ocean color measurements. In turn, the absorption spectra of algal populations are strongly dependent on their pigment composition, so that much effort has been focused on extracting pigment information from algal absorption spectra, or even directly from ocean color spectra. Such methods could provide the possibility of getting taxonomic information on phytoplankton from ocean color measured from satellite-borne sensors, which would have important biogeochemical applications.

The retrieval of pigment concentrations from absorption spectra remains a difficult inverse problem, because the variability of the absorption spectrum, at a given chlorophyll-a concentration, is dependent not only on the concentration of accessory pigments but also on the package effect. This effect accounts for the fact that, because of the packaging of pigments within phytoplankton cells (and inside cells within chloroplasts), the absorption spectrum is flattened with respect to the absorption spectrum of the same pigments in solution. The package effect varies nonlinearly with cell size and with intracellular pigment concentrations. Therefore the relationship between the spectral absorption coefficients of natural algal populations and the corresponding pigment concentrations is complex and not easily modeled. The Laboratoire d'Océanographie de Villefranche (LOV) has gathered a large set of ocean water samples for which the absorption spectra of phytoplankton, associated with detailed pigment concentrations...
measured by high-pressure liquid chromatography (HPLC) have been determined. These water samples were collected during several cruises covering different parts of the ocean in different seasons and therefore present a wide variety of situations. Their properties have been analyzed by Bricaud et al. who showed that the field variations in algal absorption coefficients (at a given chlorophyll-a content) are determined not only by the pigment composition but also by variations in the predominant cell size in the phytoplankton populations. This biological noise is likely to blur the relationships between algal absorption spectra and pigment concentrations. Similar effects can also result from photoacclimation of populations to various light conditions (various incident irradiances, or a decrease of irradiance with depth).

The objective of this paper was to extract pertinent objective information embedded in the absorption spectra of this data set from their statistical properties only, without any other a priori knowledge. Owing to the complexity of this information we propose a new and efficient way of visualizing it and consequently of facilitating its interpretation. We used a method based on neural networks, the so-called self-organizing map (SOM). This method can be considered an automatic analysis of the statistical properties of a data set. It allows an objective synthesis of the information in this data set. We propose to identify some pertinent clusters extracted from the full absorption spectra data set by providing a reduced set of characteristic spectra (the so-called prototype spectra) associated with the clusters. Each prototype spectrum (or class) provides a partition (or classification) of the data set that can be further associated with biogeophysical parameters of the water samples. Using this classification we were able to invert the absorption spectrum of phytoplankton to retrieve information concerning their pigment composition.

In Section 2 we describe the data set and the coding. In Section 3 we briefly present the SOM method and the maps resulting from the learning of the data set. Moreover, we show that the complexity of the data set is well represented by the maps. In Section 4 we investigate the possibility of retrieving some pigment concentrations by inverting the absorption spectral values with the help of the self-organizing map. In Section 5 we present a discussion and our conclusions.

2. Observations and Data Description

A. Observations

Samples were collected during ten cruises, in various seasons and various areas of the world’s oceans, between 1990 and 2002. The location of each cruise and the number of samples are displayed in Table 1. All the data considered in this study corresponded only to oceanic case 1 waters, i.e., waters for which the optical properties are dependent on matter of biological origin only and not on terrigenous (particulate or dissolved) substances.

The methods employed for particulate and algal absorption measurements are described in Refs. 8 and 10. In summary, for each sample particulate matter was collected on a 25 mm Whatman glass-fiber filter, and particulate absorption coefficients were measured stepwise (in 2 nm steps) by spectrophotometry, from 400 to 700 nm. Spectra were mea-

<table>
<thead>
<tr>
<th>Cruise Number</th>
<th>Cruises</th>
<th>Location</th>
<th>Usual Trophic State</th>
<th>Date</th>
<th>Total Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TOMOFRONT</td>
<td>Northwestern Mediterranean</td>
<td>Mesotrophic</td>
<td>April 1990</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>EUMELI3</td>
<td>Tropical North Atlantic</td>
<td>Oligotrophic,</td>
<td>October 1991</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mesotrophic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>FLUPAC</td>
<td>Equatorial and Subequatorial</td>
<td>Oligotrophic</td>
<td>September–October</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pacific</td>
<td></td>
<td>1994</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>OLIPAC</td>
<td>Equatorial and Subequatorial</td>
<td>Oligotrophic</td>
<td>November 1994</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pacific</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MINOS</td>
<td>Eastern and Western Mediterranean</td>
<td>Oligotrophic</td>
<td>May 1996</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ALMOFRONT2</td>
<td>Alboran Sea</td>
<td>Mesotrophic</td>
<td>December 1997 and</td>
<td>477</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>January 1998</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>PROSOPE</td>
<td>Eastern and Western Mediterranean</td>
<td>Oligotrophic,</td>
<td>September–October</td>
<td>606</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and Morocco Upwelling Area</td>
<td>Eutrophic</td>
<td>1999</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>POMME</td>
<td>North Atlantic</td>
<td>Oligotrophic,</td>
<td>February–May and</td>
<td>525 + 1571</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mesotrophic</td>
<td>August–October</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>BENCAL</td>
<td>Benguela Upwelling Area</td>
<td>Mesotrophic,</td>
<td>October 2002</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Eutrophic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
measured directly on the wet filter (by reference to a blank filter), by using the quantitative glass fiber filter technique (QFT), except for the FLUPAC cruise, for which the glass slide technique (Allali et al.\textsuperscript{10}) was used. With this latter technique, absorption measurements are not affected by the path-length amplification effect (\(\beta\) effect). When the QFT was used, all the spectra were corrected for the \(\beta\) effect using the algorithms given by Allali et al.\textsuperscript{11} for samples collected in oligotrophic waters, and by Bricaud and Stramski\textsuperscript{12} for other samples. Finally, the respective contributions of phytoplankton and nonalgal particulate matter to total particulate absorption were determined either experimentally (pigment extraction by methanol; Kishino et al.\textsuperscript{13}) or by numerical deconvolution (Bricaud and Stramski\textsuperscript{12}).

We applied a triangular moving window of size 3 to filter the noisy part of the spectra. We then sampled each filtered spectrum every 10 nm. Each spectrum was therefore a 31-dimension vector. The phytoplankton absorption spectral coefficients are represented by the symbol \(a(\lambda)\), where \(\lambda\) stands for the wavelength in nanometers.

Pigment concentrations were measured by high-pressure liquid chromatography (HPLC) using the procedure described by Vidussi et al.\textsuperscript{14} Up to 20 pigments were identified for each sample. For the SOM analysis, all the pigments were grouped into five main categories according to their similarity of spectral absorption characteristics: (1) chlorophyll-\(a\), divinyl–chlorophyll-\(a\), chlorophyllid-\(a\), and pheopigments (pheophytin and pheophorbide) (the sum is noted as Tchl-\(a\)); (2) chlorophyll-\(b\) and divinyl–chlorophyll-\(b\) (noted as Tchl-\(b\)); (3) chlorophyll-\(c\), chlorophyll-\(c\), and chlorophyll-\(c\) (noted as Tchl-\(c\)); (4) photosynthetic carotenoids (noted as TPSC), i.e., fucoxanthin, peridinin, 19'-HF, and 19'-BF; (5) nonphotosynthetic carotenoids (noted as TPPC), i.e., zeaxanthin, diadinoxanthin, alloxanthin, and \(\beta\)-carotene. Two photosynthetic carotenoids, lutein and ocarotene, were also included in the latter category, because their absorption spectra are similar to those of zeaxanthin and \(\beta\)-carotene, respectively.

For each station, the first optical depth was computed as \(z_{\text{opt}}/4.6\), where \(z_{\text{opt}}\) the depth of the euphotic zone, is the depth at which the photosynthetically available radiation is reduced to 1% of its value just below the sea surface. The depth of the euphotic zone was either determined in the field or computed from the chlorophyll profile according to Morel and Maritorena.\textsuperscript{15}

### B. Database

The objective of the research discussed in this paper was to analyze the absorption spectral values \(a(\lambda)\) of algal particulate matter and to try to associate them with concomitant biogeochemical and physical parameters, such as pigment concentrations or environmental factors. In this study we split the absorption spectra data set into two subsets: one for learning and the other for validation of the method. One part of the data set, corresponding to the three POMME cruises, was much larger than the remaining part containing the spectra of all the other cruises (see Table 1). In addition, waters sampled during these cruises were found on many occasions to be bio-optically different from other waters with similar chlorophyll-\(a\) concentrations.\textsuperscript{8} To avoid a possible bias, only 525 randomly selected samples from the POMME cruises were retained in the learning data set. The remaining data from these three cruises (1571 samples) were kept apart and used for the validation of the method.

The learning data set, which was used to estimate the parameters of the statistical model, comprised 2163 samples. Hereafter we refer to \(D\) when dealing with the whole data set. \(\mathcal{L}\) and \(\mathcal{V}\) will stand for the learning set and the validation set, respectively.

### C. Coding

Coding the data is the first step and plays an important role in data analysis. As topological maps process multidimensional data, it becomes possible to describe each observation by a large set of characteristics, each one introducing a particular item of knowledge about the phenomenon under study.

As shown by many studies dealing with optical properties of ocean waters, absorption is primarily dependent on chlorophyll-\(a\) concentration. Once this first-order information has been extracted, there is still more complex and more subtle information to be extracted from the data set. To investigate this second-order information it is useful to remove the first-order information. This is usually done by dividing each absorption datum by the corresponding chlorophyll-\(a\) concentration. This method is widely used by the concerned marine scientific community and has recently been used by Bricaud et al.\textsuperscript{8} to analyze the present data set. It allowed these authors to extract pertinent information from absorption spectra and pigment concentrations. Owing to the theoretical characteristics of the topological map and its highly discriminative power,\textsuperscript{16} we decided not to apply the chlorophyll-\(a\) normalization to the observations. In fact, more satisfactory results were obtained using nonnormalized data than normalized data. Normalization by Tchl-\(a\) was introduced after the learning phase for analyzing the concentration of the different pigments.

As shown in Fig. 1, which displays the weight-specific absorption spectra of the different pigments with respect to wavelength, we note that the absorption spectra of the pigments have their maxima at different wavelengths and have different shapes. Consequently, the shape of the absorption spectrum of a water sample is related to the different pigment concentrations. To help the analysis and the retrieval of the pigment concentrations from the spectral information, we decided to use the first derivative of the spectra with respect to the wavelength, which is a proxy of the spectrum shape in statistical processing. This coding was inspired by previous work by Bidigare et al.\textsuperscript{17} and Faust and Norris.\textsuperscript{18,19} Owing to the large variability (several decades) of the absorption spectral values \(a(\lambda)\), we decided subsequently to
use log10[\(a(\lambda)\)] values rather than \(a(\lambda)\) values, and for the same reasons to use the log-transformed values of the pigment concentration as well.

Each water sample spectrum is thus characterized by a 31-component vector, whose first 30 components are the spectrum derivatives computed as the difference between the log10[\(a(\lambda)\)] values for two consecutive wavelengths (i.e., log10[\(a(\lambda_{400+10i})\) − log10[\(a(\lambda_{400+10(i+1)})\)], where \(i = 1, \ldots, 30\), and the last component is the maximum value of the absorption. This latter value is a pertinent piece of information about chlorophyll-\(a\) content and more generally on the overall pigment content. As absorption coefficients were log-transformed and a few spectra presented null or slightly negative values in the green part of the spectrum, we added a small offset of 10^{-4} \text{ m}^{-1} to these spectra to have only positive absorption coefficients. Finally, since all the pigment concentrations were also log-transformed, we set the zero pigment concentrations arbitrarily to a value of 10^{-4} \text{ mg m}^{-3}.

3. Classification Method: Self-Organizing Map Algorithm

A. Introduction

We used an unsupervised classification method to extract pertinent information from the statistical structure of the data set without any \textit{a priori} information. This method has been extensively described by Niang \textit{et al.}^{20} We now give an outline of it.

The method consists of a statistical model, the so-called SOM, which was described by Kohonen\(^9\) for visualizing and clustering high-dimensional data sets. The classification aims at summarizing the information contained in the learning data set \(L\) by producing a set of reference vectors \(\mathbf{rv}\) (synthetic spectra) that are representative of the data. The set of \(\mathbf{rv}\)'s represents the data set by compressing the information contained in it. Each \(\mathbf{rv}\) matches a certain number of data of \(L\) according to a quadratic distance \(d^2\) and defines a class, the matched data being the elements of the class. Each neuron of the SOM is associated with a particular reference vector \(\mathbf{rv}\). The different neurons of the topological map \(C\) are connected and determine a topological (neighborhood) relationship among the different neurons (Fig. 2): close neurons on the map represent similar subsets of data (classes presenting similarities).

In this study we produced a 2D \((n \times p)\) topological map with a large number of neurons \((10 \times 10\) in this study), providing a highly discriminating representation of the observations. The 2163 spectra of \(L\) are thus clustered into 100 classes of spectra, each one represented by a typical spectrum characterizing the associated reference vector. We now present a first analysis of the SOM using this method and show its capacity to easily synthesize well-known results.

B. Validation of the Self-Organizing Map

We now analyze the structure of the SOM at the end of the learning phase. Each neuron of the map is associated with a class. It is represented by a prototype spectrum (the reference vector \(\mathbf{rv}\)), which is a statistical mean of all the spectra captured by the neuron. Thus, in the following, we also associate the prototype spectrum with a pigment concentration that is defined as the mean of the pigment concentrations corresponding to all the spectra selected by the neuron.
1. Relationship between the Amplitude of the Absorption Spectrum and the Pigment Concentrations

Figure 3 displays the mean log-transformed absorption spectral data associated with each \(rv\) of the SOM and allows us to visualize the topological order provided by the Kohonen map. We note that the SOM is well organized, the adjacent neurons having very similar spectra thanks to the neighborhood procedure mentioned previously. The learning phase was successful, as nearly every neuron captured a significant number of observed spectra, as shown in Fig. 3, in which the number of spectra captured by each neuron of the SOM is displayed. Each reference vector represents 22 spectra on average (60 maximum and 2 minimum) showing that the observations of \(L\) are well distributed over all the reference vectors.

The set of 10 \(\times\) 10 neurons provided by the SOM is a reduced data set that represents a rational decomposition of \(L\); it is a way to compress the information of this data set by extracting the most pertinent spectra. It is expected to be easier to extract significant objective characteristics from this reduced data set than from the whole data set. It also enables us to easily visualize the main trends in the data set learned by the map.

First we note that the mean amplitude of the spectra decreases from the top-left corner (where it is at its maximum) down to the bottom-right corner (where it is at its minimum). This variation can be schematized by the first diagonal of the SOM from the top-left corner to the bottom-right corner. Therefore the classification can order the spectra according to their amplitude. In Subsection 3.B.2 we shall see that more complex and more subtle information is embedded in this map, but various shapes of spectra are already apparent.

Figure 4 shows the spectra associated with some reference vectors. Each spectrum is displayed with its associated error bars (\(\pm 2\) standard deviations). We also added the different spectra of \(L\) captured by each neuron. All the spectra captured by a neuron may differ in amplitude but have generally similar shapes. The shape and the amplitude vary drastically among the reference vectors. The SOM is thus ordered with respect to the amplitude and the shape of the spectra.

Figure 5 shows the log-transformed concentrations of the five different groups of pigments associated with the neurons of the SOM. The concentrations of Tchl-\(\alpha\), TPSC, Tchl-\(\gamma\), and TPPC decrease from the top-left corner to the bottom-right corner. These variations in pigment concentrations are in agreement with the variations in the amplitude of the spectra [Fig. 5(f)]. The concentrations of Tchl-\(\alpha\) and of TPSC appear to be closely correlated with the spectrum mean amplitude, which suggests a strong link between the spectrum amplitude and the concentra-
tions of these two pigments. This link also exists for Tchl-α and TPPC, but it is less pronounced. The concentration gradient can be schematized by the first diagonal of the SOM (from the top left to the bottom right) showing the decrease in pigment richness of the water, which is first-order information embedded in the map and consistent with previous studies,\textsuperscript{21} which showed that the chlorophyll content is the first determinant of the amplitude of absorption spectra. Moreover, the Tchl-α concentration on the SOM also varies slightly on a line perpendicular to the Tchl-α concentration gradient, which, we believe, is the...
second-order information embedded in the map, the first-order information being along the chlorophyll-α gradient. The behavior of the Tchl-b concentration is different. It is roughly maximum along a line parallel to the second diagonal (bottom left to top right) and minimum on each side of it, which is coherent with the second-order information perpendicular to the first diagonal representing the first-order information.

From this analysis we note that the map divides the five groups of pigments into three categories according to the patterns displayed by the SOM. The first group concerns Tchl-α and the TPSC; the second group, Tchl-c and the TPPC; and the third group, Tchl-b. The first two groups have similar properties: their concentrations are well correlated with the amplitude of the spectrum. The map, which was learned only with the spectrum, captured not only the first-order information embedded in the spectrum, but also more subtle information because of the pigment composition (see Subsection 3.B.2).

The region above the first diagonal (i.e., the diagonal joining the top-left corner to the bottom-right corner) corresponds mainly to water samples taken below the first optical depth. The zone below this diagonal contains most of the first-optical-depth samples mixed with some deeper ones, as we have noted previously, whose spectra are similar in shape and amplitude. The division above and below the first diagonal divides deep and shallow samples.

2. Variations in the Accessory-Pigment/Tchl-α Ratios with Respect to Spectral Absorption
At the first-order information level, the concentrations of accessory pigments are related to that of Tchl-α (Fig. 5). To remove the strong influence of the Tchl-α concentration in the organization of the SOM, we studied the variations in the concentrations of accessory pigments normalized to Tchl-α. Figure 6 shows the different pigments-to-Tchl-α ratios for the neurons of the SOM. These ratios display well-identified patterns on the SOM, which are different from those in Fig. 5. This shows the ability of the SOM method to capture second-order information from the data set. In particular, we noted the variations in a direction orthogonal to the Tchl-α gradient. Indeed, the maps are no longer symmetrical with respect to the first diagonal. Strong differences are evident between the surface neurons and the deeper ones [compare Fig. 6(a) with Fig. 6(b), which displays only neurons above the first optical depth]. Furthermore, there are clearly identifiable patterns specific to each pigment ratio in Fig. 6. Figures 6(a) and 6(e), for instance, show that neurons corresponding to high Tchl-b/Tchl-α and TPPC/Tchl-α ratios are grouped above and below the gray line (separating the surface

Fig. 6. (Color online) (a)–(e) Ratios of the concentrations of the four pigment classes to the Tchl-α concentration for all the neurons of the SOM. The color scale is different for each pigment class. The ellipses enclose clear patterns for each of the pigment ratios. The neurons above the gray line roughly correspond to neurons associated with deep samples. The neurons below the gray line roughly correspond to neurons associated with samples in the “first optical depth.” (b) Tchl-b/Tchl-α ratios for neurons of the first optical depth.
and deep neurons), respectively. This demonstrates, consistently with previous field studies,\textsuperscript{11,22} that high Tchl-$b$/Tchl-$a$ ratios are found preferentially in deep water, whereas high TPPC/Tchl-$a$ ratios are found near the sea surface. Figures 6(c) and 6(d) have some symmetrical patterns with respect to the first diagonal, showing that high Tchl-$c$/Tchl-$a$ and TPSI/Tchl-$a$ ratios may be found both near the surface and in deeper water.

The above analysis showed the ability of the SOM method to retrieve well-known features. The weight of the map learned with phytoplankton absorption spectra clearly contains first- and second-order information consistent with the pigment composition variations shown in previous studies.\textsuperscript{8} It must be emphasized here that this information is derived from the shape and amplitude of the absorption spectra only, without any ancillary data. Nevertheless it seems difficult to extract more information on the structure of the data set from the SOMs as they are presented in Figs. 5 and 6. We suggest refining the analysis by looking at potential links between new variables. In the following section we investigate links between the derivatives of the absorption spectrum at specific wavelengths and some individual pigment concentrations.

4. Exploitation of the Self-Organizing Map

The above results have shown that the SOM method applied to the absorption spectrum derivatives provided well-organized topological maps with respect to the concentrations of the various pigments and to the pigment ratios. They allowed us to show actual relationships between pigment concentrations and the amplitude and shape of absorption spectra, some of which are already known from previous field studies. We applied the SOM method to reveal new relationships between some pigment ratios in a water sample and the derivative of the absorption spectrum of the phytoplankton in this sample.

We examined the derivatives of the \textbf{rv} spectra with respect to the wavelength and drew 30 maps (Fig. 7) corresponding to the 30 derivative values of the \textbf{rv} spectra with respect to the wavelength, each map representing the derivative values at a specific wavelength of the 100 reference spectra of the SOM. The derivatives were computed for the $\log_{10}$ phytoplankton absorption spectra as

$$
\log_{10}(a_j) - \log_{10}(a_{j-10}) = \log_{10}(a_j/a_{j-10}),
$$

where $j = (400 + 10i)$ and $i = 1, 2, 3, \ldots, 30$. \textsuperscript{(1)}
Some of these maps present well-identified patterns at specific wavelengths, which can be related to some of the pigment–ratio maps, suggesting a linear relationship between the pigment ratio and the derivative values of the spectrum at given wavelengths.

A. Fucoxanthin/Tchl-a Concentration Ratio

The fucoxanthin/Tchl-a concentration ratio provides information about the phytoplankton cell size structure, since fucoxanthin is the main carotenoid of diatoms, which are usually large cells. It must be emphasized, however, that ambiguities in this approach may occur, as fucoxanthin may also be found in some prymnesiophytes, chrysophytes, and pelagophytes. The SOM of the spectral data (provided by the SOM map) are represented by crosses (+). The regression line linking the two quantities suggests that the fucoxanthin/Tchl-a ratio is actually correlated with this derivative. We computed both a linear and a log-linear regression from the original data set (2163 samples). The most accurate fit was provided by the log-linear relationship

$$\text{fucoxanthin/Tchl-a} = 1.311 \left(\frac{a_{510}}{a_{500}}\right)^{0.19}$$

\hspace{1cm} (2)

$$(R^2 = 0.59, s = 0.284; \text{see the definitions of these coefficients in Appendix A.})$$

The log-linear regression is illustrated in Fig. 9. The dots display the observations for the individual samples, and the crosses, the different reference vectors (which are well spread among the samples, indicating that the reference vectors summarize the observations well).

The above-mentioned regression suggests that it is possible to get a rough estimate of the fucoxanthin/Tchl-a ratio from the derivative of the logarithm of the absorption spectrum at 510 nm. We nevertheless note that the dispersion is quite high for the very small values of the fucoxanthin/Tchl-a ratio. This can be explained by the fact that, for these small ratio values, the relationship may depend on other pigments. The topological map is able to take this dependence into account. The above-mentioned relationship could hardly have been derived from heuristic arguments based on the specific pigment absorption spectra displayed in Fig. 1 in which the fucoxanthin absorption is maximum at $\lambda = 490$ nm and the fucoxanthin absorption derivatives are maximum at $\lambda = 535$ nm and $\lambda = 410$ nm. Regressions with respect to the absorption at $\lambda = 490$ nm are not significant, probably because the interference of other pigments is much stronger at this wavelength (see Fig. 1). The SOM allows us to determine visually the most significant spectrum derivative according to the constraints due to the influence of the other pigments. This visual determination is more objectively confirmed by computing the correlation on the map. We note that 510 nm is just in the middle of the interval 488–532 nm, which corresponds to the wavelength interval for which Eisner et al. computed the slope of the absorption spectra of particles and showed that it was linearly related to the TPPC/TPSC ratio.

B. TPPC/TPSC Concentration Ratio

The TPPC/TPSC ratio has been studied by Eisner et al. in surface water (0–20 m depth range) samples taken off the Oregon coast in June 1998. The authors showed a strong linear relationship between this ratio and the normalized absorption spectrum slope of approximately $\lambda = 510$ nm, computed as $(a_{488} - a_{532})/\left[a_{476}(488 - 532)\right]$. By computing a proxy of this slope for the rv spectra of the SOM and the TPPC/TPSC ratio associated with the rv, we found a stronger correlation between this slope and the log-transformed TPPC/TPSC ratio than that given by Eisner et al. This correlation is illustrated in Fig. 10. It is improved if we consider samples only from the first optical depth (between 2 and 25 m). In that case we found a correlation coefficient of 0.71 (instead of 0.58), allowing us to estimate an empirical relationship between these two quantities. In the physical space, the logarithm relationship becomes

$$\left(\frac{a_{490} - a_{530}}{a_{480}(490 - 530)}\right) = -0.0416 - 0.0262 \times \log_{10}(\text{TPPC/TPSC}).$$

\hspace{1cm} (3)
The correlation coefficient between the Tchl-b/concentration ratio associated with the reference vectors and the Tchl-c/concentration ratio is 0.82. Owing to the number of data used in absorption derivative at ratio associated with the reference vectors and the Tchl-b/concentration ratio, a strong link between the two quantities. This relationship is illustrated in Fig. 11, which shows the scatterplot of the two quantities. This result may be considered as an extension of the work of Eisner et al., who processed a small amount of data (32) whose characteristics were similar. Moreover, the Eisner et al. TPSC/TPC ratio ranged between 0 and 1, whereas in our data set it ranges between 0 and 10. Even if the determination coefficient obtained in our study is smaller than that obtained by Eisner et al., (0.93), the confidence level is similar, owing to the fact that the sample size in the present data set is much higher. Note also that the similarity between the variations in the fucoxanthin/Tchl-a and the TPSC/TPC ratios with respect to the absorption spectrum slope is due to the fact that fucoxanthin is usually the predominant pigment in the TPSC.

C. Tchl-b/Tchl-a Concentration Ratio

We now compare the Tchl-b/Tchl-a ratio to the term log10(a650/a640) (Fig. 12). There is a strong similarity between the two maps, suggesting the existence of a significant relationship between the two quantities. The correlation coefficient between the Tchl-b/Tchl-a ratio associated with the reference vectors and the absorption derivative at λ = 650 nm of the reference vectors is 0.82. Owing to the number of data used in the computation of the correlations (100), these correlations are highly significant, still suggesting a potential linear relationship between these variables. This led us to determine, as before, a log-linear relationship of the following form in physical space:

\[
\text{Tchl-b/Tchl-a} = 0.019 \left(\frac{a_{650}}{a_{640}}\right)^{13.33}
\]

with \(R^2 = 0.28, s = 0.84\). (4)

This linear regression poorly models the relationship between the Tchl-b/Tchl-a concentration ratio and the derivative of the logarithm of the absorption spectrum at 650 nm [Fig. 13(a)]. This is confirmed by Fig. 13(a), which shows the regression line, the observations, with dots, and the quantities associated with the reference vectors with crosses.

Three reference vectors (crosses) corresponding to a Tchl-b/Tchl-a ratio smaller than 10−2 are far from the regression line. The physical characteristics of these three reference vectors are given in Table 2. The Tchl-b concentration was 10−4 mg m−3, which corresponds to values of Tchl-b equal to zero arbitrarily set to 10−4 mg m−3 to avoid zero values, as mentioned in Section 2. The Tchl-c concentrations are also very low. All the spectra captured by these three neurons were sampled just below the first optical depth during the PROSOPE cruise. Figure 5 shows that these three neurons are close together and form an isolated pattern on the SOM [lighter neurons at the top of Fig. 5(c)] far away from the large pattern corresponding to low concentration values of Tchl-b and Tchl-c near the bottom-right corner, which is an isolated pattern on the SOM [lighter neurons at the top of Fig. 5(c)].

### Table 2. Characteristics of Three Outliers Neurons

<table>
<thead>
<tr>
<th>Neuron Number</th>
<th>Tchl-a</th>
<th>Tchl-b</th>
<th>Tchl-c</th>
<th>TPSC</th>
<th>TPPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>0.3231</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.2210</td>
<td>0.0208</td>
</tr>
<tr>
<td>32</td>
<td>0.8815</td>
<td>0.0001</td>
<td>0.0106</td>
<td>0.5958</td>
<td>0.0336</td>
</tr>
<tr>
<td>41</td>
<td>0.3546</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.2528</td>
<td>0.0343</td>
</tr>
</tbody>
</table>

*Pigment concentrations associated with the three reference vectors considered as potential outliers in the Tchl-b/Tchl-a ratio analysis. The Tchl-b and two Tchl-c concentrations are nearly equal to zero (10−4 mg m−3).*
associated with samples of low Tchl-a content taken at depth and coming from the same cruise. Besides the zero Tchl-b concentration, the slope of the absorption spectrum is dependent on other pigment concentrations. It implies that the noise caused by other pigments plays a strong role in determining the value of the derivative that is no longer related to the Tchl/H2O ratio. These data can thus be considered as outliers with respect to the regression.

We removed the 19 samples associated with these reference vectors and determined a new regression line similar to the one above:

$$\text{Tchl-b/Tchl-a} = 0.090 \left(\frac{a_{650}}{a_{640}}\right)^{0.838}, \quad (5)$$

with an improved goodness of fit coefficient ($R^2 = 0.531$) and a smaller $s$ ($s = 0.246$).

The skill of the above regression is now improved and is presented in Fig. 13(b). Moreover, the relationships are similar when limiting the analysis to the first-optical-depth data. In these three examples, we show how we can use the SOM algorithm to establish new links between the features of the absorption spectra of phytoplankton and the pigment ratios and how to improve them by analyzing the large clustering of the data provided by the SOM. It is well known that because of the package effect the relationships linking the logarithm of the derivative of the spectrum and pigment concentrations are not strictly linear. This is confirmed by the values of $R^2$ and $s$ for the various regression analyses. Progress in understanding and establishing new links among these above-mentioned variables implies the use of methods more sophisticated than linear regression. In Subsection 4.D we compare the performance of the SOM method with those of linear regressions.

### D. Retrieval of the Pigment Composition from the Spectral Characteristics of Absorption

In Subsection 4.C we showed that visualizing the reference vectors of the SOM allowed us to establish relationships between the optical and the biological properties. Figures 8, 10, and 12 suggest that the retrieval of pigment concentration information from optical properties is a possible task but more complex than the simple regressions we tested, if we need to obtain a high accuracy in the inverted data. Nonlinearity and the combination of several variables may play an important role. The SOM method offers us a way to overcome these difficulties by using its classification capabilities that can model multivariate and nonlinear processes.

The inversion procedure is as follows. The inversion of the phytoplankton absorption spectrum of a given sample is processed by the SOM algorithm and is associated with a specific reference vector. Then the pigment concentrations of this reference vector are attributed to the sample. The proposed method is similar to the analogous methodology employed in meteorology for weather forecasting. We tested this procedure on the learning and validation sets defined in Section 2. The results of the inversion are

![Fig. 13. Tchl-b/Tchl-a ratio as a function of the slope of the log10 of the absorption spectrum at 510 nm. The water sample data are represented by dots (·). The reference vector data (provided by the SOM map) are represented by crosses (+). Log-linear regression linking the two values is shown by a solid line (—). $s$ and $R^2$, two estimators of the quality of the regression are displayed. In (a) all the data are processed ($s = 0.842, R^2 = 0.277$). In (b) the outliers are removed, which greatly improves the quality of the regression ($s = 0.246, R^2 = 0.531$).](image-url)
presented in Table 3 for the learning and the validation data sets. The error performances are given in terms of relative root-mean-square error (RRMSE) for each pigment. We have also computed the vector RRMSE (VRRMSE), which is a multivariate estimator taking into account the estimation of the error of the pigment concentration vector globally and not separately for each vector. The mathematical expressions of the RRMSE and the VRRMSE are given in Appendix B.

To assess the efficiency of the SOM approach, we also compared the SOM performances to a rough inversion given by a classical log-linear regression between the different pigment concentrations and the Tchl-\(\alpha\) content. The Tchl-\(\alpha\) content was estimated from the absorption coefficient of phytoplankton at 440 nm using a log-linear regression similar to that proposed by Bricaud et al.\(^8\) As suggested by the results in Subsection 3.B, the other pigment concentrations were estimated by a linear regression between the estimated Tchl-\(\alpha\) concentration and that of each pigment. These empirical linear functions were calibrated on the learning data set. These regressions have been computed only to provide threshold values for the SOM retrieval to assess the confidence in the retrieval procedure.

The SOM retrieval gives a better statistical estimator for each pigment than those computed from the linear regressions. The improvement is meaningful for the VRRMSE, which is a multivariate estimator taking into account the pigment concentration estimation globally and not separately for each vector, meaning that a link exists between the retrieval of the different pigments. This confirms the well-known fact that the relationship between pigments and absorption spectra is better modeled by a multivariate nonlinear procedure, such as SOM, than by a univariate linear procedure. Improving the pigment retrieval should imply the use of methods other than linear regression, which are able to take into account the nonlinearity and the multiple dimensions of the problem. A major advantage of the SOM approach is that it allows us to retrieve the pigments globally without any \textit{a priori} hypothesis with the same algorithm. One could imagine more efficient algorithms, each one being specific to a particular pigment.

5. Discussion and Conclusions

By processing a large data set of phytoplanktonic absorption spectra with the SOM algorithm, we were able to show pertinent information between the pigment concentrations of the water samples and the absorption spectrum values. When we started this work, an important question was raised on the coding of the absorption data set. To relate pigment composition variations with phytoplankton species, analysis of the absorption spectrum is usually done by normalizing it to the Tchl-\(\alpha\) concentration or to the absorption at \(\lambda = 440\) nm of the samples to remove the effect of Tchl-\(\alpha\), which is important and can hide that of the other pigments. We used nonnormalized spectra. To find the optimum coding for processing the SOM algorithm, we tested several SOMs learned with different coding for the retrieval of the pigment concentrations as described in Subsection 4.D. The results of this sensitivity

<table>
<thead>
<tr>
<th>Retrieving Algorithm</th>
<th>Tchl-(\alpha)</th>
<th>Tchl-b</th>
<th>Tchl-c</th>
<th>TPSC</th>
<th>TPPC</th>
<th>VRRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learn</td>
<td>36.7</td>
<td>70.5</td>
<td>73.3</td>
<td>128.9</td>
<td>57.8</td>
<td>27</td>
</tr>
<tr>
<td>Validation</td>
<td>38.8</td>
<td>70.9</td>
<td>43.2</td>
<td>74.8</td>
<td>84.6</td>
<td>24</td>
</tr>
<tr>
<td>Least-squares fit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learn</td>
<td>41.0</td>
<td>90.3</td>
<td>77.3</td>
<td>124.1</td>
<td>76.0</td>
<td>45</td>
</tr>
<tr>
<td>Validation</td>
<td>49.5</td>
<td>74.6</td>
<td>58.5</td>
<td>103.6</td>
<td>95.4</td>
<td>36</td>
</tr>
</tbody>
</table>

*RRMSE and VRRMSE performances (see Appendix B for definitions) expressed as a percentage of the SOM and of the least-squares fit, for the learning (\(\approx\)) and validation (\(\approx\)) data sets. The best skill values are boldface.

Table 3. Error Performances on Pigment Retrieving

<table>
<thead>
<tr>
<th>Coding of Spectra Used to Learn the Topological Map</th>
<th>Coding Number</th>
<th>Pigment Concentration Error Performance RMSE(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectrum</td>
<td>1</td>
<td>(0.726) (0.042) (0.123) (0.544) (0.1002)</td>
</tr>
<tr>
<td>(\log_{10}) (Absorption Spectrum)</td>
<td>2</td>
<td>(0.849) (0.042) (0.139) (0.611) (0.1013)</td>
</tr>
<tr>
<td>Spectrum Divided by the Chlorophyll-(\alpha) Content</td>
<td>3</td>
<td>(1.132) (0.054) (0.178) (0.764) (0.1608)</td>
</tr>
<tr>
<td>Derivatives of (\log_{10}) (Absorption Spectrum) and (\log_{10}) (Maximum Amplitude)</td>
<td>4</td>
<td>(0.728) (0.037) (0.123) (0.524) (0.0923)</td>
</tr>
</tbody>
</table>

*RMSE performances (see Appendix B for the definition) for the pigment retrieval using different SOMs educated with different codings of the spectrum. The error performances are computed for the different pigment concentrations to determine which coding best represents the pigment composition information embedded in the spectrum. The best performance for each pigment concentration is shown in boldface. The result of this sensitivity study shows that coding 4 gives the best performances.
study are presented in Table 4. We computed error performances when the coding is the raw spectrum, the log10 of the absorption spectrum, the spectrum divided by the chlorophyll-a concentration and, finally, the log10 of the derivatives of the absorption spectra plus the log10 of the maximum amplitude of the spectrum. The best performance for the pigment concentration retrieval was given for the last coding (coding number 4 in Table 4), which is the coding we used in the present study. We note that this coding is a normalized quantity since it can be written as \( \log_{10}(a(\lambda_i)/a(\lambda_{i+1})) \).

The SOM algorithm decomposed the set of absorption spectra into a reduced number of prototype reference vectors (rvs) by considering the specific statistical characteristics of the rvs for the reduced data set derived from the whole data set. It is easier to extract significant objective characteristics from the reduced data set rvs than from the whole data set.

The data processing presented in this paper was carried out by using the information embedded only in the absorption spectra. The physical and biogeochemical parameters helped us to interpret the classification of the spectra but were not used as inputs to the classification. First we validated the SOM algorithm by showing that it was able to retrieve well-established relationships corresponding to first-order variations in the absorption spectrum, i.e., covarying with the chlorophyll-a concentration. It also allowed us to expose more complex second-order patterns corresponding to variations in the accessory-pigment/chlorophyll-a ratios.

We then showed the possibility of associating the variability of some pigment concentrations with the derivative value (or slope) of the log of the absorption spectrum. By examining the SOMs of the spectral derivatives, we were able to propose two simple empirical models, one relating the fucoxanthin/Tchl-a ratio to the spectral derivative at 510 nm, the other relating the Tchl-b/Tchl-a ratio to the spectral derivative at 640 nm. These new empirical laws could hardly have been derived by reasoning alone, as discussed in Subsection 3.A for the fucoxanthin/Tchl-a ratio. The optimal wavelength for the derivative was found by carefully inspecting the maps of the derivatives provided by the SOM algorithm and comparing them to the map of the fucoxanthin/Tchl-a ratio.

We also revisited the Eisner et al.24 relationship, showing the existence of a linear relationship between the slope of the spectrum at approximately 510 nm and the log10(TPPC/TPSC), rather than with the TPPC/TPSC ratio as argued by Eisner et al.24 This was possible because of the wider range of the TPPC/TPSC ratio in the present data set.

The SOM method allowed us to retrieve the pigment concentrations with respect to the spectrum characteristics. To our knowledge, this challenging problem has never been solved in a global manner. The SOM was able to retrieve the concentrations of the main groups of pigments with reasonable accuracy. To assess its performance, we compared it with those given by a rough pigment retrieval estimation from Tchl-a by using a regression for which Tchl-a was obtained from \( a_{440} \). These regressions are used only to provide benchmarks for statistical estimator values to assess the skillfulness of the SOM method. The SOM retrieval gives better statistical estimator values for each pigment than those computed from linear regressions. A major advantage of the SOM method is that it allows us to retrieve the pigments globally without any \( a \) priori hypothesis with the same algorithm. Improving the pigment retrieval should imply the use of methodologies specific to each pigment, such as multilayer perceptrons, which are other types of neural network well adapted for inverting nonlinear processes.27

Since specific pigment concentrations characterize the phytoplankton species and modulate the spectral variation in marine reflectance,6 it is expected that this work will contribute to identify some phytoplanktonic species from space by processing the signal received by multispectral ocean color sensors. This methodology is general and can be used to analyze a large class of complex data.

Appendix A. Linear Regression

To establish the quality of a linear or log-linear relationship, two values, \( R^2 \) and \( s \), are usually calculated. We consider a data set of \( n \) samples \( (x_i, y_i^{\text{observed}}) \). The determination coefficient \( R^2 \), which is referred to as goodness of fit, represents that part of the variation in \( y \) explained by \( x \). It is given by

\[
R^2 = \frac{\sum (y_i^{\text{observed}} - \bar{y})^2}{\sum (y_i^{\text{observed}} - \bar{y})^2},
\]

where \( \bar{y} \) is the mean of the \( y_i^{\text{observed}} \); \( s \), which is referred to as the rms, is an estimation of the error on \( y \) and is given by

\[
s = \left( \frac{\sum (y_i^{\text{observed}} - y_i^{\text{estimated}})^2}{n - 2} \right)^{1/2}.
\]

Appendix B. Error Performances

The relative performances were calculated after discarding the \( x_i^{\text{observed}} \) values under 0.05 for which relative error performances are not appropriate. A small variation is still a small variation, even if it is a large percentage.

Let us consider a variable \( x_i, i = 1, 2, 3, \ldots, N \). The RMSE is defined as

\[
\text{RMSE} = \left[ \frac{1}{N} \sum_{i=1}^{N} (x_i^{\text{estimated}} - x_i^{\text{observed}})^2 \right]^{1/2}.
\]

The RRMSE is defined as

\[
\text{RRMSE} = \left[ \frac{1}{N} \sum_{i=1}^{N} \left( \frac{x_i^{\text{estimated}} - x_i^{\text{observed}}}{x_i^{\text{observed}}} \right)^2 \right]^{1/2}.
\]
and the VRRMSE of dimension $M$ is defined as

$$\text{VRRMSE} = \frac{1}{N} \sum_{j=1}^{N} \left[ \frac{1}{M} \sum_{i=1}^{M} \frac{(x_{ij} - \bar{x}_{ij})^2}{\bar{x}_{ij}^2} \right]^{1/2}.$$  \hspace{1cm} (B3)

The VRRMSE computes an estimate of the relative vector error in the following manner. Let us define the pigment concentration vector whose components are different pigment concentrations. Let us define a vector error, which is the sum of the relative error of the component of the pigment concentration error. The VRRMSE is the mean of the vector errors. The VRRMSE takes into account a potential link between the different components of the vector.

This work is a contribution to the European Neural Algorithms for Ocean Color project (EVG1-CT-2000-00034) and to several projects that have been funded by the Processus Biogéochimiques dans l’Océan et Flux national program (EUMELI, EOPE, FON-TAL, PROSOPE, POMME). The authors thank the chief scientists of the cruises during which the in situ data were collected and H. Claustre, J. Ras, K. Oubelkheir, K. Allali, C. Cailliau, J. C. Marty, N. Sadoudi, D. Tailliez, and F. Vidussi, who participated in sample collection, HPLC measurements, or absorption measurements. The authors also thank the two anonymous reviewers and H. Claustre for their help and suggestions.

References


