In situ multispectral fluorescence, a new method for the assessment of phytoplankton community taxonomic composition from BGC-Argo floats

1) Introduction

Phytoplankton play a key role in numerous biogeochemical cycles. In the current context of global changes it is crucial to monitor its dynamic on the global scale. The emergence of new observation tools such as Biogeochemical Argo (BGC-Argo) Floats opened the possibility to embark sensors to collect continuous vertical optical and chemical data on a quasi-global scale. However, the phytoplankton component is mostly monitored through its biomass using a fluorometer to estimate chlorophyll-a concentration ([chla]). However, the phytoplankton community composition plays a key role in the carbon cycle in many ways... (blabla taille cellule, efficacité photosynthèse taxa dépendant, NPP..) Some methods have been previously developed to estimate an index of the size structure of the community, either from optical data or from fluorescence profiles. Even if those measurements provide a good estimation of the size structure of the phytoplankton community we still miss the taxonomic variability hidden inside of each size class. Yet, it is established that the taxonomic composition of the phytoplankton community plays a significant role in the carbon cycle. This study aims to evaluate a new method, based on *in-situ* multispectral fluorescence, that would enable the estimation of the phytoplankton community taxonomic composition. Multispectral fluorescence relies on the diversity of pigment composition of the different phytoplankton taxon (blabla difference pigment composition induit difference de fluorescence spectra donc possible de retrouver composition du phyto à partir de différentes longueurs d’ondes). Its composition influence the absorption spectra of the cell, and so, would define its fluorescence response in different wavelength. We hypothesize that the pigment composition of phytoplankton on the global scale is contrasted enough to induce significant variability of fluorescence response between the three wavelength of excitation of the multispectral fluorometer that would allow to define a classification method from the fluorometer output.

In this study we combine laboratory and field work to investigate the possibility to retrieve a taxonomical information from multispectral fluorescence in combination with already equipped BGC-Argo float sensors. We selected ten different phytoplankton strains that correspond to the typical diversity observed during the seasonal succession of the north western (NW) Mediterranean sea. We measured the fluorometer signal on those ten strains in controlled conditions, to test the multispectral fluorometer sensitivity to taxa variation. Ultimately, we introduce a classification method to assess the different phytoplankton communities based on a year of observations of phytoplankton communities composition and multispectral fluorescence values in the NW Mediterranean sea.

2) Material & Methods

2.1) Laboratory work

2.1.1) Strain culture

Ten strains were selected for the protocol and provided by the Roscoff Culture Collection (RCC). Three Synechococcus strains that are commonly observed in the summer surface waters, nom des souches, three Prochlorococcus that are representative of deep picophytoplankton communities, nom des souches, three diatoms with different size spectra, nom des souches, one dinoflagellate and one pelagophyte….

10 souches : PCC9511, RCC156, RCC2374, RCC 2319, RCC2379, RCC76, RCC100, RCC1717, RCC4213, RCC3006.

Multispectral fluorescence measurements were taken at different concentration. The culture was diluted in different medium volume to produce a dilution scale ranging from 10 mg m-3 to 0.1 mg m-3.

All strains were grown under the same controlled conditions, \*récupérer données de temp entre 26/10/2020 et 06/11/2020\*, at 52 μE of continuous white light, in K+Si or PCRS11 medium. The growth state of the culture was monitored with a Pulse Amplification Method (PAM) fluorometer, using the Fv/Fm index. The protocol of multispectral fluorescence measurement was performed on each culture in the end of the exponential growth phase. Expliquer pourquoi.. Afterward each culture was putted back in growth condition until a new exponential growth, three times, to get biological triplicate. Expliquer pourquoi…

2.1.2) Multispectral fluorescence measurements

Pour la culture mère et chaque dilution,

- mesure au noir, après une acclimatation des cultures au noir. (Combien de temps ?)

- Prélever 2 mL dans un cuve de spectrofluorimétrie (ref) pour mesurer le F0 le Fmd et le Fm et calculer le Fv/Fm à l’aide du PAM (ref) ;

- S’assurer que le fluorimètre est droit et au centre du bécher immergé de 5mm dans la culture. S’assurer de l’absence de bulles au niveaux des canaux d’excitations et du canal de mesure, mettre la culture sous légère agitation statique. Mesurer 3 \* 1minutes avec 2 minutes de désexcitation entre chaque acquisition.

Pour la culture mère, créer une référence avec le PAM.

2.2) Time series acquisition

The MF was deployed at sea from December 2020 to October 2021, in the North Western Mediterranean sea, at the B.. BOUSSOLE site. This site was chosen for its seasonality that leads to contrasted phytoplankton community succession through the year. Its diversity of phytoplankton communities may be compared to the trophic gradient observed on the global scale going from microphytoplankton dominated communities to picophytoplankton dominated communities. Once a month, a cruise took place with a conductivity temperature depth (CTD) device mounted with the multispectral fluorometer (MF) to cast a profile from surface to 400m depth. In addition, the Wetlabs ECO-series (ref) fluorometer and a transmissiometer (ref) were mounted on the CTD providing BBP and Cp measurements concurrently to the multispectral fluorescence signal…

Profiles were binned every meter and spike were removed following a moving median method…

2.3) Statistics

2.3.1) Classification methods

The classification of the laboratory samples was determined using a Support Vector Machine Classifier (SVC) which allows a visual interpretation of the results. The SVC was designed using a Radian Based Function (RBF) kernel in order to precise the discrimination power of the model, by allowing non-linearity of the decision boundaries.

The apparent complexity of in-situ phytoplankton community composition compared to laboratory monospecific cultured encouraged us to use a more complex classification method, including new descriptors. Thus, we used an Histogram Gradient Boosting (HGB) classification model, including multispectral fluorescence data and optical measurement that can be monitored with BGC-Argo floats (i.e. BBP and Cp). This type of machine learning model is particularly well suited for tabular data (avantages de la méthode) etc.. The imbalance in the number of samples per cluster was counterbalanced by a SMOTE oversampling. Each cluster was oversampled to get as much observation as the largest cluster. In the end, each cluster is composed of 32 samples. The descriptors do not rely on absolute values to prevent any overfitting to the local environment. The hyperparameters of the model were defined using a cross validation grid search etc… etc.. détailler le modèle.

Ultimately the performance of the HGB classification model is discussed regarding 4 different possible configuration of BGC-Argo sensor packages. One with MF, transmissiometer etc…

2.3.2) Clustering method

The *in-situ* HPLC data were clusterized in order to define phytoplankton communities that are meant to be discriminated by the HGB classification method. An assemblage of pigments concentration (i.e. c chlorophyll c1, c2 and c3; peridinin; but; fucoxanthin; neoxanthin; prasinoxanthin; violaxanthin; hex; diadinoxanthin; alloxanthin; diatinoxanthin; zeaxanthin; lutein; divinyl chlorophyll b; chlorophyll b; divinyl chlorophyll a and chlorophyll a) were used as descriptors of a correspond analysis (CA). The first two dimensions of the CA were used as numerical estimation of the pigment composition resemblance of each samples. A clusterization was then made on the first 2 dimensions, using a Hierarchical Ascending Classification (HAC). The number of cluster was decided by looking at the resulting cluster dendrogram. As there is a strong difference of pigment composition between prokaryotic picophytoplankton communities and micro/nano-phytoplankton communities the same method was applied excluding picophytoplankton-dominated samples. In the end 4 different phytoplankton communities were defined.

3) Results and discussion

3.1) Discrimination of phytoplankton taxa in controlled conditions

In order to investigate the possibility of using the MF as a tool for taxa discrimination we measured multispectral fluorescence response of the cultures in controlled conditions, on a range of concentration going from 0.1 to 10 mg m-3. The ratio between the three different fluorescence signal from the three excitation wavelengths of the 3X1M is shown on the figure 1. The aims of the classification is to discriminate taxa rather than species, so the different strains are group by taxa in this study. The different taxa are spatially grouped in the scatterplot, even despite the variability of strains in each of them. This indicates that reducing the taxonomic information to this level is well suited for interpreting the fluorometer outputs. Synechococcus have a consistently high value of F532/F470 and a relatively heterogeneous values of F440/F470. Pelagophytes have low values of both F532/F470 and F440/F470. Prochlorococcus have medium to high values of F532/F470 and consistently high values of F440/F470. Finally, Diatoms are found in the middle of the scatterplot with median values of F532/F470 and F440/F470.

Higher values of F532/F470 in the Synechococcus and Prochlorococcus taxa may be explained by a higher fluorescence response with the 532 nm excitation, induced by the presence of phycobilins in Synechococcus (Six et al., 2007) and, in less extent, in Prochlorococcus (Wiethaus et al., 2010). Phycobilins extend the light harvest complex, they have a peak absorption at 560 nm and transmit absorbed energy to the photosystem, leading to a significant fluorescence response when excited at 532 nm. The difference of F440/F470 ratio values between pelagophytes and diatoms may be explained by a difference of carotenoids compositions.

The visual discrimination of the different taxa in this scatterplot indicates that the fluorometer response is taxa-dependent. Moreover, its response should be sensitive enough to be successfully used in a SVC classification model. (fig. 2) However there is some ambiguous samples, we observe the lowest precision on the diatoms…

The multispectral fluorescence at 440, 470 and 532 nm is sensitive enough to discriminate four different taxa in controlled condition. Thus we can hypothesize that it can provide significant information on the taxonomical composition of *in-situ* phytoplankton composition. To validate this assumption, we will evaluate the possibility to discriminate different phytoplankton communities in the NW Mediterranean sea using *in-situ* multispectral fluorescence and optical measurements.

3.2) Identification of phytoplankton communities

The different phytoplankton communities were defined using a pigment based clusterization that allows the clustering of samples with similar pigment composition.

To compromise between the number of samples per cluster and the ecological meaning of each cluster we used 4 clusters (figure 3). The first one correspond to late summer deep picophytoplankton and winter communities with a significant proportion of chlorophyll-b. The second cluster correspond to the bloom community with a domination of microphytoplankton with a high fucoxanthin proportion, typically associated to diatoms. The third one is associated to summer communities below the DCM with a mixed composition of micro and nanophytoplankton. Finally, the fourth one corresponds to surface summer picophytoplankton dominated communities, typically associated to Synechococcus.

In order to compare the relative pigment composition of this four cluster in comparison to the selected strains we built a correspondence analysis from the pigment composition of the strains and projected the pigment composition of the cluster samples on it. We observe three distinct poles corresponding to the different strains taxa. One is composed of the Diatoms and Pelagophytes while the two other poles correspond respectively to *Synechococcus* and *Prochlorococcus*. The *in-situ* samples are evenly spread in the center of the plan, indicating that we find the same variability of pigment composition as the laboratory strains. Moreover, the four cluster are well distinguished in this plan. As we already showed that the MF can be used to discriminate the different poles of this CA we can expect that it will be performant enough to discriminate the *in-situ* clusters.

However, the *in-situ* samples are not as far from the middle as the laboratory strains. This results from the complexity of environmental samples that are composed of different species instead of one monospecific strains, leading to mixed pigment composition, which may reduce the discrimination efficiency of the MF.

3.3) Discrimination of phytoplankton communities from *in-situ* measurements

The pigment composition of the *in-situ* phytoplankton communities as been showed to be in the range of the MF discrimination capacity. However, the complexity of environmental samples implies to use two optical descriptors, BBP and Cp, that are commonly used as proxies of phytoplankton size class carbon contribution. Moreover, the use of a HGB classifier instead of a SVC classifier has been chosen to favor global performance and generalization of the model.

The mean global accuracy of the HGB classification using all the descriptor on a stratified shuffle split cross validation with 20 splits is 72% (+/- 10%). This demonstrates the possibility to use multispectral fluorescence and optical measurements to discriminate the phytoplankton communities taxonomical composition from embarked sensors on BGC-Argo floats.

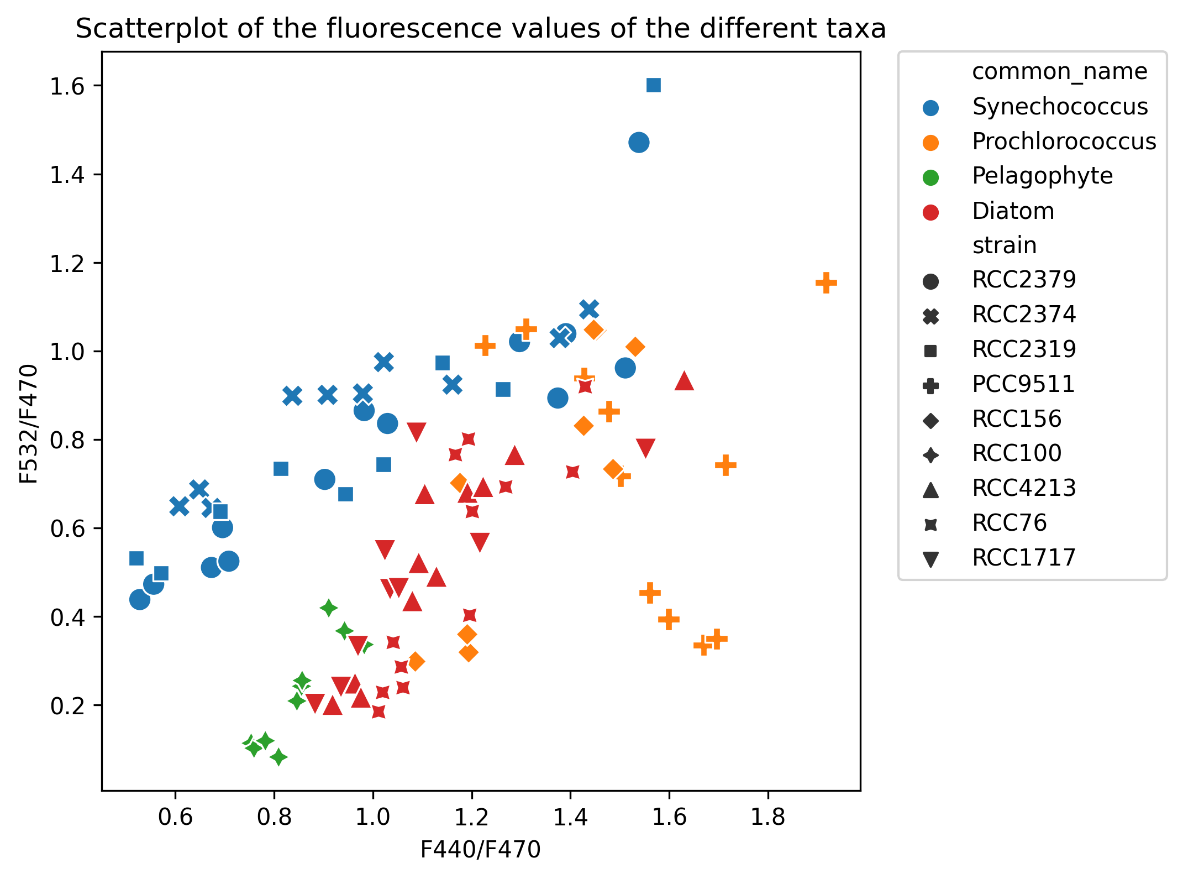
The mean impurity of the descriptors (table 1) indicates a significant role of all the descriptors with a particularly important effect of the ratio between F470 and Cp. The relatively low impurity of descriptors involving the F532 measurement (table 1) may be explained by the relatively low sensitivity of the MF…

The best accuracy and recall score are obtained on the cluster 1 and 4, indicating that this method best perform to discriminate the prokaryote to eukaryote communities. This is of a strong interest regarding biogeochemical cycles….

Equipping all of this sensors on each floats is expensive and sometimes demands a significant logistic (???). Thus, we examined the sensitivity of the model on different BGC-Argo configurations (figure 5), in view of adapting this method on other configurations. Removing transmissiometer leads to a significant decrease in the global accuracy and recall of the model (figure 5), as well as removing the 532nm excitation fluorescence or BBP signal. The performances of the model is not good enough to discriminate the fours defined taxa.

However, we may reduce the level of discrimination of the model by reducing the number of cluster in view of increasing the performances of the models with less sensors. In doing so, we can evaluate the performances on two other cluster configurations. Reducing the number to three clusters by grouping the cluster 2 and 3, leading to a discrimination between surface summer picophytoplankton with zeaxanthin, deep and winter picophytoplankton with chlorophyll and divinyl chlorophyll b and mixed communities of micro and nanophytoplankton. The lower level of discrimination consists of having only two clusters corresponding to picophytoplankton with zeaxanthin or chlorophyll b and microphytoplankton and nanophytoplankton. The performance of the model is evaluated by looking at the mean balanced recall on the same cross validation method as used before (figure 6). In the best sensor package configuration the number of cluster does not interfere on the performance with a consistent balanced recall of around 75%. On the contrary, with fewer sensors reducing the number of clusters leads to an increase in performances. While most of the configurations are still around 60% of recall, the prediction of two clusters with F440 F470 and BBP seems to stand out with a recall of 70% (figure 6 and table 2). This results enlighten the possibility to retrieve a taxonomical information with fewer sensors.

4) Conclusion and perspectives

Figure 1: Scatterplot of the laboratory measured values of the ratios F440/F470 and F532/F470 of each strains of the different taxa.

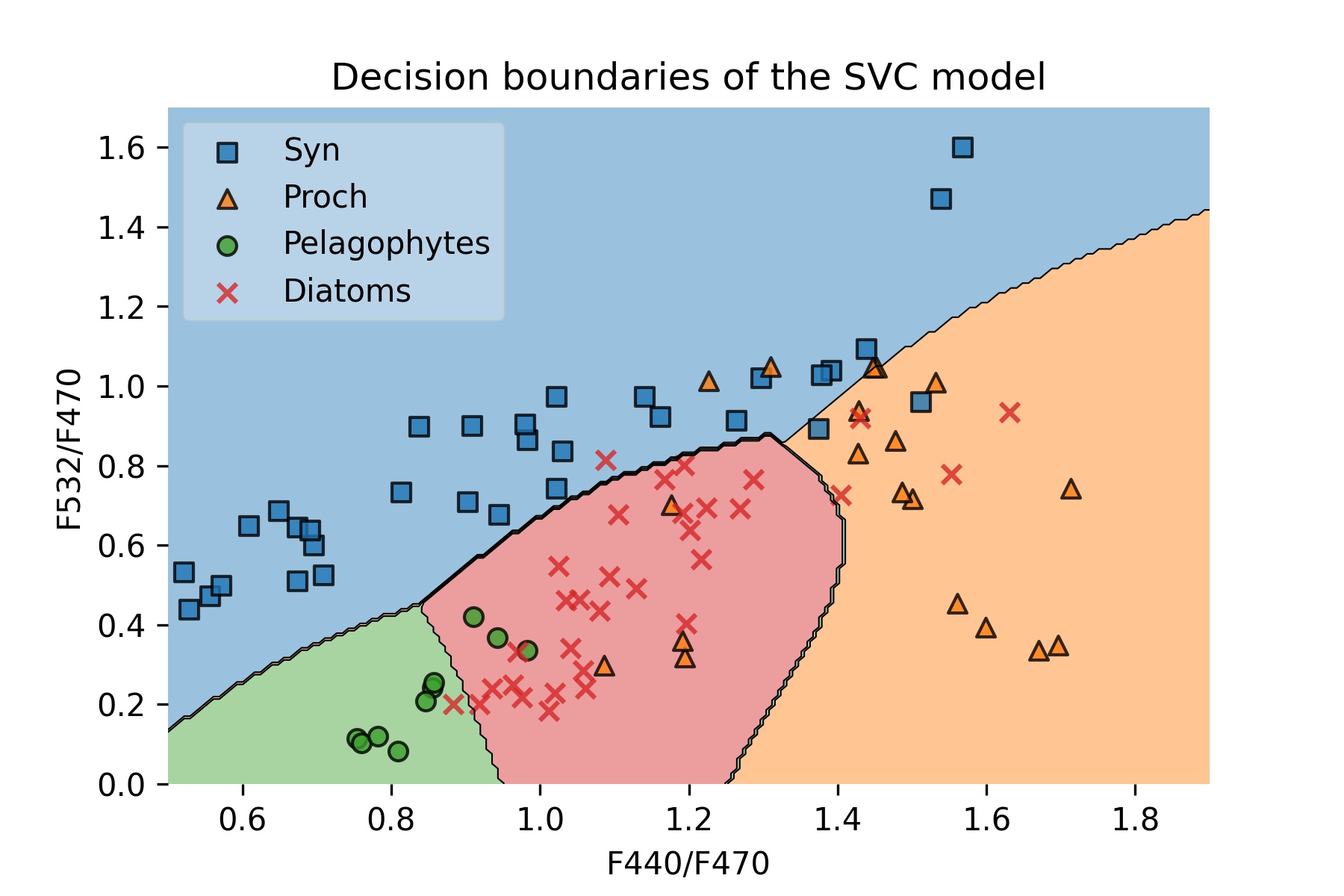


Figure 2 : Decision boundaries of the Support Vector Classifier with a Radial Basis Function kernel. Balanced accuracy : 86.7% ; balanced recall : 86.8%

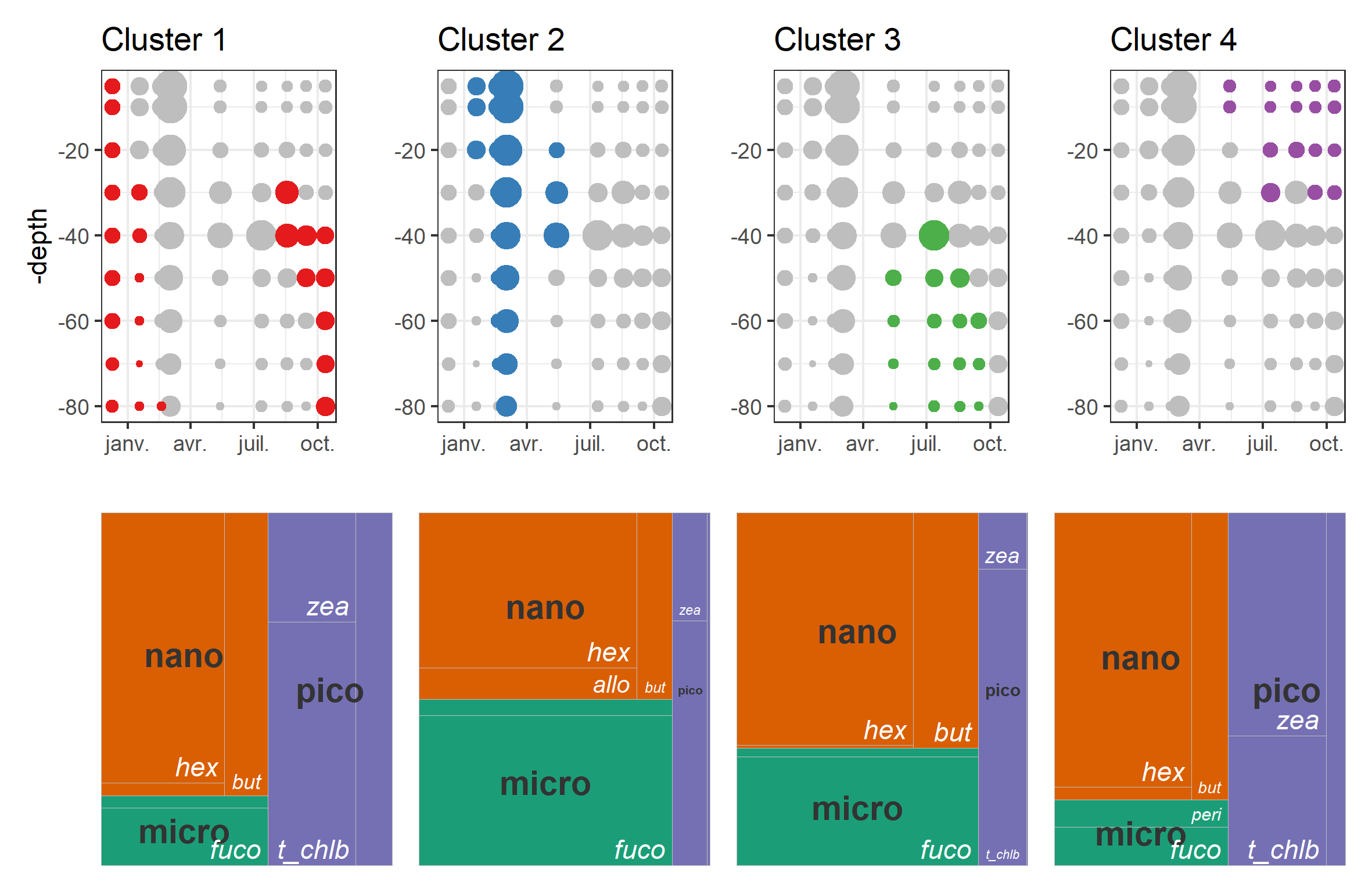
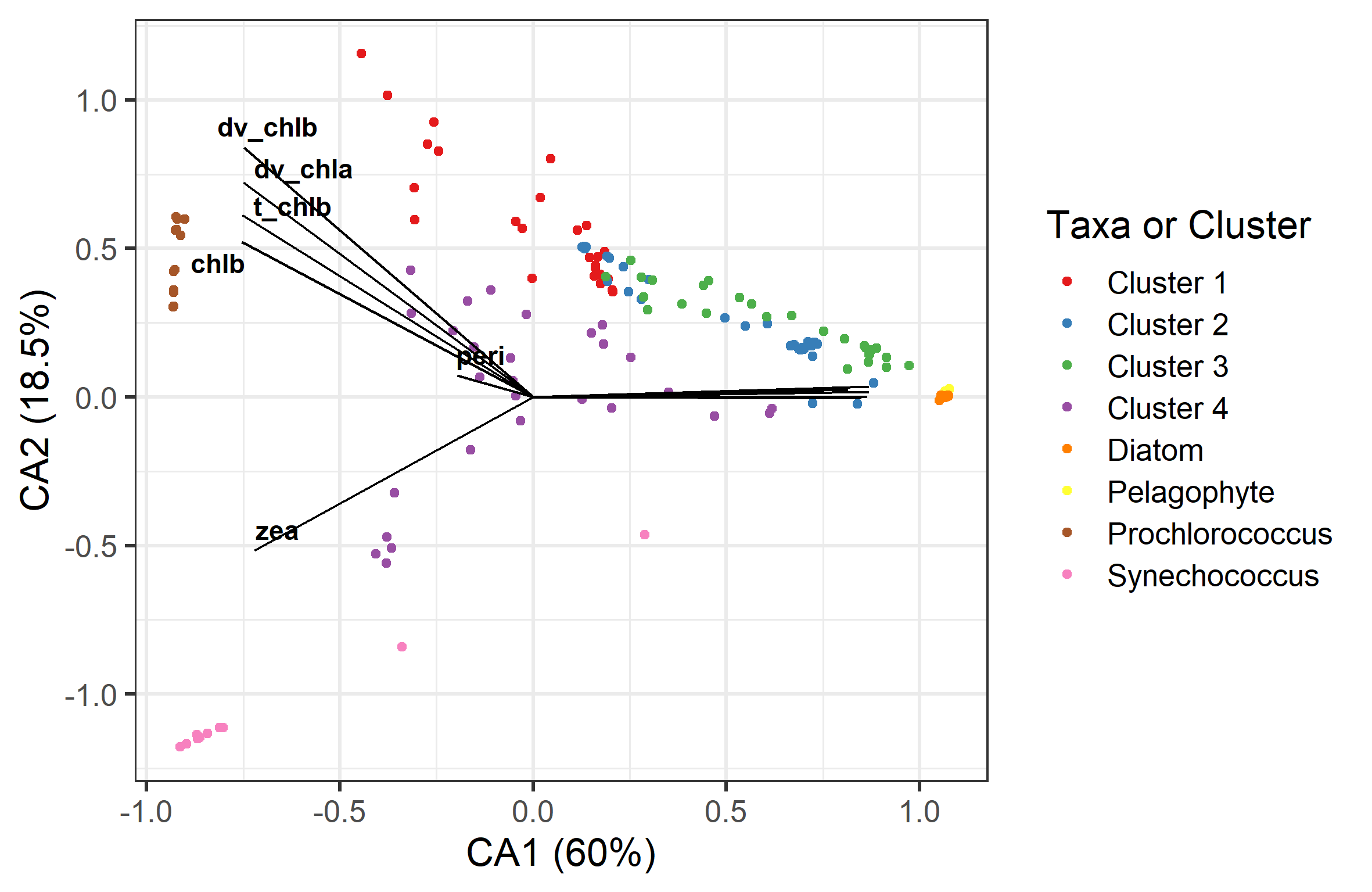
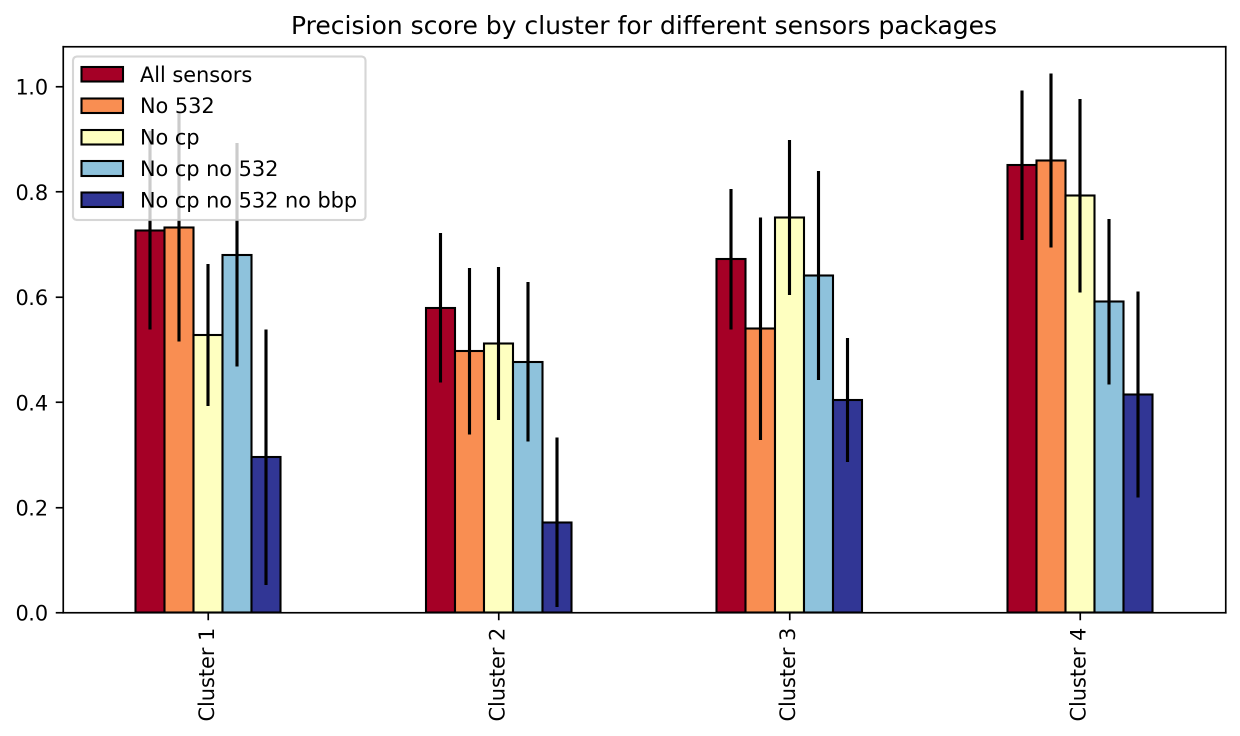


Figure 3 : Localization of each cluster in the time series, regarding the depth and time of the year. Relative pigment composition of each cluster.

Figure 4 : Correspondence analysis of the pigment concentration of the selected taxa, with the projection of in-situ phytoplankton communities pigment composition.



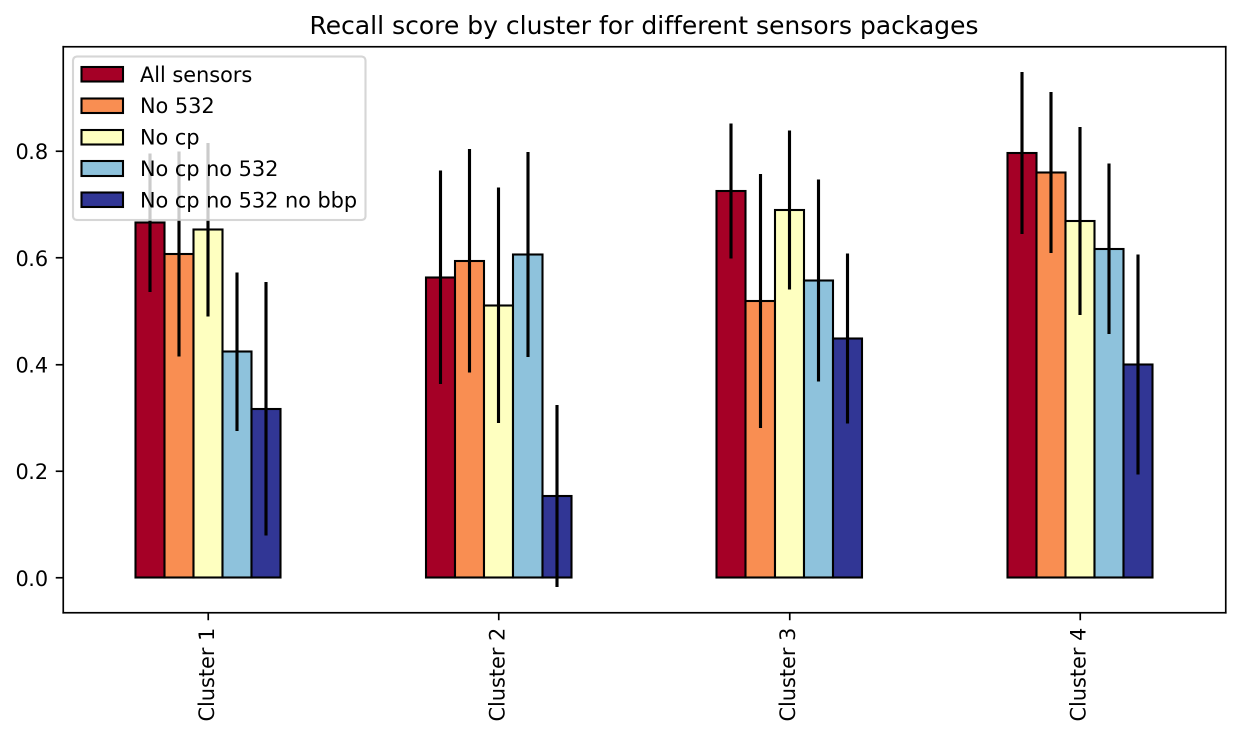


Figure 4 : Accuracy and recall of the HGB classification model for each cluster on each configuration of BGC-Argo sensors package.

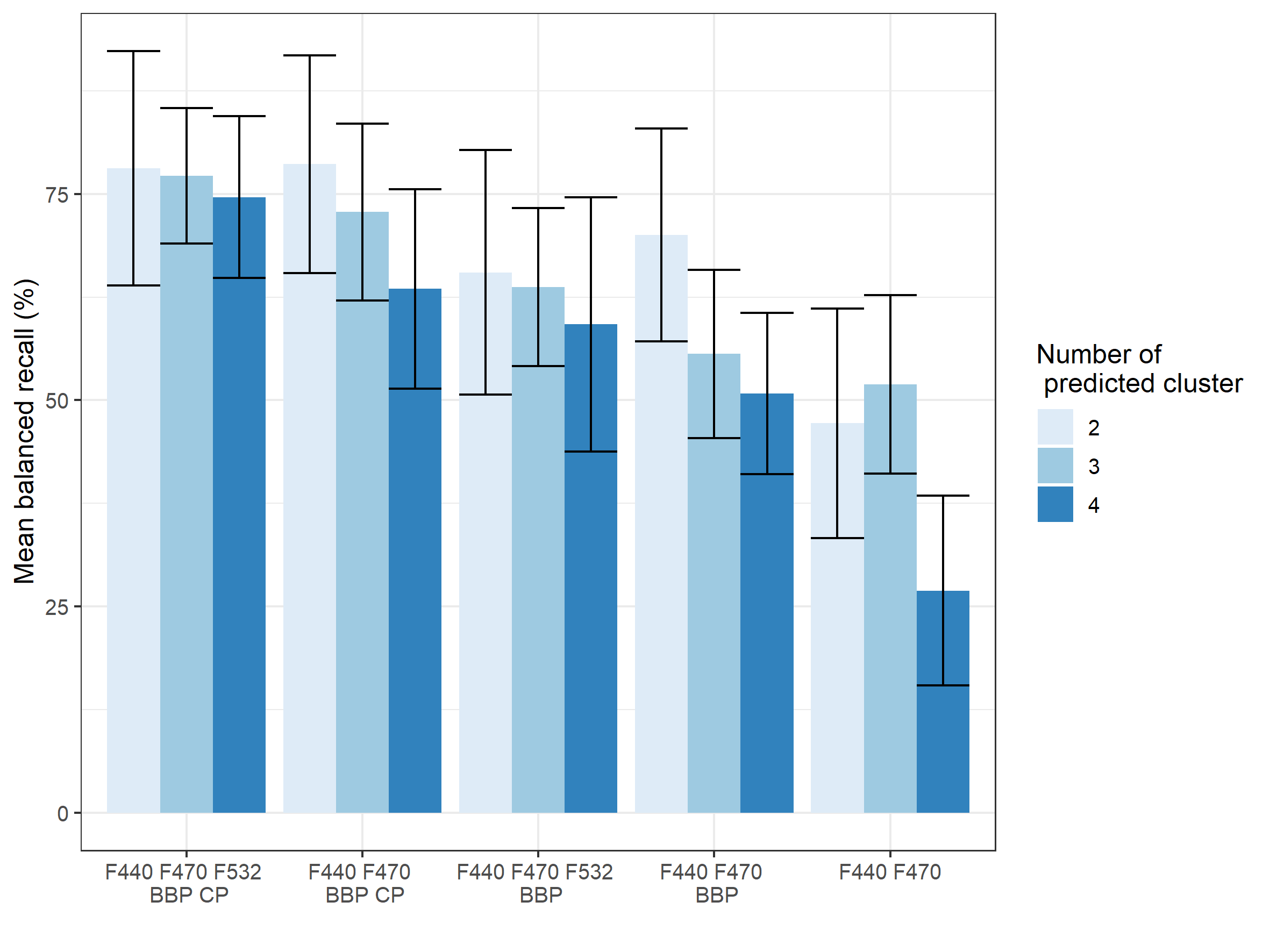
Figure 5: Values of the mean weighted recall resulting from a cross validation with different number of cluster and different sensor package.

Table 1: Description and impurity of the different descriptors.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **F440/F470** | **F532/F470** | **BBP/Cp** | **F440/BBP** | **F470/BBP** | **F532/BBP** | **F440/Cp** | **F470/Cp** |
| mean | 0.9 | 0.6 | 894 | 0.73 | 0.84 | 0.45 | 609 | 685 |
| std | 0.09 | 0.25 | 424 | 0.28 | 0.4 | 0.16 | 276 | 314 |
| min | 0.63 | 0.19 | 0.1 | 0.37 | 0.35 | 0.3 | 0.05 | 0.05 |
| max | 1.18 | 1.1 | 1983 | 1.68 | 2.22 | 1.46 | 1369 | 1527 |
| Impurity | 0.12 | 0.07 | 0.09 | 0.1 | 0.07 | 0.1 | 0.22 | 0.2 |

Table 2: Mean and standard deviation of precision and recall for different configuration of clusterization and sensors packages in percentage. The recall and precision values are obtained by a stratified shuffle cross validation with 20 splits.

|  |  |  |  |
| --- | --- | --- | --- |
| Number of cluster | Sensor package | Mean precision (+/- SD) in % | Mean recall (+/- SD) in % |
| 4 | F440 F470 F532 BBP CP | 73.3 (10.9) | 74.6 (9.8) |
| 4 | F440 F470 BBP CP | 64.5 (13.2) | 63.5 (12.1) |
| 4 | F440 F470 F532 BBP | 59.9 (16.1) | 59.2 (15.4) |
| 4 | F440 F470 BBP | 50.5 (8.9) | 50.8 (9.8) |
| 4 | F440 F470 | 27.2 (12.4) | 26.9 (11.5) |
| 3 | F440 F470 F532 BBP CP | 76 (7.9) | 77.2 (8.2) |
| 3 | F440 F470 BBP CP | 72.9 (10.1) | 72.8 (10.7) |
| 3 | F440 F470 F532 BBP | 63.3 (10.7) | 63.7 (9.6) |
| 3 | F440 F470 BBP | 55 (10.7) | 55.6 (10.2) |
| 3 | F440 F470 | 51.9 (10.5) | 51.9 (10.8) |
| 2 | F440 F470 F532 BBP CP | 78.5 (14.8) | 78.1 (14.2) |
| 2 | F440 F470 BBP CP | 77.7 (14.5) | 78.6 (13.2) |
| 2 | F440 F470 F532 BBP | 63.7 (15.4) | 65.5 (14.8) |
| 2 | F440 F470 BBP | 69 (13.1) | 70 (12.9) |
| 2 | F440 F470 | 47.4 (13.2) | 47.2 (13.9) |