Définition multivariée et multi-échelle d'états environnementaux par Machine Learning : caractérisation de la dynamique phytoplanctonique

Kelly Grassi

Ifremer/ULCO-LISIC
WeatherForce

École doctorale **ED SMRE - 104**
Unité de recherche **Ifremer**

Thèse présentée par **Kelly Grassi**

Soutenue le **19 novembre, 2020**

En vue de l'obtention du grade de docteur de l'Ifremer/ULCO-LISIC

Discipline **Biologie de l'environnement, des populations, écologie**

Définition multivariée et multi-échelle d'états environnementaux par Machine Learning : Caractérisation de la dynamique phytoplanctonique.

Thèse dirigée par	Alain Lefebvre	directeur
	Andre Bigand	co-directeur
	Émilie Poisson-Caillault	co-encadrante

Composition du jury

Rapporteurs	David Nérini	MCF-HDR au MIO Marseille
	Véronique Creach	DR au CEFAS Lowestoft
Examinateurs	François Cabestaing	professeur à l'Université de Lille
	Pascal Claquin	professeur à l'Université de Caen
	Cédric Bacher	CR-HDR à l'Ifremer Brest
Directeurs de thèse	Alain Lefebvre	CR-HDR à l'Ifremer Boulogne-sur-mer
	Andre Bigand	MCF-HDR à l'ULCO-Lisic Calais
	Émilie Poisson-Caillault	MCF-HDR à l'ULCO-Lisic Calais

Colophon

Mémoire de thèse intitulé « Définition multivariée et multi-échelle d'états environnementaux par Machine Learning : Caractérisation de la dynamique phytoplanctonique. », écrit par Kelly Grassi, achevé le 2021-09-14, composé au moyen du système de préparation de document LaTeX et de la classe yathesis dédiée aux thèses préparées en France.

Ifremer/ULCO-LISIC
WeatherForce

École doctorale **ED SMRE - 104**
Unité de recherche **Ifremer**

Thèse présentée par **Kelly Grassi**

Soutenue le **19 novembre, 2020**

En vue de l'obtention du grade de docteur de l'Ifremer/ULCO-LISIC

Discipline **Biologie de l'environnement, des populations, écologie**

Définition multivariée et multi-échelle d'états environnementaux par Machine Learning : Caractérisation de la dynamique phytoplanctonique.

Thèse dirigée par	Alain Lefebvre	directeur
	Andre Bigand	co-directeur
	Émilie Poisson-Caillault	co-encadrante

Composition du jury

Rapporteurs	David Nérini	MCF-HDR au MIO Marseille
	Véronique Creach	DR au CEFAS Lowestoft
Examinateurs	François Cabestaing	professeur à l'Université de Lille
	Pascal Claquin	professeur à l'Université de Caen
	Cédric Bacher	CR-HDR à l'Ifremer Brest
Directeurs de thèse	Alain Lefebvre	CR-HDR à l'Ifremer Boulogne-sur-mer
	Andre Bigand	MCF-HDR à l'ULCO-Lisic Calais
	Émilie Poisson-Caillault	MCF-HDR à l'ULCO-Lisic Calais

Ifremer/ULCO-LISIC
WeatherForce

Doctoral School **ED SMRE - 104**

University Department **Ifremer**

Thesis defended by **Kelly Grassi**

Defended on **19th November, 2020**

In order to become Doctor from Ifremer/ULCO-LISIC

Academic Field **Biologie de l'environnement, des populations, écologie**

Multivariate and multi-scale definition of environmental states by Machine Learning: Characterization of phytoplankton dynamics

Thesis supervised by	Alain Lefebvre	Supervisor
	Andre Bigand	Co-Supervisor
	Émilie Poisson-Caillault	Co-Monitor

Committee members

Referees	David Nérini	MCF-HDR at MIO Marseille
	Véronique Creach	Senior Researcher at CEFAS Lowestoft
Examiners	François Cabestaing	Professor at Université de Lille
	Pascal Claquin	Professor at Université de Caen
	Cédric Bacher	CR-HDR at Ifremer Brest
Supervisors	Alain Lefebvre	CR-HDR at Ifremer Boulogne-sur-mer
	Andre Bigand	MCF-HDR at ULCO-Lisic Calais
	Émilie Poisson-Caillault	MCF-HDR at ULCO-Lisic Calais

Cette thèse a été préparée dans les laboratoires suivants.

Ifremer

150 quai Gambetta
62200 Boulogne-sur-mer
France

✉ centre-de-boulogne@ifremer.fr
Site https://wwz.ifremer.fr/manchemerdunord/Implantations/Boulogne-sur-Mer

LISIC

Maison de la Recherche Blaise Pascal
50, rue Ferdinand Buisson
CS 80699
62228 Calais Cedex
France

✉ secretariat@univ-littoral.fr
Site https://www-lisic.univ-littoral.fr/

WeatherForce (financeurs)

30 rue de Metz
31000 Toulouse
France

✉ info@weatherforce.org
Site https://weatherforce.org/

Le paradis, c'est le moment où les rêves que nous formons aujourd'hui seront réalisés

---

Simone de Beauvoir

Il faut d'abord savoir ce que l'on veut, il faut ensuite avoir le courage de le dire, il faut enfin l'énergie de le faire

---

Georges Clémenceau

Remerciements

"Collecter des données n'est pas savoir. La véritable connaissance est un processus de réflexion, un savoir-faire, et non une accumulation de données, à rebours de l'idée reçue que la connaissance se trouverait toute faite dans quelque chose : un livre, un ordinateur, une encyclopédie en ligne."
La longue montée de l'ignorance - Dimitri Casali

Pour ma part, cette réflexion ne s'est pas faite toute seule, elle est le fruit d'un travail d'équipe, d'un partage de connaissances et de données de nombreuses personnes, sans lesquelles cette thèse n'aurait pas pu aboutir. C'est pourquoi, je tiens à remercier toutes les personnes qui ont participé de loin ou prêt à mon projet de thèse.

Je souhaite donc remercier en premier lieu mon directeur de thèse, Alain Lefebvre, responsable du laboratoire Environnement Ressources de l'Ifremer de Boulogne-sur-Mer pour m'avoir accueilli au sein de son équipe et de son projet. Je lui suis également reconnaissante pour le temps conséquent qu'il m'a accordé, pour le savoir et la rigueur qu'il m'a fait partager. Et surtout, je le remercie pour ne pas avoir trop réveillé la force de Dark Vador qui sommeille en lui face au Bébé Yoda que j'ai pu être.

J'adresse de chaleureux remerciements à mon encadrante de thèse, Émilie Poissons-Caillault, pour son énergie et son soutien, pour ses conseils et son écoute qui ont été prépondérants pour la bonne réussite de cette thèse. La team Girl Power a été un élément moteur pour moi et j'ai pris un grand plaisir à travailler avec elle. Et surtout, merci de m'avoir redonné confiance dans les moments difficiles.

Je tiens aussi à remercier mon troisième encadrant, André Bigand, Maître de Conférence HDR au Laboratoire d'informatique Signal et Image de la Côte d'Opale (LISIC) pour nos discussions et ses conseils qui m'ont accompagné durant ces 3 ans.

J'adresse tous mes remerciements à Monsieur David Nerini, Maître de conférences à l'Institut Méditerranéen d'Océanologie (MIO), ainsi qu'à Madame Véronique Creach directrice de recherche au Centre for Environment, Fisheries and Aquaculture Science (CEFAS), de l'honneur qu'ils m'ont fait en acceptant d'être les rapporteurs de cette thèse. Leurs remarques m'ont permis d'envisager mon travail sous un autre angle.

De même, merci à François Cabestaing, Professeur à l'Université de Lille, Pascal Claquin, Professeur à l'Université de Caen Normandie (ICN), et Cédric Basher, responsable de l'Unité de Recherche 'Dynamique des Écosystèmes Côtier' à l'Ifremer de Brest, d'avoir accepté de participer à mon jury de thèse. Ils ont pris le temps de me lire, de m'écouter et de discuter avec moi. Pour tout cela, je les en remercie.

J'adresse de sincères remerciements à Gérald Grégori, Chargé de recherche à l'Institut Méditerranéen d'Océanologie (MIO), et Yann Leredde, Maître de conférences au laboratoire de Géoscience Montpellier pour avoir accepté de suivre mon travail lors de mes comités de thèse. Leurs remarques m'ont permis de faire avancer mes réflexions et j'ai pu à ces occasions apprécié leur enthousiasme et leur sympathie.

Je remercie l'ensemble de la communauté CAOST-HF qui a joué un rôle important dans la construction et la diffusion de ma thèse. Je souhaite remercier en particulier, Guillaume Charria responsable du Laboratoire Océan Côtier à l'Ifremer Brest, pour les échanges et l'évaluation de mes idées, ainsi que l'intégration de ma thèse au sein de la communauté ainsi que Ivane Pairaud, chercheuse au Laboratoire Océan Côtier à l'Ifremer Brest, Peggy Rimmelin-Maury, responsable des séries d'observation physico-chimiques au Centre National de Recherche Scientifique (CNRS) et Rosalie Fuchs cadre de recherche à l'Ifermer de la Seyne-sur-mer, pour les données, les informations et les précieux conseils, fournis pour les stations MesuRho et MAREL-Iroise. Et enfin, Francis Gohin chercheur au Laboratoire d'Écologie Pélagique de l'Ifremer Brest pour l'extraction des données satellites.

Mes remerciements vont également à toute l'équipe de WeatherForce, qui a tout fait pour m'aider et qui m'a suivie de près durant ces 3 ans. Merci à Pascal Venzac et Christine David les co-dirigeants de WeatherForce sans qui tout ce projet n'aurai pas vu le jour. Merci à toi Julien pour ta collaboration active et pour ton sourire, merci à Dark Peter d'avoir suivi dark Kelly, à la petite Anouk pour nos « balades nocturnes », à Emma ma coloc préférée et à la joyeuse Morgane pour nos moments potins et pas que...

Je remercie toute l'équipe du LISIC pour leur accueil, merci pour vos conseils lors des séminaires internes. Un grand merci à Pierre-Alex pour le SAV latex, tu m'auras sauvé la mise au dernier moment. Merci à Baptiste pour notre rencontre et notre amitié.

Je tiens aussi à remercier Monsieur Dominique Godefroy, directeur du centre IFREMER Manche Mer du Nord et à l'UMR BOREA de m'avoir accueilli au sein du centre de Boulogne-sur-Mer.

Je remercie tous mes collègues du centre Manche Mer du Nord de l'Ifremer pour m'avoir accueillie chaleureusement. Un merci particulier pour l'équipe de LER et mes collègues de bureau Camille Dezecache, Mathias Li, Thomas et surtout David Devreker qui m'a tenu compagnie durant les trois ans. Merci de m'avoir supportée. Merci aussi à mes partenaires lors des campagnes en mer auxquelles j'ai eu la chance de participer : Camille, Michèle et Remi, ces 20 jours sur la Thalassa avec vous resterons gravés dans ma mémoire. Merci à Vincent pour la mise en pratique des prélèvements REPHY sur la Somme. Et surtout merci à mes amis de tous les instants la team techniciens GeofyGeof, Duduss, thibaut et Faboucher, merci d'avoir été présent dans tous les moments off de ma thèse. Que ce soit pour écouter mes potins ou mes plaintes vous m'avez permis de me changer les idées et pour cela merci.

Je tiens aussi à remercier toute l'équipe de doctorants qui a partagé et je dirais même vécue cette thèse avec moi : Merci à Martin, Maria, « la team bassin », d'avoir parlé de physique, de latex et de tous autres sujets que le team biologie ne voulais pas entendre parler. Je me suis sentie moins seul dans ce nouveau monde et merci Maria pour tes corrections.

Merci aux « anciens » de m'avoir accueillie si chaleureusement : merci Pierre de m'avoir affublée d'un nouveau surnom « Kellox » qui est resté même après ton départ. Merci Matthew

pour les relectures de mes articles, Khalilou pour les différentes sorties hors de Boulogne et surtout merci à Julien pour tous les moments qu'on a passé et ceux qu'on passera encore, il y en a trop pour que j'en choisisse un.

Un GRAND merci aux trois fous de mon année, j'ai nommé Petite tortue, JuJu et Carlito (alias mon coloc préféré), on a vécu du bon et du mauvais, mais on a toujours été là les un pour les autres sans vous, c'est clair que mon moral n'aurait jamais tenu le coup, mais je peux le dire maintenant, on l'a fait ! Alors merci pour vos skills, vos tutos et tous les samedis « travail » et bien plus encore. Et bien sûr, je n'oublie pas les petites jeunettes Léa et Alaia merci pour votre bonne humeur et nos fou-rires.

Merci aussi à vous tous mes autres compatriotes Boulonnais Carole, Aurélie, Sophie, FanFan, Lola, Valentine, Ines, Éric, la team AccroYoga, keke, momo et j'en passe ! Vous m'avez changé les idées chaque soir et chaque week-end, grâce à vous, j'ai trouvé une nouvelle famille dans le Ch'Nord.

Merci à mes amis de toujours, ceux qui me suivent depuis bien des années et qui ont toujours été derrière moi. Merci au gang des frisé Camcam, Coco, Marjo et Claiiireee les études et notre amitié sont indissociables sans vous, je ne serai pas là ou j'en suis merci pour vos relectures, vos corrections, vos conseils et surtout pour votre soutien. Merci à tous les amis du sud Audrey, Doriane, Nelly, Steven, Attila, Pierre, Andrew, Laurent et Paolo pour leurs encouragements même si ce que je fais est toujours un mystère pour eux.
Ces remerciements ne peuvent s'achever, sans une pensée pour ma famille. Merci à mes deux plus grands fans : mes parents. Merci Papa, tu es le correcteur officiel des fautes d'orthographe de cette thèse ! Je pense que tu as lu et relu ma thèse plus que n'importe qui alors merci. Maman, merci pour ton coaching moral toujours là pour m'apporter des solutions et du réconfort. Votre présence et vos encouragements ont toujours fait partie de moi et m'ont permis et me permettront toujours d'accomplir mes souhaits et mes ambitions. Merci, à ma sœur Julia, tu me supportes depuis 24 ans et je sais que ce n'est pas toujours facile, avec toi, j'ai appris la collaboration, le self-contrôle et plein d'autres qualités indispensable au travail en équipe, alors merci, tu es toujours avec moi dans mes actions et mes pensées. Merci à mes cousines Marie et Alice qui ont suivi mon travail de loin et qui ont facilité mon intégration dans le nord, merci d'être là tout simplement. Et enfin merci à mes grands-parents, merci à toi Mamie Marie-jo, pour ton soutien inconditionnel et pour toutes les petites attentions que tu me portais même si j'étais loin de toi. Et je tiens à adresser un mot particulier à mon papi Alain qui est parti un peu avant l'achèvement de ma thèse, mais qui reste dans mon cœur pour toujours. Je le sais, tu étais et tu serais fier de moi aujourd'hui, alors merci d'avoir cru en moi jusqu'au bout. Une pensée aussi pour toi mamie Christiane mon porte-bonheur qui est depuis longtemps dans mon cœur.

Définition multivariée et multi-échelle d'états environnementaux par Machine Learning : Caractérisation de la dynamique phytoplanctonique.

Résumé

Les systèmes automatisés de mesures à haute fréquence (HF) déployés dans des écosystèmes contrastés sont censés permettre une meilleure compréhension de la dynamique de l'environnement (et du phytoplancton) en réponse aux pressions d'origine naturelle et anthropiques, ainsi que les effets directs et indirects des proliférations du phytoplancton nuisible pouvant conduire à des dysfonctionnements des écosystèmes. La compréhension de cette dynamique est également importante afin de proposer des indicateurs conformes aux objectifs des directives européennes et des conventions des mers régionales. Alors que les données basses fréquences (BF) continuent de livrer leurs secrets, la complexité des données HF (bases volumineuses, convexes, avec des données manquantes,. . .) rend leur exploitation difficile.
Dans ce contexte, cette thèse a pour objectif de développer un système numérique Open Source basé sur plusieurs méthodes du Machine Learning. Ce système doit permettre (i) de définir des schémas de fonctionnement des efflorescences multi-sources et multi-échelles à partir de données multivariées, (ii) de disposer d'un système de prédiction et d'alerte, (iii) de pouvoir adapter en temps (quasi) réel les stratégies d'échantillonnage pour les besoins de l'Observation, de la Surveillance et de la Recherche. La mise en évidence d'états environnementaux favorables ou pas aux efflorescences algales permet de hiérarchiser les facteurs de contrôle et d'identifier des schémas de fonctionnements. Ainsi, cette thèse a été conduite en différentes phases. Tout d'abord, une étude exploratoire des différentes variables, sources et échelles de données et des sites d'études a permis de définir les spécificités de chaque environnement et d'appliquer une stratégie d'étude multicritères. Ensuite, une nouvelle méthode de classification adaptée aux problématiques écologiques et HF est développée : La Classification Spectrale Multi-Niveaux. Cette approche appliquée aux données de la station MAREL-Carnot (IR ILICO, SNO COAST-HF) a permis la description (au niveau biogéochimique et taxonomique) d'événements récurrents mais aussi extrêmes, dont la période peut être infra-hebdomadaire ou même horaire. En plus de la caractérisation par état, cette approche multicritères identifie des schémas relationnels (états pressions et réponses). Ainsi, elle met en évidence des schémas de succession des facteurs d'influences, et des communautés. Enfin, une comparaison des réponses du modèle de classification à d'autres bases de données (MAREL-Iroise et MesuRho) est présentée. Pour aller plus loin, la méthodologie proposée est étendue à la prédiction d'événements météorologiques à partir de réanalyses (prévisions ERA5) et à un cas d'étude spatialisé (campagne océanographique CGFS). Elle démontre des classes cohérentes avec les expertises et s'ouvre ainsi vers une variété d'applications.

Mots clés : manche, atlantique, méditerranée, phytoplancton, nuisible, bouée instrumentée, machine learning, classification spectrale.

Ifremer
150 quai Gambetta – 62200 Boulogne-sur-mer – France

Multivariate and multi-scale definition of environmental states by Machine Learning: Characterization of phytoplankton dynamics

Abstract

Automatic observing high frequency (HF) systems should allow a better understanding of the environment dynamic (and phytoplankton) in response to natural and anthropogenic pressures, as well as direct and indirect effects of phytoplankton blooms, leading to ecosystem dysfunctions. Improving knowledges is important in order to propose indicators consistent with the aims of the European directives and regional seas conventions. While low frequency (LF) data are widely used, processing HF data is a difficult task, due to their complexity (heavy, non linear, missing, etc. . .).

In this context, the aim of this thesis is to propose an open source numeric system that should be able to (i) define multi-scale efflorescence functioning patterns from multivariate and multi-source data, (ii) have a forecasting and warning system, (iii) adapt in near real time the sampling strategies for the Observation, Monitoring and Research needs. Therefore, this thesis was divided in different steps. First of all, a study of the different available databases on each study sites allows defining a multi-criteria study strategy. Then, a new non-supervised clustering method, enabling environmental states labelling, was developed: the Multi-Level Spectral Clustering. The detection of environmental states allows the classification and comparison of blooms controlling factors and the identification of functioning patterns. This multi-criteria method, applied to data from the instrumented station MAREL-Carnot, leads to the biogeochemical and taxonomic description of recursive and extreme events. Moreover, it detects temporally structured relational patterns (pressures and responses) and highlights successive patterns of influencing factors, and communities. These states provide a learning base allowing the development of a learning model, through supervised classification methods, in order to analyze and predict these states. Thus, a semi-supervised approach, coupling a multi-level spectral clustering (unsupervised) and a classification method (supervised) is proposed. A comparison of the responses from the supervised classification model to other databases (MAREL-Iroise and MesuRho) is presented. Furthermore, this method is applied to the prediction of meteorological events, based on reanalyzes (ERA5 predictions) and a spatialized case study (CGFS oceanographic campaign). It demonstrates clusters consistent with the expertise and opens up to a variety of applications.

Keywords: english channel, atlantic, mediterranean sea, phytoplankton, harmful, high frequency, buoy, deep learning, spectral clustering.

Note à l'attention des lecteurs : Les pages 197 et 198 peuvent être imprimées individuellement et servir de marque-page et de support afin de faciliter la lecture du manuscrit en rappelant les principales notions.

Table des matières

Liste des tableaux

Table des figures

Introduction Générale

Les travaux présentés dans ce manuscrit s'intègrent dans un cadre pluridisciplinaire, écologie marine - traitement de signal et informatique - météorologie et pluri-institut. Le sujet de cette thèse est né d'une collaboration bi-partie Ifremer-LER Boulogne-sur-mer et ULCO-LISIC ancrée dans plusieurs projets INTERREG IVa 2 mers DYMAPHY 2009-2013 et JERICO-NEXT et CPER MARCO depuis 2014-2020, JERICO S3 depuis février 2020. Il a été étendu à la société WEATHERFORCE basée à Toulouse, spécialisée dans le développement de solutions et expertises météorologiques qui, après un entretien d'embauche en 2017 et leur avoir décrit le projet de thèse ont proposé de cofinancer ces travaux à travers un financement CIFRE. Les travaux présentés dans ce manuscrit ont donc pour objectif de proposer une méthodologie numérique pour aboutir à une définition multivariée et multiéchelle d'états environnementaux par des techniques d'apprentissage guidées par les données (*Machine Learning*). Les méthodologies et outils développés sont construits pour permettre la compréhension et le suivi de phénomènes tels que la dynamique phytoplanctonique en zone côtière ou encore la prédiction d'épisodes météorologiques comme les périodes de moussons,...

Contexte

La zone côtière : Zone à forte variabilité

Les zones côtières sont des zones socio-économiques très importantes. Elles sont le lieu de nombreuses activités (*e.g.* pêche, aquaculture, tourisme, développement des énergies marines renouvelables, agriculture, production industrielle) et jouent un rôle dans de nombreux processus environnementaux (cycle des nutriments, stockage de carbone, ...). Ainsi, bien que représentant seulement 7 % de la surface de l'océan, les écosystèmes côtiers constituent environ 14 à 30 % de la production primaire globale [Gattuso et al. 1998] et comptent donc parmi les zones les plus productives. Parallèlement, ses écosystèmes et ses habitats sont très vulnérables. De par sa position d'interface entre le continent et la zone hauturière océanique, la zone côtière est soumise à de nombreux forçages à fortes variabilités. Les écosystèmes côtiers sont en effet régis, d'une part, par des forçages "écologiques" tels les mécanismes physiques, chimiques et " biologiques ou naturels" comme la géomorphologie de la côte, les variations saisonnières, les changements atmosphériques [Wefer 2015]. D'autre part, ils subissent énormément de pressions anthropiques via des activités directes comme les activités portuaires, l'élimination des déchets, les modifications du littoral et via l'usage en amont des bassins versants par l'agriculture, l'industrie, l'urbanisation. Cette activité humaine, qui augmente d'année en année, cause des modifications et perturbe le milieu. Elle engendre de nombreux changements, à savoir des phénomènes d'eutrophisation, l'arrivée et la prolifération des espèces envahissantes/toxiques et des effets combinés sur les pêcheries locales, les pertes de biodiversité ou encore des changements phénologiques.

Ce milieu dynamique et complexe change beaucoup plus rapidement que ce qui était prévu

il y a une décennie [Cloern, Abreu et al. 2016], créant un défi de taille pour gérer ces zones de manière durable. Pour répondre à cet enjeu, des outils d'évaluation et de gestion de la qualité de l'environnement national et international comme la Directive Cadre sur l'Eau (DCE) [DCE 2000/60/CE], la Directive Cadre Stratégie pour le Milieu Marin (DCSMM) [DCSMM 2008/56/CE] et la convention d'Oslo et de Paris (OSPAR) [OSPAR 1992] ont été mis en place. Ces outils d'évaluation écologique visent tous à établir un cadre d'action pour la protection du milieu marin et/ou fluvial. Ces stratégies sont définies pour maintenir ou atteindre un « Bon État Écologique »(BEE). Ainsi, l'amélioration des connaissances du fonctionnement du système côtier sous influence anthropique et sous contrôle de divers facteurs environnementaux est un enjeux sanitaire et écologique important [Smayada 1990a].

Le phytoplancton : indicateur écologique

Le phytoplancton, principal producteur primaire, a une fonction essentielle pour la productivité et le cycle des sels nutritifs de la zone côtière. Constitué en majeure partie d'organismes unicellulaires autotrophes, il transforme la matière inorganique en matière organique via le processus de photosynthèse et joue un rôle de pompe biologique. Le dioxyde de carbone fixé par le phytoplancton lors de la photosynthèse est rendu disponible pour d'autres organismes. Lorsque les organismes périclitent, leurs squelettes externes ou enveloppes (*i.e.* frustules, thèques, coccolithes) chargées en carbone et autres matières inorganiques (comme le Silicium), coulent doucement pour atterrir au fond de l'océan. Ainsi, le phytoplancton est responsable du transfert vertical de carbone de la surface des océans vers les zones plus profondes [Reynolds 2006] et fonctionne comme une pompe biologique. De par cette caractéristique, il tient une place importante dans les cycles géochimiques des sels nutritifs ce qui le positionne à la base du réseau trophique.

De plus, les périodes productives sont marquées par une augmentation de la biomasse phytoplanctonique appelée une efflorescence phytoplanctonique (*bloom*). Les efflorescences phytoplanctoniques sont des événements rapides qui répondent à des changements liés à des forçages environnementaux et anthropiques. Ainsi sa place centrale dans l'écosystème et ses capacités de réponses rapides aux multiples changements de son milieu le définissent comme un indicateur à haute réactivité des variations de la qualité des eaux côtières.

Les efflorescences : processus spatio-temporels multi-échelles.

Les forçages liés aux changements environnementaux et aux perturbations anthropiques s'expriment à différentes échelles spatiales et temporelles. Par conséquent, les réponses du phytoplancton s'observent aussi à des échelles spatio-temporelles diverses. De nombreuses études ont montré que le phytoplancton, sa biomasse, sa composition spécifique et sa phénologie sont influencés par les conditions abiotiques du milieu à grandes et petites échelles. [Edwards et Richardson 2004] démontrent une relation de dépendance entre les variations de l'amplitude des efflorescences et le changement climatique. D'autres mettent en évidence des changements hydrologiques et physico-chimiques conditionnant des modifications de la distribution et de la composition du phytoplancton [Gailhard et al. 2002 ; Lefebvre, Guiselin et al. 2011] ou des changements de la phénologie de l'efflorescence [Breton, Rousseau et al. 2006 ; Racault et al. 2012].

Par ailleurs, il doit aussi être pris en considération la possibilité de transferts d'échelle. Dans le contexte de la dynamique phytoplanctonique, cela peut être représenté par la relation entre un impact à petite échelle tel que l'absorption des sels nutritifs et un phénomène à plus grande échelle tel que la succession saisonnière des communautés. Le schéma de Dickey 2003a (Figure 1) met en évidence les processus biologiques et physico-chimiques (ellipses) qui interagissent dans

le milieu marin à des échelles multiples et imbriquées. Il rend compte de la répercussion des phénomènes petites échelles tels que les processus moléculaires sur des phénomènes plus grandes échelles tels que les cycles saisonniers (molecular processes et seasonal cycles sur la figure 1). Dans le cas du phytoplancton, il permet d'aborder les phénomènes petites échelles (*e.g. l'absorption des sels nutritifs*) impactant la croissance du phytoplancton et qui eux-mêmes ont des répercussions sur des phénomènes méso-échelles (*e.g. successions d'assemblages phytoplanctoniques*). A chacun de ces processus peuvent être associées différentes stratégies d'échantillonnage. Les rectangles sur le schéma indiquent les échelles de temps et d'espaces approximatives pour une variété de plates-formes d'échantillonnage et d'observation [Karl et Church 2017].

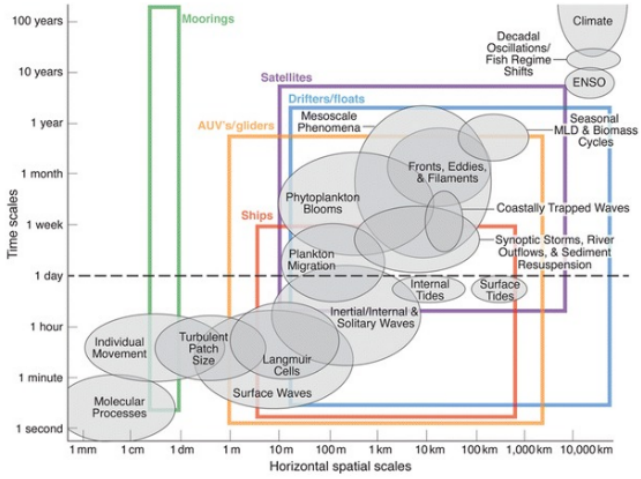


Figure 1 – Échelles spatiales et temporelles (Log/log) des différents processus impliqués dans les efflorescences du phytoplancton (d'après Dickey 2003a, tiré de Karl et Church 2017).

Stratégies d'échantillonnages : infra-horaires.

Le caractère variable du milieu côtier et multi-échelles des processus rend difficile la qualification et la quantification de la dynamique phytoplanctonique. Une des explications de la connaissance très générique de ces variations tient à la difficulté d'effectuer des observations régulières de cet environnement. Une des clefs pour appréhender cette problématique réside dans la stratégie d'échantillonnage (Figure 1). Le plus souvent, les programmes d'observation et de surveillance proposent une approche conventionnelle dite à Basses Fréquences (BF) c'est-à-dire avec des fréquences d'échantillonnages bimensuelles, mensuelles voire quelquefois hebdomadaires. Mais les approches dites à Hautes Fréquences (HF) se sont largement multipliées au cours des dernières décennies. Plusieurs groupements scientifiques portent un intérêt particulier à cette thématique et des initiatives à l'échelle nationale comme le consortium COAST-HF de l'Infrastructure de Recherche Littorale et Côtière (IR-ILICO) ou à l'échelle européenne comme la structuration de la communauté scientifique dans le cadre du projet H2020 JERICO-NEXT.

Pour aider à la compréhension de cet environnement dynamique, l'Institut français de recherche pour l'exploitation de la mer (Ifremer) développe depuis de nombreuses années des systèmes automatisés de mesures à HF ainsi que des méthodes numériques optimisées pour traiter ces données [Schmitt et Lefebvre 2016]. L'Ifremer est notamment doté d'un réseau de stations de mesures côtières multi-paramètres qui permettent un suivi bio-géochimique HF avec une acquisition de l'ordre de la dizaine de minutes (toutes les 20, 30 minutes) : le réseau Mesures

Automatisées en Réseau pour l'Environnement Littoral (MAREL) (Figure 1.5 Chapitre 2 Section 2.1.1). Ce réseau a été créé en complément des réseaux de mesures BF et fait désormais partie du COAST-HF de l'IR-ILICO. Ces stations mesurent les paramètres physico-chimiques et biologiques classiquement mesurés dans les systèmes aquatiques (détails Chapitre 2 section 2.1.1). Or le caractère HF et autonome de ce type de stations exige que les mesures soient infra-horaires voire infra-journalières et automatiques. Ainsi, les variations du phytoplancton sont mesurées, sur ces stations, à partir de la fluorescence, proxy de la biomasse phytoplanctonique. L'évolution de cette variable sera donc à la base des analyses de la dynamique phytoplanctonique de cette étude.

Même si la mise en place de ce type de réseau permet une vision plus précise de cet environnement variable et des phénomènes associés, les résultats ne sont pas toujours exploités de façon optimale. L'utilisation de ces données reste limitée dans la littérature académique, en raison de multiples contraintes associées à ces bases de données : forte variabilité, large gamme d'échelle pour la variabilité des séries temporelles, nombreuses données manquantes (liées aux maintenances et épisodes de pannes des capteurs), nombreux paramètres à prendre en compte obéissant à des forçages et processus différents, etc. En effet, cette quantité très importante de données n'est pas forcément facile à appréhender avec des méthodes d'investigations classiques. Par conséquent, les séries de données HF sont généralement dégradées pour revenir à des fréquences considérées comme interprétables. Cette dégradation engendre une perte importante d'informations, la valorisation scientifique de ces séries est par conséquent insuffisante par rapport à l'information offerte initialement.

Objectifs de l'étude

Ainsi les problématiques liées à ce contexte et aux contraintes de l'exploitation des données HF sont :

(i) de pouvoir appréhender ces nombreuses contraintes, en utilisant des méthodologies et des outils adaptés et normalisés, pouvant être appliqués à différents jeux de données HF ;

(ii) de définir des états environnementaux multicritères : globaux et spécifiques, aux variations régulières, épisodiques ou extrêmes ;

(iii) d'en étudier la dynamique au regard de l'existence de perturbations multi-échelles environnementales (apports de nutriments, tempêtes, ...) à différentes échelles de temps et dans des écosystèmes contrastés.

De ces problématiques (i, ii, iii) a émergé l'axe de recherche de cette étude qui est d'identifier des outils numériques et de mettre en place des protocoles performants de traitement de données afin de définir des schémas de fonctionnement des efflorescences (facteurs de contrôles, conditions d'initiation et de terminaison et de contrôles de l'amplitude) par le biais de l'étude d'écosystèmes contrastés. Pour y répondre, des objectifs méthodologiques et écologiques suivant ont été fixés :

1. Méthodologie

   M1 Développement d'un outil numérique Open Source pour le traitement et l'analyse de données HF (bouées instrumentées, Ferry Box,...).

   M2 Développement d'un système de définition d'états environnementaux multicritères (combinaison de paramètres physico-chimiques et biologiques) et de prédiction en réponse aux forçages naturels et anthropiques.

2. Écologie, observation et surveillance

E1 Définir des schémas de fonctionnement (facteurs de contrôle, conditions d'initiation et de terminaison, contrôle de l'amplitude) des efflorescences du phytoplancton dans des écosystèmes contrastés.

E2 Comparaison des schémas de fonctionnement pour une hiérarchisation de l'effet des perturbations.

E3 Compréhension du rôle respectif de chaque état environnemental récurrent ou extrême sur la caractérisation et la dynamique des événements (efflorescences, régénération des sels nutritifs, ...)

E4 Développement d'un système de prédictions et d'alertes face aux efflorescences d'algues potentiellement nuisibles.

Mise en place d'une stratégie d'étude

Ainsi nous nous sommes demandés : Quelles stratégies mettre en place pour développer des outils numériques de traitement et d'analyse des données HF (objectif M1) qui permettent de définir des états multicritères (objectif M2) ? Et comment définir, à partir de ces états multicritères, des schémas de fonctionnement des efflorescences du phytoplancton (Objectif E1) pour ensuite effectuer des comparaisons des facteurs d'influence (Objectif E2) et essayer de comprendre le rôle des événements sur la dynamique et la structuration du phytoplancton dans l'écosystème (Objectif E3).

Approche Multicritère

Premièrement, la réponse à ces questions passe initialement par la mise en place d'une base de données pertinente. En effet, il est difficile de répondre précisément à la question : "Comment évolue la dynamique environnementale en fonction des forçages locaux, régionaux et globaux des différents écosystèmes pélagiques ?", sans une base de données complète et cohérente. Nous avons donc choisi de sélectionner une liste de variables explicatives facile d'accès (c'est-à-dire couramment mesurées et en téléchargement libre) mais de manière la plus exhaustive possible. De plus, pour "comparer les facteurs d'influences et leurs rôles", nous avons décidé d'évaluer les résultats de notre méthode sur les jeux de données de plusieurs sites d'études qui correspondent à des écosystèmes contrastés. Cette démarche permettra de voir les réponses de la méthode lorsque les conditions environnementale changent. Une approche multicritère (a),(b),(c) a donc été construite dans cette optique. Une bases de données a donc été composée sur la base des critères :

(a) Multi-paramètres via l'identification et la sélection des paramètres d'intérêts, ceux les plus utilisés dans la littérature ;
(b) Multi-échelles et multi-sources avec une intégration de données basse et haute fréquence provenant de nouvelles sources de données *in-situ*, météorologiques, satellites et de modèles ;
(c) Multi-sites avec des observations issues de trois zones géographiques enregistrées sur le réseau MAREL.

Approche Multi-paramètres / Multivariées

Comme exposé dans la section 3, l'écosystème côtier et notamment la dynamique du phytoplancton est régie par des processus inter-connectés, multifactoriels et multi-échelles. Afin de prendre en compte cet aspect dans notre étude, il était essentiel que la base de données intègre un ensemble de variables cohérentes et adaptées.

La mise en place d'une base de données multivariées complète et standardisée nécessite l'identification de variables d'intérêts. Pour ce faire, nous avons utilisé le concept de Variables Océanographiques Essentielles (EOVs) : le groupe de variables indispensables pour comprendre et diagnostiquer la "santé" de l'Océan et les changements dans l'environnement marin.

Inspiré par l'impact positif qu'a eu, sur la science du climat, la définition des Variables Climatiques Essentielles (*Essential Climate Variables* (ECVs)), le Cadre d'observation de l'océan (*Framework for Ocean Observing* (FOO), [Lindstrom et al. 2012]) a suggéré d'organiser les activités de surveillance des océans autour des EOVs définies par un panel d'experts.

Plusieurs études généralistes telles que [Miloslavich et al. 2018 ; Kissling et al. 2018 ; Bojinski et al. 2014 ; Magdalena 2018] font le bilan des différentes variables EOVs et/ou du minimum requis pour les études du fonctionnement de chaque compartiment des écosystèmes marins. De la liste des EOVs du compartiment phytoplanctonique reprise dans le Tableau 1), nous avons extrait les variables d'intérêts disponibles, écrites en rouge.

Tableau 1 – Liste des EOVs relatives aux études sur la diversité et de la biomasse du phytoplancton établie à partir des recommandation dans [Lindstrom et al. 2012] ; En rouge les paramètres pris en compte dans notre étude.

Biomasse Phytoplanctonique		
Variables directes	**Variables supports**	**Variables relatives**
- **Biodiversité (Distribution : présence ou absence, Abondance, Densité et Abondance relative)** - **Diversité-Taxonomie** - **Fluorescence**, Concentration cellulaire - Concentration pigmentaire par spectrophotometrie (Chlorophylle a, Phaeopigment)	- **Nutriment (Silicate, Phosphate, Nitrate)** - **Température, Salinité, Oxygène, Oxygène saturé, Carbone organique dissout** - Matière organique particulaire, **Matière en suspension, Turbidité** - **pH** - **Variables bio-optiques (reflectance, couleur de l'eau-télédétection**, coefficient d'absorption)	- Biodiversité (Production primaire, biomasse de zooplancton, aire de répartition géographique) - Espèces introduites (présence, fréquence d'apparition ...) - **Variables atmosphériques (Température de l'air, vitesse et direction du vent, pression, précipitation)** - Variables terrestres (Débit et rejets des rivières, utilisation de l'eau, eaux souterraines) - Activités anthropiques (Pêche, Pollution par les hydrocarbures, pollution chimique, **Dragage, ouverture de barrages**, trafic maritime, fermes aquacoles ...)

Approche Multi-échelles/ Multi-sources

Cette approche intégrée multi-paramètres nécessite un cadre interdiciplinaire. En effet, pour rassembler l'ensemble des paramètres (Tableau 1) les sources de données ont été multipliées. La plateforme Coriolis a été choisie comme source principale de données océanographiques *in situ* HF parce qu'elle centralise, depuis 2014, toutes les stations du réseau MAREL. À ce socle

de données biogéochimiques HF, ont été ajoutées des données BF (campagnes de mesures) qui seront prises en compte comme nouvelles sources de données. Ces nouvelles bases de données vont permettre l'ajout de variables supplémentaires, telles que la diversité taxonomique, et des variables complémentaires, telles que les concentrations en sels nutritifs, qui serviront d'information de contrôle.

Ainsi, la mise en place de cette stratégie multi-paramètres et multi-sources va permettre l'intégration d'une grande partie des paramètres à HF ou BF de cette liste. Les paramètres en **rouge/gras** correspondent aux paramètres qui seront utilisés lors de cette étude.

Approche Multi-sites

Un même facteur de pression peut engendrer des réponses très différentes en fonction de la zone d'étude et de sa dynamique propre (environnement homogène ou stratifié, eutrophe ou oligotrophe). L'étude et la comparaison des schémas de fonctionnement des efflorescences nécessitent d'approfondir ces recherches sur plusieurs écosystèmes. Afin d'identifier et de hiérarchiser les potentiels schémas fonctionnels dans des écosystèmes contrastés, trois sites de mesures à HF, MAREL-Carnot, MAREL-Iroise et MesuRho situés respectivement en Manche, en mer d'Iroise et en mer Méditerranée, seront étudiés. Ils assurent ainsi la représentation des trois façades maritimes de l'hexagone français et vont permettre une inter-comparaison des schémas qui seront mis en évidence. Ainsi, les réponses d'un environnement homogène ou stratifié, eutrophe ou oligotrophe ... pourront être observées (Figure 2). Par ailleurs, ces trois stations sont positionnées dans des sites d'étude aux caractéristiques communes. Par exemple, elles sont toutes des zones côtières à proximité d'un affluent, dans des zones avec une forte activité anthropique. Ces points communs offrent l'avantage d'avoir le même type de facteur pression pour chaque site d'étude.

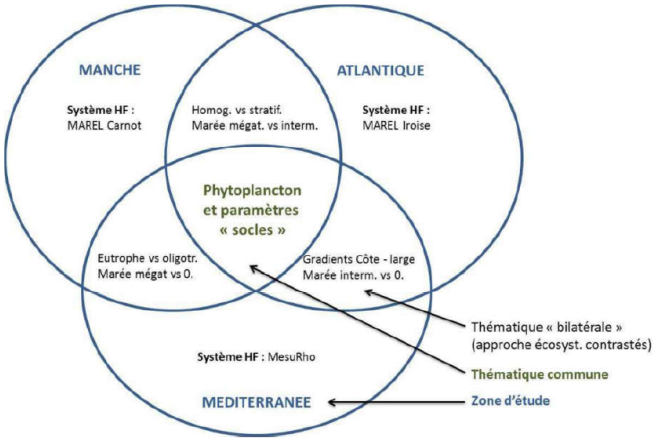


FIGURE 2 – Stratégies d'étude de la dynamique phytoplanctonique à Hautes Fréquences (HF).

Méthodologie

Ensuite, caractériser et prédire la dynamique des efflorescences du phytoplancton implique, d'une part, d'identifier des états environnementaux et, d'autre part, d'apprendre la dynamique interne des états. Or nous n'avons aucune "connaissance *a priori*" sur ces états. Dans notre cas, cela signifie qu'il n'y a pas de classification connue (sous forme par exemple d'étiquettes (*label*)), ni d'informations sur leurs nombres, leurs formes ou leurs distributions. Il est donc nécessaire,

dans un premier temps d'employer des méthodes capables de séparer les données en différents groupes sans aucune connaissance *a priori* pour déterminer des schémas de fonctionnement. Puis il convient de choisir une méthode de prédiction performante pour apprendre ces schémas et prédire un état pour de nouvelles observations ou nouveaux jeux de données. Nous avons donc opté pour une approche numérique en plein essor : l'apprentissage artificiel (*Machine Learning*) pour répondre à nos problématiques numériques et environnementales.

De manière générale, la notion d'apprentissage artificiel (*Machine Learning*) englobe toutes méthodes permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle pré-existant, partiel ou moins général (*Transfert Learning*), soit en créant complètement le modèle.

On différencie trois types de méthodes :

le classement ou classification supervisée, en anglais *Classification*, désigne le regroupement des classes à partir de leurs descripteurs en fonction des informations sur les données (de leurs étiquettes/*labels*).

la classification non supervisée, en anglais *Clustering*, consiste, *a contrario*, à identifier des groupes d'objets (classes/*clusters*) tels que les objets (ici les observations) les plus similaires soient dans le même groupe. Ces groupes appelés par abus de langage aussi classes sont souvent organisées en structures. Si cette structure est un arbre alors on parle de taxinomie (*taxonomy*).

La régression, en anglais *Regression*, traite, quant à elle, des cas où les valeurs à prédire sont numériques à partir d'un ensemble de données uni- ou multivariées.

Ces différents types de classification introduisent deux notions essentielles de l'apprentissage : l'apprentissage non supervisé et l'apprentissage supervisé [Cornuéjols et Miclet 2010].

L'apprentissage non supervisé se construit sur des exemples non étiquetés (sans *labels*). À partir d'une base d'apprentissage $X = (x_i)_{1 \leq i \leq N}$, les données vont être associées au même groupe classe en fonction des régularités trouvées. Cette recherche de régularité/similarité est obtenue à partir d'hypothèses sous forme, par exemple, de fonctions (régression), de nuages de points (mixture de gaussiennes) ou de modèles complexes (réseau bayésien).

L'apprentissage supervisé se construit sur des exemples étiquetés (*label*) "connus" ; c'est-à-dire que l'étiquetage (labellisation) des données a été réalisé par un oracle ou expert. À partir de la base d'apprentissage $X = (x_i; l_k)_{1 \leq i,k \leq N}$ avec ($X = echantillons$ et $l = Etiquettes$), une loi de dépendance entre x et l est recherchée. Par exemple, une fonction h aussi proche de f (fonction cible) que possible tel que $l_k = f(x_i)$, avec une distribution de probabilité $P(l_k|x_i)$

En écologie numérique, les méthodes de classification, supervisées ou non, sont souvent utilisées pour pour synthétiser l'information, comprendre la structure des données et en extraire le maximum de renseignements afin d'améliorer les connaissances et, émettre des recommandations en matière de gestion de l'environnement. Dans notre cas d'étude, nous n'avons pas de connaissances *a priori* (pas de label) sur le jeu de données. Notre choix s'est donc orienté vers les méthodes non supervisées.

Dans les travaux de recherche issus d'une précédente thèse [Rousseeuw et al. 2015a] une approche d'apprentissage non supervisée a été proposé, sans utilisation de connaissances environnementales *a priori*, pour construire un système numérique de détection d'états environnementaux par classification spectrale de données HF, couplée à une Modelisation Markovienne Caché (MMC). Ce système MMC permet de définir la dynamique d'états environnementaux multicritères caractéristiques de différentes phases des efflorescences phytoplanctoniques en réponse aux modifications environnementales et/ou des modifications du milieu en réponse à ces efflorescences.

Dans cette étude, une des étapes de travail a consisté (1) à étendre l'approche précédente de caractérisation directe à une classification non supervisée multiéchelle et (2) à apporter une labellisation environnementale des états obtenus et construire un modèle de prédiction. L'idée est donc de coupler une phase d'apprentissage supervisée avec une étape antérieure de classification non-supervisée (Schéma 3). Ainsi, notre stratégie d'étude va se diviser en deux phases :

- une phase de classification non supervisée (classes c_i) suivi d'une étape de labellisation (labels l_i). Elle a pour but d'identifier des groupes (classes) et de définir pour chaque classe un label à l'aide de l'avis d'experts (hiérarchisation des pressions, formulation d'hypothèses de réponses aux perturbations, ...) et de données complémentaires HF comme BF (ex., composition de la communauté phytoplanctonique).
- une phase d'apprentissage supervisé et de prédiction des données pourra être réalisée une fois les labels définies. Cette étape permettra l'analyse et la prédiction d'états environnementaux en temps *quasi* réels ou futurs.

Ces connaissances définies lors de la phase de labellisation seront alors utilisées comme informations clefs dans la phase d'apprentissage. Une attention particulière est donc apportée sur ce point. Ainsi, il est mis en place lors de ce travail de thèse une approche multi-niveaux, soit une approche profonde avec plusieurs niveaux qui couple plusieurs méthodes de *clustering*. Cette approche a pour objectif de mieux définir les limites des états et d'approfondir la détermination de chacun (états plus petits et plus spécifiques). Cette étape permettra d'améliorer l'expertise et donc la qualité des informations données en entrées de la phase d'apprentissage.

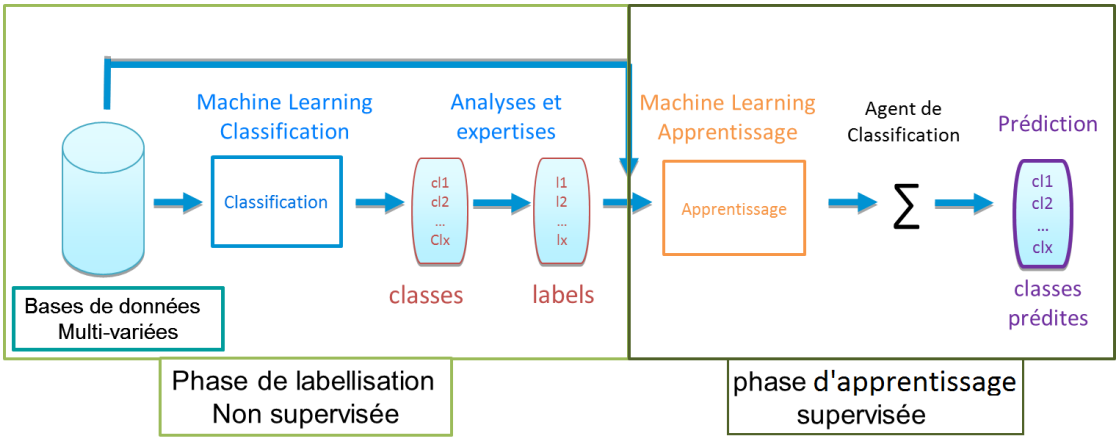


FIGURE 3 – Approche semi-supervisée.

D'après les objectifs et les réflexions exposées, précédemment, cette thèse se construit de la manière suivante :

Le **Chapitre 1** est une partie introductive qui présente le contexte de cette étude liée aux compartiments hydrologiques, phytoplanctoniques et aux zones d'études. Une première section présente donc les caractéristiques écologiques du compartiment phytoplanctonique. Il détaille les caractéristiques générales du phytoplancton et les relations entre les schémas de croissance et son environnement (forçages hydroclimatiques et physico-chimiques). Ensuite, les différentes caractéristiques hydrodynamiques, physico-chimiques et la distribution phytoplanctonique de chaque zone d'études sont détaillées dans une autre section.

Le **Chapitre 2** décrit l'acquisition des données et les méthodes d'analyses existantes utilisées pour la construction de notre approche de segmentation multi-niveaux. Dans une première section, les réseaux de surveillance, leurs systèmes instrumentés et leurs stratégies d'échantillonnage sont présentés. Dans une seconde section, les différentes méthodes d'analyses descriptives des données utilisées sont exposées. Dans une troisième section, le principe et les spécificités de plusieurs méthodes usuelles d'apprentissage non supervisé sont détaillées. Chaque méthode présentée a été utilisée directement ou comme point de comparaison avec les méthodes développées lors de cette thèse.

Dans le **Chapitre 3**, une nouvelle approche non-supervisée, la classification spectrale multi-niveaux, est proposée. Cette classification spectrale composée d'une architecture profonde vise à améliorer la définition et la compréhension des variations dans les séries temporelles multivariées. Ce développement méthodologique a pour but, dans le cadre de cette étude, de faciliter la caractérisation de la dynamique phytoplanctonique. Tout d'abord la méthode ainsi que les tests de validation de celle-ci seront présentés. Une fois validée, la méthode a été utilisée de manière automatique sur les données issues de la station MAREL-Carnot (considéré ici comme le cas de référence). Le protocole de pré-traitement et les sorties de la classification sont donc présentés. Cette classification a permis une première phase de labellisation des classes en états écologiques. Les résultats ont, ensuite, été confrontés à des données taxonomiques BF et une relation étroite entre les assemblages phytoplanctoniques et les états environnementaux a été mise en évidence.

Dans le **Chapitre 4** est proposée une approche semi-supervisée couplant apprentissage (supervisé) et classification spectrale multi-niveau (non supervisé). Cette combinaison apprentissage et de classification spectrale a pour but l'analyse et la prédiction d'états. La première phase de classification (Chapitre 3) a permis la construction d'une base de données labellisée. Ce jeu de données est utilisé comme base d'apprentissage. Le modèle crée à partir des méthodes d'apprentissage est ensuite appliqué sur les autres jeux de données (MAREL-Iroise et MesuRho). Ainsi, les résultats permettent de valider l'approche semi-supervisée.

Le **Chapitre 5** correspond à l'application de l'approche de classification spectrale sur d'autres jeux de données. Il est démontré, dans une première partie, l'efficacité de la méthode dans un contexte non pas temporel mais spatial. Ainsi la méthode a été appliquée à des données issues d'un système type FerryBox qui présente la particularité d'être couplé à un fluorimètre spectral (permettant la détection des classes d'algues). Les résultats ont ensuite été reliés à plusieurs éco-régions telles que les paysage marins établis par la DCSMM ou encore la répartition spatiale de communauté de poissons. La seconde partie de ce chapitre s'intéresse aux données HF issues d'un modèle météorologique : les données de réanalyses du modèle ERA5. L'objectif était l'identification et la caractérisation des événements pluviométriques dans des régions touchées par des événements extrêmes (fortes pluies, le début des moussons) et où les informations *in-situ* sont peu présentes.

Chapitre 1

Contexte environnemental : Introduction des réponses du phytoplancton aux changements environnementaux

1.1 Introduction

L'objet de nos travaux est de construire un outil permettant d'améliorer nos connaissances sur la dynamique du phytoplancton dans les écosystèmes côtiers. La zone côtière est un milieu complexe, à forte variabilité, soumis à de nombreux forçages écologiques (*e.g.* mécanismes liés au milieu), naturels (*e.g.* variations saisonnières, changements climatiques) et anthropiques (*e.g.* l'industrie, les activités portuaires). Les réponses du phytoplancton à ces variations s'appliquent à tous les niveaux : de l'individu (*e.g.* physiologie, taille des cellules) jusqu'au niveau des communautés (*e.g.* biomasse, structure, phénologie). Ces variations concernent aussi bien la biomasse globale [Boyce et al. 2010], la diversité spécifique [Hernández et al. 2014] ou fonctionnelle [David et al. 2012] ou encore leur phénologie [Caillault-Poisson et Lefebvre 2017].

Ce chapitre a pour but de définir les différents facteurs qui interagissent et jouent un rôle de contrôle quant à l'évolution du phytoplancton. Une première partie décrit l'écologie du phytoplancton : ses caractéristiques générales, le rôle des facteurs abiotiques. Afin de caractériser les réponses du phytoplancton à ces différents facteurs, la dynamique des efflorescences est étudiée sur trois sites de mesures à HF, MAREL-Carnot, MAREL-Iroise et MesuRho situés respectivement en Manche, en mer d'Iroise et en mer Méditerranée. La seconde partie expose pour chaque zone d'étude sa géographie, son hydrodynamique et ses caractéristiques physico-chimiques. Un accent est mis sur les taxons dominants dans chaque zone et les particularités (le ou les points clefs) propres à chaque zone.

1.2 Caractéristiques écologiques du phytoplancton

1.2.1 Caractéristiques générales

Le phytoplancton constitue l'ensemble des organismes végétaux, présents dans le milieu pélagique, entraînés passivement par les mouvements d'eau dans le plan horizontal. Ce sont des

cellules microscopiques rarement visibles à l'œil nu. Elles sont présentes dans le milieu sous une multitude de formes (Sphériques, ellipsoïdes, prismes triangulaires ...), de tailles (2 μm à plusieurs centimètres) et peuvent être solitaires ou sous formes coloniales.

En 1995, Tett et Barton estiment le nombre d'espèces décrites à plus de 5000 [Tett et Barton 1995]. Aujourd'hui, la diversité du phytoplancton s'exprime au travers de plusieurs millions d'espèces. On retrouve dans ce groupe des organismes eucaryotes, et des organismes procaryotes (Figure 1.1). Les cyanobactéries, qui ne présentent pas d'organites cytoplasmiques, représentent la majeure partie des organismes photosynthétiques procaryotes. Cette caractéristique a été un facteur déterminant dans l'évolution des organismes photosynthétiques. En effet, l'acquisition de la photosynthèse chez les algues eucaryotes est basée sur des processus d'endosymbioses impliquant les cyanobactéries et des protistes eucaryotes [Reynolds 2006]. Ensuite, les groupes majoritaires d'Eucaryotes sont les diatomées (Classe *Bacillariophyceae*), les dinoflagellés (Classe *Dinophyceae*) et prymnésiophycées (Classe *Prymnesiophyceae*) qui représentent respectivement 40, 40, et 10 % des espèces phytoplanctoniques eucaryotes décrites [Simon et al. 2009].

Super-règne	Règne (Sous-règne)	Division	*Classe* (**nom usuel**)
PROKARYA (procaryotes)	Bactéries (Eubactéries)	Cyanobactéries	**cyanobactéries** (ou cyanophycées)
EUKARYA (eucaryotes)	Protoctistes (Protistes)	Dinophyta	*Dinophyceae* (**dinoflagellés**)
		Heterokontophyta	*Chrysophyceae* (**chrysophycées**)
			Bacillariophyceae = *Diatomophyceae* (**diatomées**)
			Dictyochophyceae (**silicoflagellés**)
			Raphidophyceae (**chloromonadines**)
		Prymnesiophyta = Haptophyta	*Prymnesiophyceae* = *Haptophyceae* (**prymnésiophycées** ou **haptophycées**)
		Cryptophyta	*Cryptophyceae* (**cryptophycées**)
		Chlorophyta	*Chlorophyceae* (**chlorophycées**)
			Prasinophyceae (**prasinophycées**)

Figure 1.1 – Classification simplifiée des organismes phytoplanctoniques (d'après Margulis et Schwartz 1998 et Hoek et al. 1995).

L'une des caractéristiques majeures du phytoplancton est qu'il joue un rôle primordial dans le cycle du carbone [Longhurst et Glen Harrison 1989]. Il est responsable d'environ 50 % de la production primaire mondiale. Il est capable, en effet, de synthétiser de la matière inorganique via la photosynthèse. La matière produite est ensuite rapidement transférée vers les échelons supérieurs. Et par conséquent, il est à la base des réseaux trophiques marins.

Ces périodes productives sont caractérisées par une augmentation de la biomasse phytoplanctonique appelée une efflorescence (bloom) phytoplanctonique. Ces efflorescences, bien qu'essentielles pour les écosystèmes marins peuvent aussi avoir des conséquences négatives. Quelques organismes produisent des toxines (*e.g.* neurologiques, paralysantes, diarrhéiques) nocives pour la faune

marine ou pour l'homme après consommation de crustacés. Une accumulation trop importante de biomasse peut modifier les caractéristiques iogéochimiques et provoquer par exemple des phénomènes d'anoxie.

Ainsi, les variations de son abondance, de sa phénologie ou de sa composition spécifique risquent de modifier profondément le fonctionnement des écosystèmes. C'est pourquoi la compréhension, la quantification et la qualification sa diversité mais aussi de son abondance et sa dynamique sont essentielles [Garmendia et al. 2013].

1.2.2 Rôles des facteurs abiotiques sur le cycle de vie

La dynamique du phytoplancton décrit historiquement par Margalef 1978 est dépendante des facteurs de contrôles hydroclimatiques (vent, courant, turbulence, ...), physicochimiques (concentration en nutriments, luminosité, température, turbidité) et biologiques (taille, flagelle, frustule ...). Tous ces facteurs vont affecter de manière directe ou indirecte les éléments nécessaires à la croissance du phytoplancton et donc générer une variation de la biomasse plus ou moins importante.

Les apports en sels nutritifs : L'un des principaux facteurs limitants est la concentration en nutriments. En effet, le phytoplancton a besoin d'éléments chimiques, tels que du carbone (C), des composés azotés dissous (N : nitrates, nitrites, ammonium), du silicate (Si) ou encore du phosphate (P), en abondance suffisante pour se développer et réaliser la photosynthèse. Redfield (1958) définit le ratio des composés océaniques Carbone/Azote/Phosphate ($C/N/P = 106 : 16 : 1$) nécessaire à la photosynthèse. En période de forte croissance, le stock de nutriments est rapidement consommé jusqu'à atteindre des taux inférieurs aux rapports de Redfield. À ce moment, la croissance du phytoplancton devient limitée par les concentrations en nutriments. Ainsi, cette limitation de la biomasse va permettre le renouvellement progressif des stocks de nutriments. Ce renouvellement est assuré par divers processus : apports éoliens et fluviaux (au niveau des estuaires), le taux de précipitations et les processus de lessivages, par des processus de régénération (activité bactérienne) ou grâce à la fixation d'azote (cyanobactéries). Toutefois, il existe parmi les groupes phytoplanctoniques différentes stratégies de nutrition. Bien que majoritairement autotrophes via la photosynthèse, certaines espèces ont développé des capacités hétérotrophes (capture de particules ou de proies). Ainsi, ces spécificités taxonomiques vont influencer l'organisation des communautés au sein de la masse d'eau [Cloern et Dufford 2005]. Par exemple, la mixotrophie est un avantage quand la disponibilité en sels nutritifs est limitée.

La régulation par la température et la lumière : Quand les populations ne sont pas limitées en éléments nutritifs, la croissance est dépendante des propriétés liées à la croissance cellulaire soit, principalement, les conditions de température et d'éclairement du milieu. La température agit directement sur le métabolisme des organismes. Chaque température optimale et limite de tolérance est spécifique à l'espèce. Toutefois, la majorité des espèces phytoplanctoniques sont eurythermes et sont donc tolérantes aux grandes variations de température. Eppley 1972 a défini l'influence de la température sur la croissance spécifique comme une augmentation progressive et exponentielle de la croissance avec une température allant jusqu'à environ 40 degrés. La température va, aussi, agir sur le niveau de stratification du milieu et sur la profondeur de la couche de mélange. Cette couche plus ou moins profonde (en fonction de la thermocline) constitue une épaisseur d'eau maximum dans laquelle le phytoplancton peut évoluer. Elle définit donc les conditions environnementales et nutritives pour le phytoplancton.

L'apport d'énergie lumineuse est, quant à lui, indispensable à la réalisation de la photosynthèse et donc indispensable pour la croissance et la reproduction du phytoplancton. Pour cela, seulement

une partie du rayonnement lumineux est utilisée : le *Photosynthetically active radiation* (*PAR*) soit les radiations comprises entre 400 et 700 nm. L'intensité lumineuse varie en fonction du taux d'absorption et de diffusion de la masse d'eau. Elle diminue donc avec la profondeur. La profondeur de la zone euphotique sera ainsi dépendante de la composition de l'eau. Le but principal du phytoplancton sera donc de maximiser ses possibilités de suspension dans cette zone euphotique [Reynolds 2006]. Chaque espèce va présenter des adaptations photosynthétiques (ex : nombre de chloroplastes, vacuoles, ou d'organes pour augmenter leurs surfaces) en fonction de l'environnement lumineux dans lequel elle se trouve. Certains organismes ont une efficacité hotosynthétique (rendement photosynthétique) plus importante ce qui leur permet l'utilisation de l'intensité lumineuse quand celle-ci est faible. Au contraire, certaines espèces ont un potentiel photosynthétique (potentiel métabolique) plus fort mais une efficacité hotosynthétique plus faible. Ces adaptations étant plus ou moins modulables en fonction des organismes.

Le rôle de la turbulence : Un autre facteur qui joue lui aussi un rôle primordial au niveau métabolique et sur la stratification de la colonne d'eau est la turbulence. Elle va jouer un rôle indirect par le biais des autres facteurs présentés ci-dessus. Premièrement, le taux de mélange turbulent va déterminer le temps de résidence et la profondeur d'un organisme dans la masse d'eau. Le phytoplancton va subir des variations de niveau ce qui induit des variations d'intensité lumineuse. Ainsi, le potentiel photosynthétique d'un organisme sera dépendant du rapport entre la vitesse de diffusion turbulente et du temps de photo-adaptation des cellules [Lewis et al. 1984]. Deuxièmement, la turbulence agit sur la profondeur de la couche de mélange. Ainsi, l'augmentation de la turbulence a pour effet de mélanger les masses d'eau, ce qui a pour conséquence d'augmenter l'épaisseur de la couche de mélange et donc l'importation des éléments nutritifs dans la zone euphotique. L'importance respective de ces facteurs a été mise en évidence par Sverdrup 1953 qui les a définis dans son modèle au travers de la profondeur critique. Il détermine une valeur critique de la profondeur de la couche de mélange pour laquelle le phytoplancton peut se développer. La croissance du phytoplancton est donc observée lorsque la profondeur de mélange est inférieure à la profondeur critique.

En zone côtière, les faibles profondeurs vont accroître l'impact de la turbulence (générée par les vents et/ou le régime de marée). Ainsi, l'épaisseur de la couche de mélange va parfois couvrir la totalité de la colonne d'eau. La dynamique phytoplanctonique peut toujours être associée au modèle de Sverdrup 1953 avec cependant quelques nuances. Généralement, la profondeur critique se rapproche ou dépasse la profondeur de l'écosystème côtier ce qui favorise le développement de la biomasse phytoplanctonique sur les plateaux continentaux. De plus, le milieu côtier est sous l'influence plus prononcée des apports d'eau douce et des ondes de marées. Les ondes de marées de fortes amplitudes vont rajouter une variabilité temporelle par exemple semi-diurne (Basse mer / Pleine mer) ou semi-mensuelle (vive-eau / morte-eau). Enfin, les arrivées d'eau douce dans l'écosystème vont avoir deux impacts fondamentaux : (i) l'enrichissement du milieu par le biais d'apports en sels nutritifs et en matières organiques et (ii) une augmentation de la stratification et de la turbidité.

Ainsi, ces facteurs environnementaux seront donc des phénomènes structurants pour la biomasse phytoplanctonique. Ils auront un impact sur la dynamique à différentes échelles de temps et d'espace. Les réponses observées aux échelles hebdomadaires ou infra-hebdomadaires étant déterminées plus spécifiquement par les taxons présents et à leurs capacités métaboliques.

1.2.3 Distribution temporelle

Les forçages hydroclimatiques et physicochimiques ne sont pas des phénomènes stables. Cette instabilité se répercute sur la distribution spatiale et temporelle du phytoplancton. De plus, un transfert d'échelles peut se créer entre un forçage de petites échelles et un phénomène dynamique plus grand. Dickey 2003b indique l'importance des échelles spatiales et temporelles en océanographie en illustrant l'imbrication des processus physiques et biologiques en fonction de ces échelles (Figure 1.2).

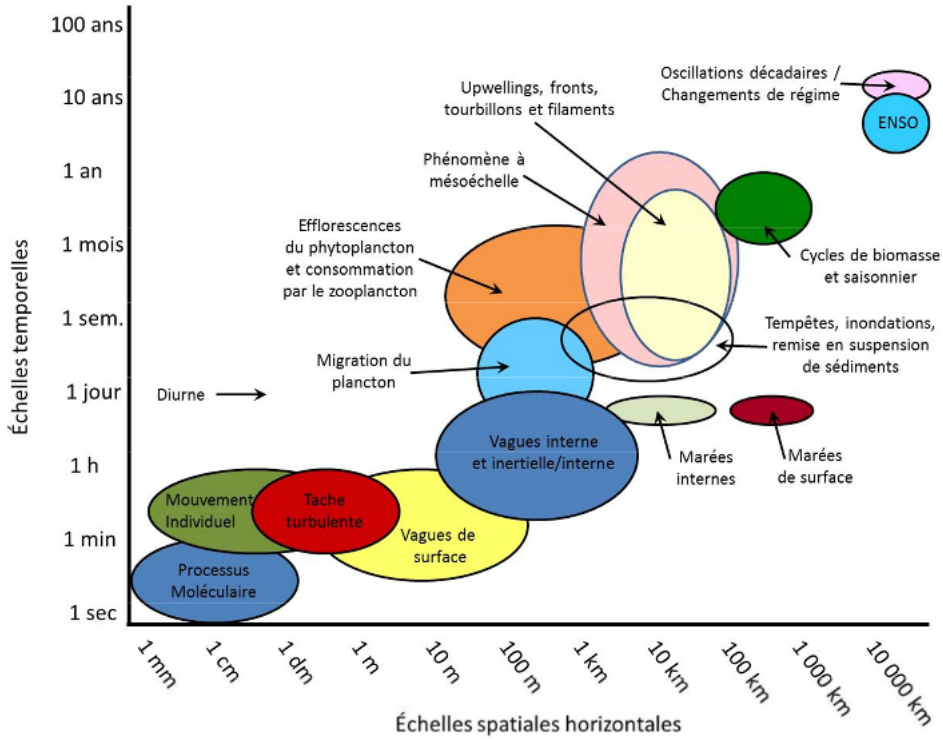


Figure 1.2 – Imbrication des échelles spatiales et temporelles (log/log) des différents processus impliqués dans les efflorescences du phytoplancton (Traduction Lefebvre A., d'après Dickey 2003b).

Ainsi, parallèlement à tous les facteurs exprimés précédemment, plusieurs dynamiques de croissance peuvent être caractérisées suivant différents types de variabilités périodiques [Cloern 1996] (Figure 1.3).

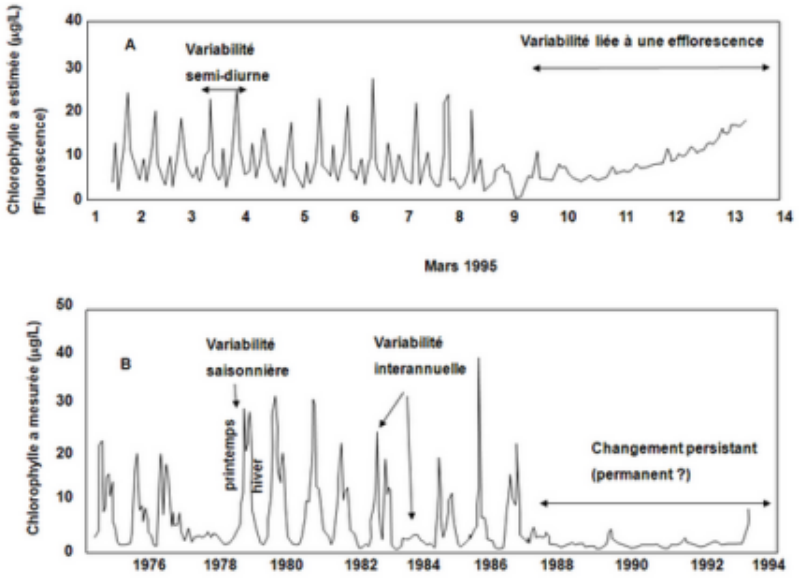


FIGURE 1.3 – Illustration des échelles temporelles impliquées dans les variations de la biomasse phytoplanctonique, estimées à partir de concentrations en chlorophylle a ($\mu g\, l^{-1}$), dans la baie de San Fransisco, avec en a), les variabilités rencontrées au cours d'un demi-mois et en b), les variabilités rencontrées sur plusieurs années. D'après CLOERN 1996.

Premièrement, une variabilité pluriannuelle liée à des ruptures majeures comme des apports anthropiques de sels nutritifs ou encore le réchauffement des masses d'eau est définie. Sur la figure, on note un pic étroit pour l'année 1986, des pics étalés de biomasse cumulée plus élevés pour 1982 ou inférieurs pour 1985, et des années avec des cumuls très bas notamment de 1987 à 1994. Plusieurs études ont mis en évidence, pour cette dernière décennie, des changements de communautés liées à ces ruptures [PHILIPPART et al. 2000 ; BRETON, ROUSSEAU et al. 2006 ; GIESKES et al. 2007 ; GOMEZ et SOUISSI 2008 ; LEFEBVRE, GUISELIN et al. 2011].

Deuxièmement le phytoplancton présente une variabilité saisonnière [CLOERN 1996, WINDER et CLOERN 2010]. Ces variations saisonnières diffèrent en fonction des écosystèmes. CUSHING 1996 décrit trois types de schémas de variations saisonnières : tropicales, polaires et tempérées (Figure 1.4).

Dans les régions tropicales sont observées une croissance et une production stable. Les niveaux de biomasse restant plus ou moins équilibrés. Au niveau des pôles, est décrite une efflorescence annuelle qui coïncide avec la disparition de la glace et l'augmentation de l'irradiance.

Dans les régions océaniques tempérées, comme c'est le cas dans notre étude, deux efflorescences annuelles sont décrites : une efflorescence printanière et une automnale. Cette dynamique est rythmée par les variations des différents facteurs abiotiques (cf. section 1.2.2). Au printemps des conditions optimales de croissance entraînent une première efflorescence, généralement dominée par des diatomées. En été, la biomasse diminue et les dinoflagellés, mieux adaptées à de faibles concentrations en sels nutritifs sont souvent dominantes à cette période. En automne, des apports secondaires en sels nutritifs favorisent une deuxième efflorescence.

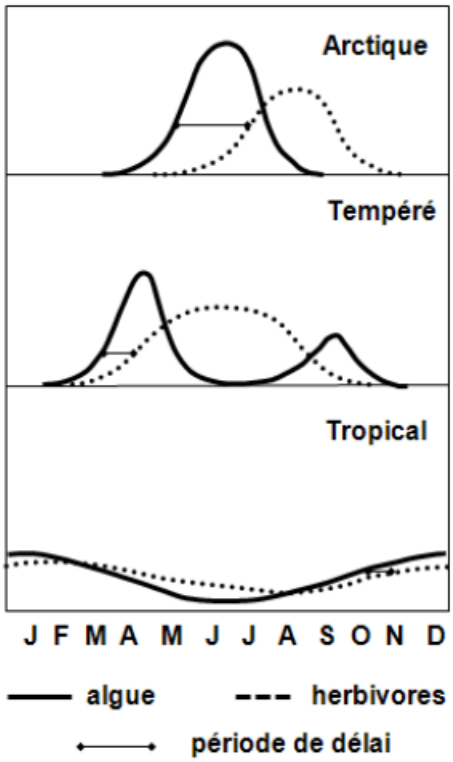


FIGURE 1.4 – Représentation des trois types de cycles saisonniers du phytoplancton et des herbivores en fonction de leur latitudes d'après CUSHING 1996. La période de délai ("delay period") fait référence à la capacité de broutage.

Troisièmement une dynamique contrainte par les régimes diurnes et semi-diurnes des marées sera à prendre en compte comme illustré sur la quinzaine de mars 1995 sur la figure 1.3. En effet, la marée est une source régulière de turbulence. Les courants de flots et de jusants génèrent une remise en suspension importante et induisent également une variabilité temporelle importante de la couche de mélange créant des périodes de stratifications et de dé-stratifications [LAGADEUC et al. 1997].

L'effet de ce mélange n'agit pas directement sur le phytoplancton, mais il agit surtout sur la disponibilité en lumière et en nutriments dans l'écosystème. L'effet de la marée est variable selon les écosystèmes (océan ouvert, côtier ou estuaire) et selon le marnage. Certains écosystèmes côtiers seront propices au piégeage des sédiments tandis que d'autres présenteront toujours des eaux bien mélangées et turbides.

Pour finir, une variabilité inter-journalière peut aussi être considérée. Ces variations sont liées aux conditions d'éclairement. La migration verticale de certains organismes est transposable à une alternance jour et nuit. Pendant la journée, les organismes utilisent la lumière à la surface et durant la nuit, ils exploitent les ressources nutritives présentes plus en profondeur [REYNOLDS 2006].

Toutefois, il reste difficile de définir un schéma précis. En effet, s'ajoutent à ces variabilités

des ruptures dues à des évènements qualifiés d'extrême c'est-à-dire des évènements aléatoires, de courtes durées : chaque écosystème côtier présentant des évolutions spécifiques à sa géolocalisation. Ainsi, des blooms phytoplanctoniques multiples sous-jacents sont observés. Les efflorescences peuvent être des évènements d'amplitudes variables, de courtes ou plus longues durées, ponctuelles ou récurrentes, associés à des conditions hydrologiques, météorologies et chimiques toutes aussi ponctuelles, récurrentes ou exceptionnelles (Figure 1.3).

En conséquence, la distribution du phytoplancton est hétérogène avec une datation des épisodes variable, voire sporadique, et la description de ces assemblages dépend de l'échelle d'observation. Il est donc essentiel de baser les estimations de la biomasse phytoplanctonique à partir de mesures avec une résolution adaptée aux processus étudiés.

1.3 Caractéristiques environnementales relatives aux zones d'études

Afin d'identifier et de hiérarchiser les potentiels schémas fonctionnels, trois sites de mesures à HF du réseau national COAST-HF : MAREL-Carnot, MAREL-Iroise et MesuRho, situés respectivement en Manche, en mer d'Iroise et en mer Méditerranée, seront étudiés (Figure 1.5). Ils assurent ainsi la représentation des trois façades maritimes de l'hexagone français et vont permettre une inter-comparaison des schémas qui seront mis en évidence.

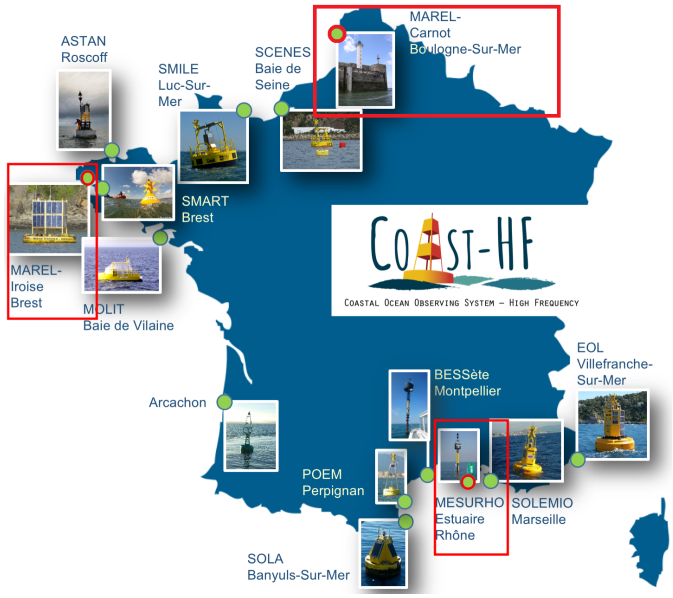


FIGURE 1.5 – Localisation des stations instrumentées du réseau national COAST-HF. L'encadrement rouge correspond aux stations sélectionnées dans cette étude.

La compréhension et la description de ces schémas passent donc par une bonne connaissance de ces trois zones d'études. Afin d'avoir une meilleure compréhension du milieu, cette partie décrit chaque zone.

1.3.1 Station MAREL-Carnot : Détroit du Pas-de-Calais, Manche

1.3.1.1 Hydrodynamique

La station MAREL-Carnot est positionnée en Manche. La manche est une mer épicontinentale qui couvre 40 % du littoral français. Elle est située entre le Royaume-Uni et la France. La presqu'île du Cotentin (France) sépare la Manche en deux bassins : la Manche occidentale et la Manche orientale (Figure 1.6). La Manche occidentale a une profondeur moyenne supérieure à 50 m avec plusieurs dépressions telles que la fosse d'Ouessant (200 m), la fosse de l'île Vierge (130 m), la fosse centrale (174 m) ou encore la fosse de la Hague (110 m). La Manche orientale s'étend sur 77 000 km^2 avec des profondeurs passant de 30 m près des côtes à 120 m, en moyenne, au centre du bassin. Le bassin étant principalement composé d'importants bancs combinés à des dunes (sans présence de dépressions marquées), la profondeur augmente progressivement de la côte vers le large. En forme d'entonnoir, elle est large de 80 km entre l'île de Wight et le Cotentin et de 35 km dans le détroit du Pas-de-Calais [Dauvin et Lozachmeur 2006 ; Dauvin 2008].

Figure 1.6 – Carte de la région marine de MAREL-Carnot. La Positions de la bouée MAREL-Carnot est indiquée par une croix rouge.

La station MAREL-Carnot se trouve dans le bassin oriental au niveau du littoral méridional du détroit du Pas-de-Calais (dans la rade de Boulogne-sur-Mer). Il est constitué de deux façades quasi rectilignes : la façade Ouest orientée Nord-Sud entre la baie d'Authie et le cap Gris-Nez, la façade Nord, orientée Sud-Ouest/Nord-Est entre le cap Gris-Nez et la frontière franco-belge (Figure 1.6).

Cette zone est caractérisée par un régime mégatidale semi-diurne et un marnage allant jusqu'à 9 mètres. La forme d'entonnoir du détroit présente une forte hydrodynamique de marée. De ce fait, sur la façade Ouest, les grandes marées provoquent d'importants courants parallèles à la côte : un courant de flot (vive-eau) vers le nord-est et un courant de jusant (morte-eau) vers le sud-ouest [Sournia et al. 1990 ; Reynaud et al. 2003].

La morphologie du bassin et ce régime de marée favorisent la création d'une zone côtière fortement influencée par les apports fluviaux de plusieurs fleuves : la Seine, la Somme, la Canche, la Liane et le Wimereux) [Sournia et al. 1990]. Cette zone côtière est délimitée sur sa « rive » droite par le littoral et sur sa « rive » gauche par une zone frontale (Figure 1.7). Dans cette zone sont observés des chutes de la salinité et des apports importants en nutriments, Matière En

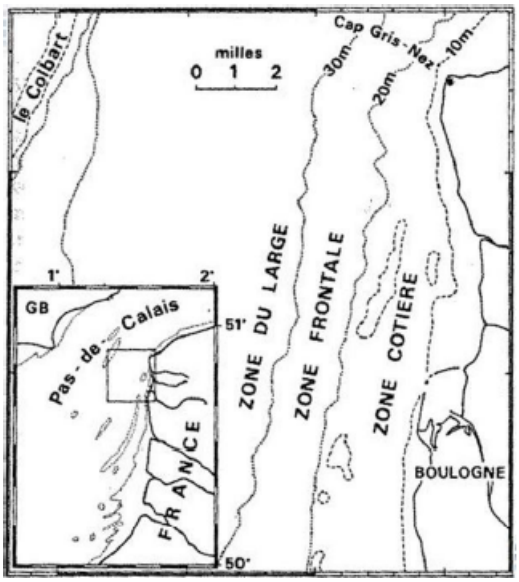


FIGURE 1.7 – Structuration du fleuve côtier en Manche orientale en trois zones : zone du large,zone frontale et zone côtière [SOURNIA et al. 1990].

Suspention (MES) et Matière Organique (MO) [DAUVIN 2008] liés aux affluents. La Seine, du fait de son débit important (moyen annuel de $510\,\mathrm{m\,s^{-3}}$), est l'affluent principal. Il détermine en majeure partie les caractéristiques de cette zone. Cependant, Brylinski et al. (1996) indiquent qu'il n'est pas le seul acteur de la zone côtière, la Somme dont ses apports se font ressentir jusqu'au détroit du Pas-de-Calais influence également les caractéristiques de cette zone côtière [LEFEBVRE, GUISELIN et al. 2011].

Au moment de la morte-eau, les eaux de la zone côtière passent au-dessus des eaux de la zone du large, ainsi une stratification verticale s'établit. Lors de la vive-eau, une barrière turbide verticale se crée entre le fleuve côtier et l'eau du large : la stratification est détruite. Cette zone, avec son hydrodynamique et ses caractéristiques physico-chimiques particulières, est un lieu idéal pour la prolifération du phytoplancton. Le phytoplancton peut ainsi être transporté sur de très longues distances, de la baie de Somme au détroit du Pas-de-Calais.

Un autre apport d'eau douce plus local est à prendre en compte. En effet, la rade de Boulogne-sur-Mer est située à la sortie de la Liane qui y déverse ses eaux à chaque ouverture du barrage Marguet. La Liane prend sa source à Quesques à 101 mètres d'altitude et fait 36 kilomètres de long. Son débit mensuel moyen se situe entre 3,06 et 5,33 m^3/s. Ces apports dépendent des ouvertures de barrage et constituent un forçage aléatoire avec, cette fois, une dessalure, un apport en nutriments, MES et MO plus directs et occasionnels.

Pour résumer, la zone d'étude rattachée à la station MAREL-Carnot est donc caractérisée par un régime mégatidale générant une structure frontale importante et des panaches fluviaux, globalement liés à la Seine et plus localement à la Liane.

1.3.1.2 Physico-chimie

La station MAREL-Carnot est située en zone tempérée où les conditions de température et d'éclairement suivent une dynamique saisonnière. De plus, elle est implantée dans un milieu

eutrophe sous influence du régime de marée. La morphologie du bassin et ce régime de marée créent une zone côtière sous influence d'une structure frontale. Cela a pour conséquence de concentrer les apports terrigènes naturels et anthropiques à la côte (Section 1.3.1.1). Cette zone regroupe donc les conditions optimales à la croissance du phytoplancton : apport en nutriments, pas de profondeur critique, turbidité acceptable pour le développement. Bien que riche en nutriment, cette zone est soumise à des variations. Ces changements sont marqués par des périodes clefs avec des nutriments potentiellement limitants. De manière générale, on observe une limitation par les silicates tout au long de l'année suivie par une limitation en Phosphate P pour les périodes de janvier à avril et octobre à décembre. Enfin, une limitation en Nitrate N de mai à septembre est observée [Lefebvre, Guiselin et al. 2011] (Tableau 1.1).

Tableau 1.1 – Pourcentages d'occurrence des principaux nutriments limitatifs potentiels (nitrate, phosphate ou silicate, par ordre de priorité) pour les trois stations côtières au cours de la période 1992-2007 pour n=226 observations (tiré de Lefebvre, Guiselin et al. 2011).

P :Si :N	Si :P :N	Si :N :P	n
6	41	36	226

Ainsi, il est aussi observé, dans cette zone, une dynamique phytoplanctonique saisonnière [Breton, Brunet et al. 2000 ; Schapira et al. 2008]. Les concentrations en Chl-a (proxy de la biomasse phytoplanctonique) sont faibles en période hivernale, puis des efflorescences se déclenchent en mars pour atteindre des maxima en avril. Les concentrations vont ensuite chuter progressivement en fonction de la consommation en nutriment jusqu'à atteindre des valeurs minimales en période estivale.

1.3.1.3 Variations phytoplanctoniques

La dominance des espèces est considérée comme un critère important dans la description des assemblages du phytoplancton [Ignatiades 1994]. En Manche et en mer du Nord, la diversité des assemblages phytoplanctoniques suit le schéma saisonnier des zones tempérées (détails Section 1.2.3).

Les diatomées (*Bacillariophyceae*) sont les espèces prédominantes. Breton, Brunet et al. 2000 souligne que les diatomées contribuent à la plus grande partie de la biomasse du phytoplancton. La présence en abondance du genre *Chaetoceros* a été enregistrée en période estivale [Rodríguez et al. 2000], l'espèce *Rizosolenia* associée à ce genre a aussi été observée [Ryckaert et al. 1983]. En période hivernale, ce sont plutôt les diatomées de petite taille qui sont dominantes telles que *Thalassiorira sp* ou encore *Lauderia sp* [Breton, Brunet et al. 2000 ; Vantrepotte et al. 2007]. En période printanière, lors de l'efflorescence principale, est identifié le genre *Pseudo-nitzschia* en concomitance avec le bloom de *Phaeocystis globosa* (*Prymnésiophycae*) [Delegrange et al. 2018]. Ces espèces ont toutes deux été répertoriées comme espèces potentiellement toxiques ou nuisibles (*Harmful Algal Blooms* (HABs)). Les dinoflagellés (*Dynophyceae*) sont la deuxième classe la plus abondante en zones tempérées. Leur développent suit souvent le bloom printanier. Les diatomées sont remplacées par les dinoflagellés au fur et à mesure que la concentration en silice diminue. Les genres *gymnodinilaes, gyrodinuim et prorocentrum heterocapsa* et d'autres sont ainsi retrouvés. Les Cryptophycées (*Cryptophyceae*) sont quant-à-elles dominées par le genre *Teleaulax / Plagioselmis* ou encore *Geminigera spp* et forment généralement des blooms en période automnale [Medlin et al. 2017 ; Breton, Brunet et al. 2000].

Bloom printanier et algues nuisibles.

Sur le littoral Est de la Manche, *Phaeocystis globosa* contribue à plus de la moitié de l'abondance

totale, notamment lors des efflorescences printanières. Ces blooms sont des phénomènes liés à l'eutrophisation des côtes [Lancelot et Mathot 1987 ; Schoemann et al. 2005]. Ainsi, ces organismes peuvent supplanter la dominance des diatomées lors de dystrophie du milieu. Dans ce cas c'est l'apport de sels nutritifs azotés qui joue un rôle fondamental dans le développement de cette algue [Gentilhomme et Lizon 1997]. L'apport accru d'azote entraîne un déséquilibre (la dystrophie) du rapport Si/N. Ce déséquilibre entraîne une limitation par le silicium de la croissance des diatomées et permet aux espèces non siliceuses tel que *Phaeocystis* de se développer.

Or, son stade biologique est caractérisé par une forme coloniale palmelloïde, c'est-à-dire agrégée dans un mucilage. Lors de la dégradation de ces colonies, une écume de plusieurs centimètres se forme sur le littoral. Cette étape consomme une grande quantité d'oxygène ce qui peut conduire à des phénomènes d'hypoxie, voire d'anoxie. Des effets combinaient avec d'autres espèces toxiques tel que Pseudo-nitzchia. Les blooms ont donc un effet négatif sur le fonctionnement de l'écosystème et les activités aquacoles et de pêches [Weisse et al. 1994 ; Schoemann et al. 2005]. Par conséquent, les colonies de *Phaeocytis* ont été catégorisées comme algues nuisibles [D. M. Anderson et al. 1998 ; Veldhuis et Wassmann 2005].

Une autre espèce toxique, dont les blooms se manifestent pendant ou à la fin du printemps, est à considérer : les diatomées du genre *Pseudo-nitzichia*. Ce genre comprend plus de 50 espèces dont 40 sont connues pour produire de l'acide domoïque (AD) [Delegrange et al. 2018 ; Lundholm 2018]. L'acide domoïque est une neurotoxine qui provoque des amnésies suite à la consommation de mollusque, soit une toxine du type Amnesic Shellfish poisoning (ASP). Ces espèces de diatomées sont souvent épiphytes d'autres espèces. Dans ce cas, elles sont souvent observées à l'intérieur des colonies de *Phaeocystis* [Sazhin et al. 2007]. Il est envisagé qu'elles utilisent la MO produite par les colonies de *Phaeocystis* pour favoriser sa croissance [Loureiro et al. 2009].

1.3.2 Station MAREL-Iroise : Rade de Brest, Mer d'Iroise

1.3.2.1 Hydrodynamique

La plateforme d'observation MAREL-Iroise est localisée à l'entrée du goulet dans la Rade de Brest (Figure 1.8).

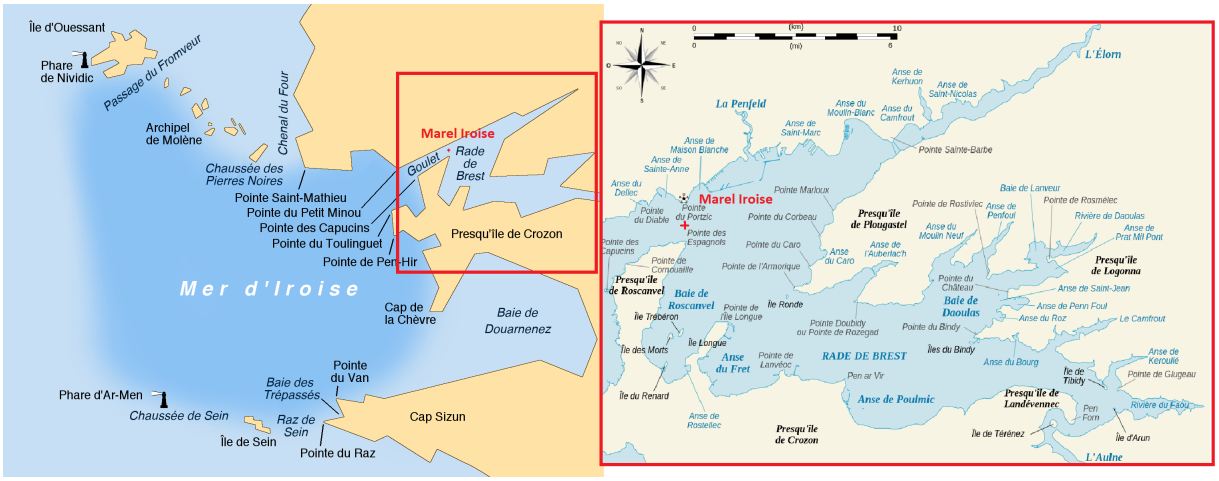


Figure 1.8 – Carte de la région marine de MAREL-Iroise, position du goulet, de la rade et la mer d'Iroise. Les différents affluents sont aussi tracés. La position de la bouée MAREL-Iroise est indiquée par une croix rouge.

La rade de Brest est un bassin peu profond (8 m) d'une surface de 180 km^2. Celle-ci est une baie semi-fermée connectée à la mer d'Iroise par le biais du goulet (d'une largeur de 1,8 km) [Le Pape et Menesguen 1997]. C'est donc un lieu de transit entre les deux masses d'eaux.

La rade de Brest est un milieu fortement soumis à l'influence maritime [Roger 1981]. En effet, elle est animée par de forts courants de marée et d'intenses tempêtes hivernales qui favorisent les échanges entre la rade et le large. Ces marées de type semi-diurne, avec un marnage moyen de 4 mètres et un marnage de vives-eaux pouvant aller jusqu'à 8 mètres, exercent un brassage continu sur la rade. Le brassage est tel qu'une grande partie (jusqu'à 40 %) des eaux évacuées de la rade pendant le jusant sont ré-introduites lors du flot suivant [Le Pape et al. 1996] à travers le goulet. Les courants tidaux ont donc un impact conséquent sur le taux de turbidité et le volume d'eau de la rade.

Une autre des spécificités de la rade est : l'influence des apports fluviaux de l'Aulne au sud et de l'Elorn au Nord-Est [Le Jehan et Tréguer 1984]. L'Aulne prend sa source dans le Mont d'Arrée et se jette dans la rade à Chateaulin après un parcours de plus de 120 km. L'Elorn débute dans la région de Sizun-Comana et termine son cours à Landerneau. Les bassins de l'Aulne et L'Elorn représentent respectivement 70 et 15 % des apports annuels cumulés en eau douce dans la rade [Troadec et Le Goff 1997]. Ces deux fleuves alimentent la zone en eau douce et fertilisent la rade lors des périodes de crues hivernales (d'octobre à mars). Ainsi, le taux de salinité peut être séparé en deux périodes : période de crue marquée par une diminution de la salinité, avec des taux jusqu'à $\tilde{3}0$ dans les eaux de surface et $\tilde{3}3$ dans les eaux de fond et une période d'étiage avec une salinité homogène sur la colonne d'eau allant au-delà de 35 [Del Amo et al. 1997 ; Savoye 2001].

Toutefois, cette influence doit être nuancée, car le goulet reste principalement soumis à l'influence des courants de marée. En effet, la stratification induite par les apports d'eau douce peut être réduite par les courants de marée, surtout en période de vives-eaux. Ainsi, en période de fortes crues, la stratification est plus marquée vers l'embouchure des rivières et nulle dans le goulet de Brest [Ragueneau, De Bias Varela et al. 1994 ; Chauvaud, Thouzeau et al. 1998].

De manière générale, les caractéristiques biochimiques et physiques de la rade sont donc fonction des fortes variations mégatidales et des divers apports fluviaux de la zone lors des épisodes de forte crue en périodes hivernales, ainsi que des cycles saisonniers.

1.3.2.2 Physico-chimie

La station MAREL-Iroise se situe dans un milieu eutrophe profondément influencé par le phénomène de marées qui gouverne la répartition des masses d'eaux marines et fluviales. De manière générale, le régime macrotdial de la zone génère un fort brassage et une remise en suspension importante. Ce régime turbulent va donc favoriser l'accumulation hivernale et la remise en suspension en période estivale des sels nutritifs et induire des concentrations en Matières En Suspension (MES) relativement constantes. La présence de MES va amplifier les phénomènes biogéochimiques d'absorption et désorption [Delmas et Tréguer 1983] et donc de régénération des nutriments.

À cette dynamique macrotidale, s'ajoute l'impact important des apports fluviaux. En effet, il est possible d'observer une évolution des éléments nutritifs relatifs aux débits fluviaux, dans la zone estuaire. Roger 1981 puis Le Jehan et Tréguer 1984 constatent que la teneur en nitrate varie linéairement avec le débit de l'affluent principal.

Or depuis les années 70, les concentrations de nitrates ont considérablement augmenté dans les rivières. Selon Le Pape et al. 1996, elles ont doublé dans les rivières Aulne et Elorn entre les années 1970 et les années 1990. Cette augmentation est principalement liée aux activités

agricoles de la zone. Ainsi, les apports anthropiques fluviaux (d'azote et phosphore) injectés dans les eaux côtières induisent une modification des rapports N : P : Si dissous dans les eaux côtières qui peut conduire à des conditions d'eutrophisation. Des perturbations importantes de l'écosystème comme une diminution à long terme du rapport Si :N sont constatées [Del Amo et al. 1997 ; Le Pape et al. 1996 ; Le Pape et Menesguen 1997]. Toutefois, même si le déversement fluvial de nitrate dans la baie a considérablement augmenté, aucune augmentation de biomasse phytoplanctonique significative (marées vertes) n'est constatée, à l'exception de zones très localisées près de l'embouchure des rivières [Le Pape et Menesguen 1997]. La résistance de cet écosystème à l'eutrophisation a été expliquée par le taux de renouvellement rapide de l'eau de la baie, liée au régime macrotidal de la zone [Roger 1981] et par l'exportation de 94 % d'azote inorganique dissous vers les eaux côtières avant le printemps [Le Pape et al. 1996], liée au découplage entre les apports d'azote (hiver et printemps) et l'optimum de température (été) pour le développement des macroalgues. Bien que la zone soit globalement non-eutrophisée, les rapports N : P : Si de la baie sont définitivement perturbés et conduisent à une modification des schémas de successions phytoplanctoniques (1.3.2.3).

1.3.2.3 Variations phytoplanctoniques

La dynamique phytoplanctonique de la Rade de Brest a fait l'objet de nombreuses études depuis les années 80 [Quéguiner et Tréguer 1984 ; Ragueneau, De Bias Varela et al. 1994 ; Del Amo et al. 1997 ; Beucher et al. 2004 ; Ragueneau, Chauvaud, Moriceau et al. 2005 ; Claquin et al. 2010]. Elle suit un schéma classique des zones tempérées avec une efflorescence printanière principale (démarrant autour du mois avril) suivie de plus petites efflorescences jusqu'au mois d'octobre. Elles sont majoritairement dominées par les diatomées.

Le groupe des diatomées (*Rhizosolenia , Nitzschia et Chætoceros* principalement) est présent une majeure partie de l'année avec une dominance des diatomées de petite taille en périodes hivernales et des diatomées de grandes tailles en périodes printanières [Quéguiner et Tréguer 1984].

L'épuisement des silicates et l'enrichissement en azote et phosphore vont créer un déséquilibre des rapports N : P : Si qui engendre un déplacement des diatomées vers d'autres espèces phytoplanctoniques [Ragueneau, De Bias Varela et al. 1994]. Ainsi, les dinoflagellés (comme *Gyrodinum , Gymnodinium , Prorocentrum* ou encore *Alexendirum*) sont fréquemment rencontrés en quantité importante, notamment en été et en période printanière succédant à un bloom de diatomée. Les espèces *Alexendrium* et *Gymnodinium* sont répertoriées comme des espèces potentiellement toxiques ou nuisibles (HABs), elles ont un impact sur certaines espèces, notamment les coquilles Saint-Jacques [Chauvaud, Thouzeau et al. 1998] ou mortelles pour la faune marine. Les cryptophycées (*fibula, Distephanus speculum, Ebria tripartia*) sont présentes plus ponctuellement tout au long de l'année.

Écosystème perturbé : Impact du cycle du silicium sur la dynamique phytoplanctonique. L'augmentation de la concentration d'azote (N) et de phosphore (P) d'origine anthropique dans les rivières, a pour conséquence directe la diminution des rapports Si : N et Si :P. Les conséquences indirectes de cette augmentation des apports (N et P) se retrouvent au travers du cycle du silicium (Si) et affectent la dynamique du phytoplanctonique dans les eaux côtières [Officer et Ryther 1980 ; Ragueneau, Chauvaud, Moriceau et al. 2005 ; Billen et Garnier 2007]. Dans cette configuration, l'acide silicique est rapidement épuisé lors de l'efflorescence printanière de la mi-mai et devient le principal facteur limitant, responsable de l'effondrement de l'efflorescence. Cette limitation génère une transition entre une production primaire basée sur les diatomées

à une production primaire dominée par d'autres groupes de phytoplancton, par exemple les dinoflagellés (Figure 1.9 B) qui comprennent de nombreuses espèces toxiques. La limitation du Si des efflorescences printanières de diatomées est devenue une caractéristique commune des écosystèmes côtiers qui reçoivent des apports d'eau douce [Smayada 1990b].

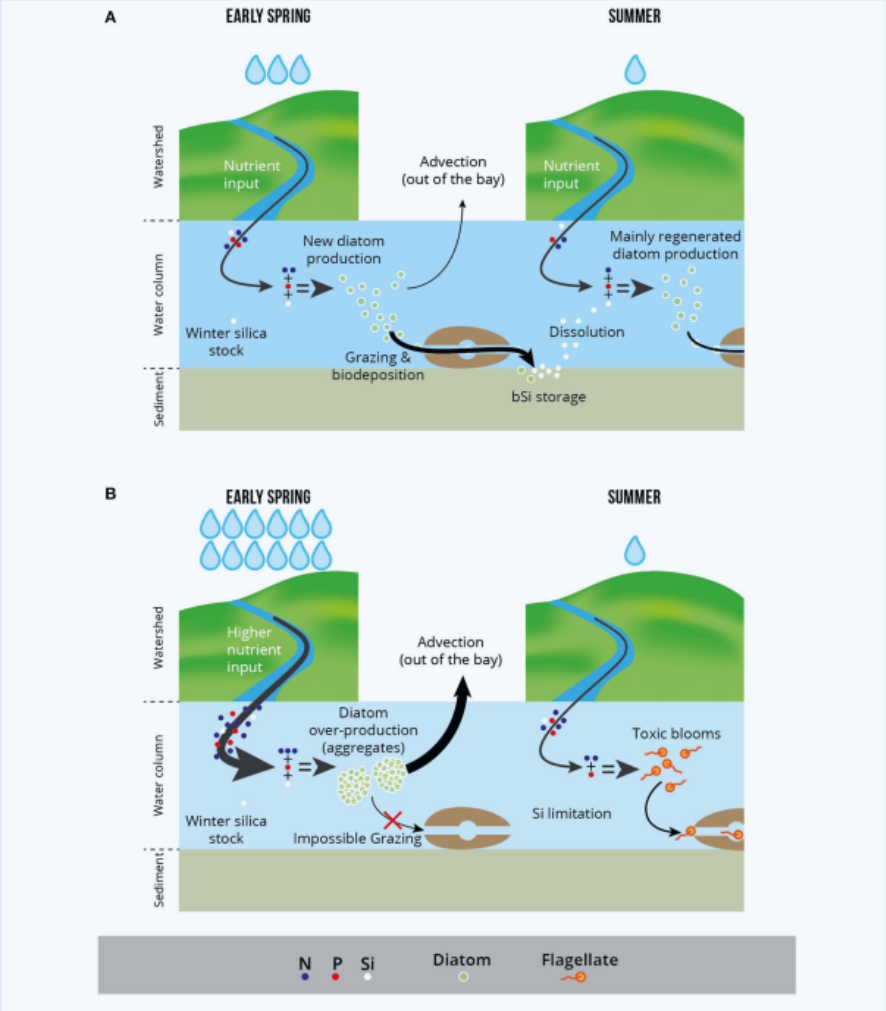


Figure 1.9 – Schématisation de l'hypothèse de travail de Chauvaud, Jean et al. 2000 sur combinaison de l'activité de suspensivore (Crepidula) et le recyclage du Si. Deux situations contrastées sont affichées : A) La production printanière de diatomées est broutée par des suspensivores benthiques (dominés par les crépidules) ; le Si est stocké dans les sédiments et lentement libéré sous forme d'acide silicique pendant l'été, permettant le maintien des diatomées dans le système. B), la production de diatomées printanières ne peut pas être broutée et la majeure partie de la production de diatomées est exportée hors de la baie, appauvrissant le système en acide silicique et favorisant une production estivale de dinoflagellé (Schémas issus de Ragueneau, Raimonet et al. 2018).

La particularité du cas de la limitation du Si en baie de Brest est que l'écosystème semble résister à ces perturbations. Ainsi, un rallongement de la période productive, plus fractionnée et plus étalée sur l'année [Chauvaud, Jean et al. 2000] et un maintien d'une dominance par les diatomées en période de production [Ragueneau, Chauvaud, Leynaert et al. 2002] ont été

constatés. Une des hypothèses récemment émise [Chauvaud, Jean et al. 2000] et testée dans plusieurs autres études [Ragueneau, Chauvaud, Leynaert et al. 2002 ; Martin et al. 2006] pour expliquer ce phénomène est : la combinaison de l'activité des suspensivores et le recyclage du Si via "la pompe des silictes". En effet, la prolifération de suspensivores, tel que *Crepidula fornicata*, serait un des moteurs importants de la pompe [Del Amo et al. 1997 ; Ragueneau, Raimonet et al. 2018]. La filtration benthique retiendrait le Si dans la baie au printemps ce qui aurait pour effet direct de diminuer la biomasse à cette période, via une concurrence directe pour l'accès aux nutriments et les phénomènes de prédations (broutage). Mais l'effet secondaire qui serait le recyclage de silice biogénique pendant l'été, fournirait le Si nécessaire pour maintenir les populations de diatomées (Figure 1.9 A). Ainsi, les organismes benthiques joueraient un rôle important sur la composition spécifique des floraisons secondaires.

Cependant, même si l'écosystème ne semble pas propice à prolifération des HABs, depuis 2012 leurs fréquences et leurs ampleurs augmentent dans le sud de la Rade avec *Alexandrium minutum* [Annie et al. 2015]. De nombreuses raisons peuvent induire cette nouvelle prolifération et il est difficile de l'attribuer à un seul facteur. Néanmoins, de fortes diminutions de la biomasse totale de *C. fornicata* sont aussi constatés à ses périodes (données cité dans Stiger-Pouvreau et Thouzeau 2015). Ceci laisse penser que la présence de *C. fornicata* doit sûrement empêcher ou ralentir le développement des HABs.

1.3.3 Sation MesuRho : Golf du Lion, Méditerranée

1.3.3.1 Hydrodynamique

La station MesuRho est située, sur le littoral Nord du bassin occidental de la mer Méditerranée, dans l'Est du Golfe du Lion, au niveau de l'embouchure du Rhône (Figure 1.10).

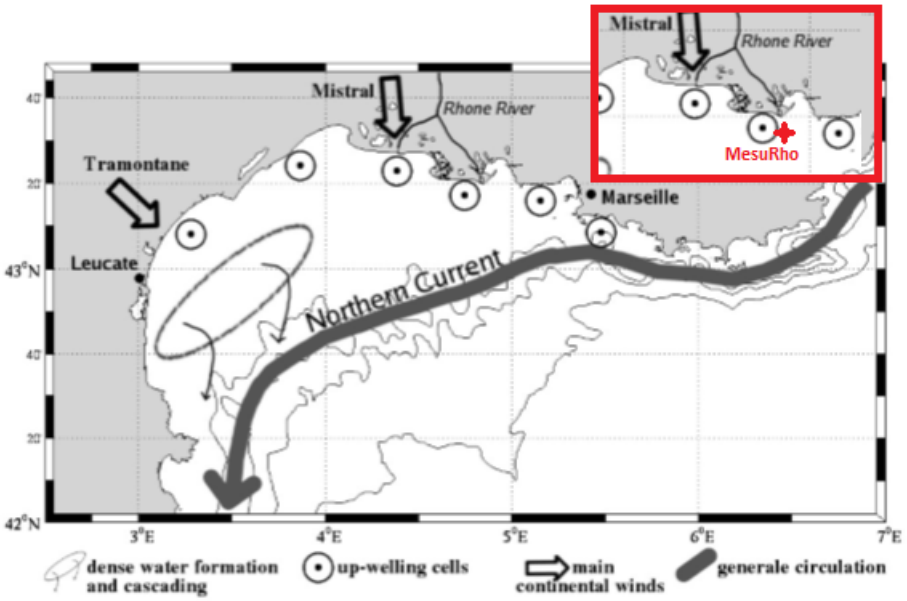


Figure 1.10 – Carte de la région marine de MesuRho, principales caractéristiques de la circulation (redessinée de Millot (1990)). Les isobathes 20, 50, 90, 160, 500, 1000 et 2000 m sont tracées. La position de la bouée MesuRho est indiquée par une croix rouge.

Le Golfe du Lion s'étend le long des côtes françaises, de Marseille au Cap Creus. Il est caractérisé par un grand plateau de forme semi-circulaire, d'une profondeur moyenne de 90 m et des régimes de vents importants tels que la Tramontane et le Mistral. Ces caractéristiques vont impacter la circulation générale du Golfe du Lion et induire plusieurs processus hydrodynamiques [Millot 1990] (Figure 1.10).

Le principal courant présent dans cette zone est le courant Liguro-Provençal également appelé Courant Nord (Figure 1.10). Ce courant est fortement contraint par le plateau continental qui forme une barrière entre les eaux côtières du plateau continental et les eaux de la mer ouverte [Albérola, Millot et Font 1995 ; Petrenko 2003]. Toutefois des intrusions sur le plateau ont été observées lors de conditions de vents particulières ou lors de situations hydrologiques du plateau et d'activités méso-échelles spécifiques du CN [Pairaud et al. 2011]. Ses instabilités à méso-échelle se produisent principalement au printemps et en hiver lorsque le courant est maximal [Flexas et al. 2002 ; Albérola et Millot 2003]. Même si ces instabilités peuvent affecter la circulation du plateau, le plateau est essentiellement forcé par les vents et par le débit de flottabilité du Rhône.

En effet, les vents continentaux intenses et fréquents (le Mistral et la Tramontane) [Millot 1990] repoussent les eaux de surface au large des côtes et créent 6 cellules principales d'upwellings. Ces vents froids et secs qui apparaissent en hiver peuvent aussi créer une zone de formation d'eau dense. Cette eau dense se forme principalement sur la partie Ouest du plateau (Figure 1.10).

Ces forçages atmosphériques et hydrologiques vont aussi influencer les apports du Rhône et leurs répartitions géographiques [Broche et al. 1998 ; Marsaleix et al. 1998 ; Reffray et al. 2004] en modifiant la forme et l'étendue de son panache. Par exemple, en période de vent faible, le panache est entraîné vers l'ouest de l'embouchure par la force de Coriolis mais en cas de vent de secteur nord (Mistral et la Tramontane), le panache est décollé de la côte et entraîné vers le sud/sud-ouest de la côte [Amandine 2010].

Le Rhône est le principal affluent du bassin Nord Méditerranée et se jette au niveau de la zone d'étude par le détroit de Camargue, dans le Golfe du lion. Avec un débit annuel moyen de $\pm$ 1700 $\mathrm{m\,s^{-3}}$ et une crue annuelle typique avec un débit $> 5000\,\mathrm{m\,s^{-3}}$ [Maillet et al. 2006], c'est la source d'eau douce la plus importante de Méditerranée [Pont et al. 2002].

Du fait de son débit, il a un impact important sur une grande partie du Golfe du Lion. Il fournit 80 % des apports sédimentaires du golfe [Courp et Monaco 1990 ; Bourrin et Durrieu de Madron 2006]. Maillet et al. 2006 ont estimé un flux total de solides en suspension d'environ 7×10^6 tonnes par an avec une variabilité annuelle élevée de 1,2 à 19,7$\times 10^6$ tonnes par an. Ces apports de matières organiques et inorganiques d'origine naturelle et anthropique vont avoir des effets sur l'écologie et la sédimentation de la région surtout en période de crues importantes. Deux mécanismes particuliers de transport sédimentaire sont connus : pour les périodes de crues et de tempêtes (Figure 1.11). En régime de crue, un système à deux couches se crée à l'embouchure du Rhône. Les MES sont essentiellement exportées vers le large en surface, et chutent au cours de leur transport, créant une couche de fond. En régime de tempête, le matériel sédimentaire est transporté près du fond, il provient d'apports fluviaux et/ou de la remise en suspension [Lorthiois 2012].

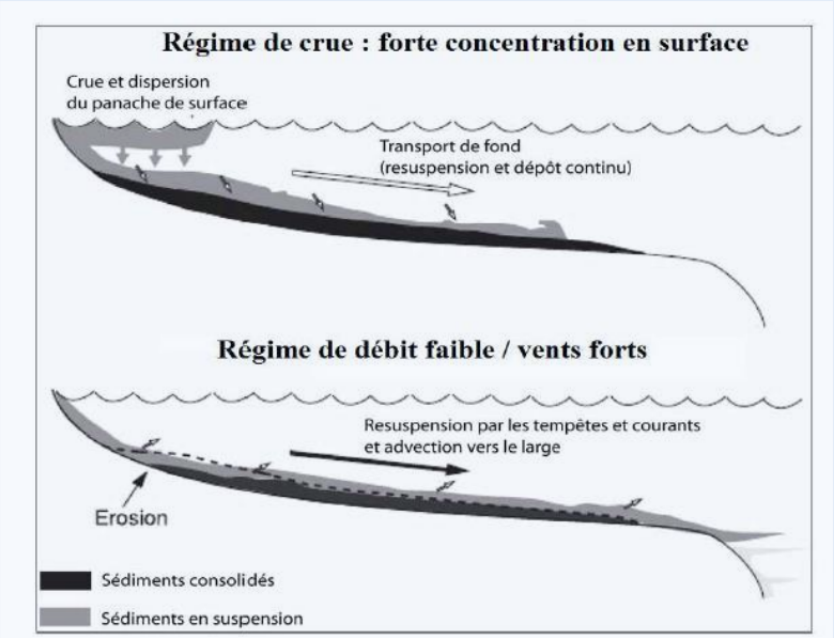


FIGURE 1.11 – Mécanisme de transport intervenant pour un régime de crue (haut) et pour un débit liquide faible associé à des vents forts (bas). (issus de LORTHIOIS 2012)).

1.3.3.2 Physico-chimie

Les eaux de méditerranée sont connues pour leur caractère oligotrophe. De manière générale, elles sont marquées par un déficit de phosphore relativement à l'azote. Ce composé se rencontre en effet à des concentrations 20 à 25 fois plus faibles que le Nitrate, or le rapport idéal est estimé à N/P=16/1 pour la croissance du phytoplancton [MONACO 2009].

Mais la station MesuRho se situe au cœur de la circulation du Golfe du Lion sous influence Rhodanienne forte. Cette zone est caractérisée par une forte variabilité, marquée par des dessalures et des apports de nutriments importants du fait de l'influence des eaux du Rhône. Les apports en eau douce et en nutriments du Rhône sont les plus importants de la Méditerranée, le Nil arrivant en seconde place depuis la construction du barrage [SEMPÉRÉ et al. 2000]. Il représente un tiers des apports totaux reçus par les eaux de surface de la Méditerranée. Ces apports sont estimés à 80-100 $\mathrm{kt\,an^{-1}}$, dont 70 % de nitrate, et 6,5 $\mathrm{kt\,an^{-1}}$, soit 6 % de l'apport total [MONACO 2009]. Ainsi, ce fleuve constitue un forçage principal de la zone d'étude.

D'après MONACO 2009, une augmentation des nitrates et une diminution continue des flux en phosphate sont constatées. En effet, sur les 30 dernières années les concentrations en azote ont doublé (en 2009 des taux de 1,8 $\mathrm{mg\,l^{-1}}$ sont enregistrées). Contrairement au nitrate, les taux de phosphore ont une tendance à diminuer depuis les années 1990. Cette diminution est liée aux mesures prises par les industriels. En 2009 des taux de l'ordre de 0,1 $\mathrm{mg\,l^{-1}}$ sont enregistrés.

Ainsi, les concentrations moyennes en nutriments du panache créent des conditions plus favorables aux développements phytoplanctoniques. Ces apports peuvent supporter entre 23 % et 69 % de la production primaire moyenne du Golfe du Lion [LUDWIG et al. 2009].

1.3.3.3 Variations phytoplanctoniques

Ainsi, les concentrations moyennes en nutriments du panache créent des conditions plus favorables aux développements phytoplanctoniques. Minas 1968 puis Lefevre et al. 1997 ont défini 4 zones hydrologiques en fonction de la production primaire dans le Golfe du Lion (Figure 1.12) : (i) la zone du golfe de Marseille délimitée au nord par le littoral, à l'ouest par le panache, et au sud par le courant nord. C'est un système oligotrophe où la densité de population est très faible (5.10^3 Cell.l^{-1}) [Blanc et al. 1969] (ii) le panache du Rhône, limité à quelques kilomètres de l'embouchure, est une zone très productive, contenant de fortes concentrations en nutriments (463.10^3 l^{-1}) [Blanc et al. 1969] (iii) la zone de dilution sous influence intermittente du panache. Cette zone est considérée comme zone productive avec un gradient de l'embouchure vers l'ouest. (iv) la zone sud, zone du large du Golfe du Lion incluant le Courant Nord oligotrophe.

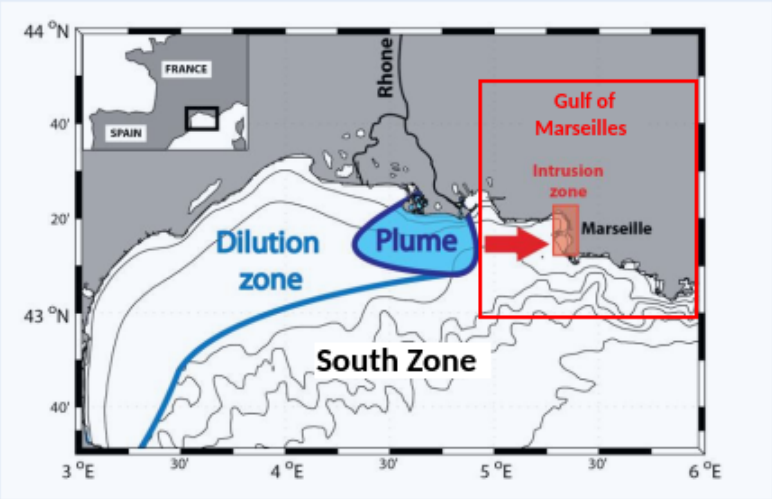


Figure 1.12 – Représentation des 4 zones hydrologiques définies par Lefevre et al. 1997. (i) En rouge le Golfe de Marseille délimité au nord par le littoral, à l'ouest par le panache, et au sud par le courant nord. Zone généralement oligotrophe. En rouge clair est définie la zone d'intrusion dans la baie de Marseille. (ii) En bleu le panache du Rhône, limité à quelques kilomètres de l'embouchure. Elle est considérée comme une zone très productive (iii) En cyan la zone de dilution limitée à l'est par le panache, à l'ouest et au nord par le littoral, et au sud par le courant nord. La production de la zone est marquée par un gradient de l'embouchure vers l'ouest au large. (iv) En noir la zone sud, représentée par une zone au large de 3 ° de longitude par 2 ° de latitude, centrée sur 42°N 5° E et bordée par le courant nord (issus de Fraysse 2014).

Chaque zone comprend des caractéristiques propres au niveau de son taux de production. Ces différences vont aussi se retrouver au niveau communautaire. Mais de manière générale, d'après le rapport DCE [Belin et al. 2012], au niveau du Golfe du Lion le nanophytoplancton domine, avec une abondance haute toute l'année et une concentration très importante à la fin du mois d'avril et en mai. Pour le microphytoplancton, ce sont principalement les diatomées qui dominent. On retrouve durant l'année et par ordre d'importance : *Leptocylindrus sp., Chaetoceros sp., Pseudo-nitzschia sp. et Rhizosolenia sp.* En juillet, avec les apports du Rhône, la communauté microphytoplanctonique est presque entièrement dominée par deux espèces de diatomées, *Leptocylindrus sp. et Pseudo-nitzschia seriata.* Les dinoflagellés, essentiellement représentés par *Gymnodinium sp. et Gyrodinium sp.* sont présents tout au long de l'année avec une abondance moyenne relative de 36 %. Les silicoflagellés (*Dictyocha sp.*) constituent le groupe le moins

abondant.

Ainsi, une dynamique saisonnière classique est retrouvée. Le déclenchement de l'efflorescence en mars et la diminution en période estivale sont bien observés dans le Golfe du Lion. Toutefois cette variabilité est moins marquée que dans les autres stations d'études (MAREL-Carnot, MAREL-Iroise) en raison des apports Rhodaniens et de la cellule d'upwelling à proximité. Ces conditions hydrologiques modifient la production primaire et la composition des communautés phytoplanctoniques en fonction de chaque zone hydrologique [Lefevre et al. 1997 ; Ouillon et Petrenko 2005]. Il faut donc s'attendre à de nombreuses variations sporadiques supplémentaires dépendantes des crues et du mélange vertical. C'est variations engendrent, à première vue, une certaine constance des taux de concentration en biomasse phytoplanctonique au niveau de notre zone d'étude ((ii) le panache du Rhône).

Pro-delta du Rhône : Effet Eutrophique Le panache du Rhône, où se situe la station MesuRho est la zone la plus eutrophe du Golfe du Lion. Dans cette zone les apports du Rhône ont un impact considérable sur la croissance du phytoplancton. Ainsi cette zone a été définie comme la zone la plus productive (300 à plus de 1500 $\mathrm{g\,m^{-2}}$ de carbone par an)[Lefevre et al. 1997]. Des concentrations en chlorophylle a constant d'environ 1 à 3 $\mu\mathrm{g\,l^{-1}}$ ont été constatées quelles que soient la forme et la propagation du panache [Morel et André 1991].

Conan et al. 1998 (part II) définissent pour la zone du panache 3 états : (1) une situation hivernale, (2) une situation d'efflorescences (correspondant au *bloom* printanier), et (3) une situation estivale.

En (1), les taux de concentration en Chl-a sont relativement élevés. En effet, les conditions d'éclairement et de température ne sont pas limitantes pour les communautés phytoplanctoniques présentes. Le régime de marée étant très faible voir quasi inexistant, ce sont les apports fluviaux et la convection thermohaline de la saison [Coste et al. 1972] qui génèrent des apports en nutriments. Dans ces conditions, tous les nutriments rejetés par le Rhône sont disponibles pour les phytoplanctons marins [Lochet et Leveau 1990].

Ainsi, les niveaux élevés de nutriments dans la région provoquent des proliférations phytoplanctoniques (dominées par les diatomées) pendant l'hiver [Karydis et Kitsiou 2012]. L'état (2) commence entre mars et avril, c'est à ce moment que le niveau de biomasse est le plus élevé. L'efflorescence précède l'apparition de la thermocline lors du réchauffement des eaux de surface. Elle est dominée par de fortes concentrations de diatomées, particulièrement adaptées aux conditions trophiques mésotrophes [Herrmann 2007]. En contact direct avec le Rhône, des espèces dulçaquicole comme *Staurosiracapucina*, saumâtres comme *Melosira juergensi et varians* sont présentes [Blanc et al. 1969]. À ce stade, la quantité de nutriments diminuent jusqu'à atteindre des taux limitants la croissance. À l'état (3), les stocks de nutriments en surface sont faibles, voire épuisés. Ainsi la biomasse de phytoplancton diminue selon les taux de nutriments disponibles [Lohrenz et al. 1988 ; Verity et Smetacek 1996]. Lors de l'appauvrissement en nutriment, une transition des diatomées vers une dominance par les dinoflagellés est observée, dominance favorisée par les eaux statiques et les températures élevées [Karydis et Kitsiou 2012].

Bien que le Golfe du Lion ne soit pas exposé aux risques d'eutrophisation en raison du temps de séjour court des eaux [Pairaud et al. 2011]. Le golfe de Fos, dont la zone pro-deltaïque du Rhône, est caractérisé par de faibles profondeurs et des concentrations plus élevées de chlorophylle a (jusqu'à 10 $\mu\mathrm{g\,l^{-1}}$), ce qui indique que le risque d'eutrophisation est plus élevé [Fraysse 2014]. De plus, les phénomènes d'eutrophisation et de prolifération d'algues nuisibles, liés aux apports anthropiques, sont un sujet de plus en plus présent pour certaines régions de la Méditerranée [Karydis et Kitsiou 2012].

1.4 Hypothèses de comparaisons

Le développement du phytoplancton dans l'écosystème est impacté par de multiples facteurs (Hydroclimatique, physico chimique, et biologique) (Section 1.2.2). La variation d'une grande partie des facteurs dépend des forçages liés aux caractéristiques du milieu (géomorphologie de la zone, température, éclairement, activités touristiques, industrielles, ...). Les forçages propres aux 3 sites d'études (Section 1.3) peuvent être résumés de manière qualitative par le tableau 1.2. Bien que localisées dans des bassins différents, les stratégies d'implantations induisent des conditions environnementales en moyennes assez proches (zones côtières, proches d'un affluent, peu profondes, ...) (Tableau 1.2).

Tableau 1.2 – Récapitulatif des caractéristiques hydrodynamiques, physico-chimiques et biologiques de chaque zone. Pour chaque caractéristique est donnée une description qualitative (le détail et les taux sont décrits dans le chapitre).

Hydrodynamique	**Carnot**	**Iroise**	**MesuRho**
Façade Maritime	Mer du nord	Atlantique	Méditerranée
Bassin	détroit du Pas-de-Calais peu profond, ouvert	Rade de Brest peu profond, fermé	Golfe du Lion peu profond, fermé
Régime de marée	Mégatidale semi-diurne	Mégatidale semi-diurne	Microtidale semi-diurne à inégalité
Affluents	Seine, Somme + liane	Aulne, Elorn	Rhône
Stratification	morte eau	brassage relativement constant mais stratification en période de crue	période estivale, période de crue
Marnage vive eaux	9m	8m	50cm
Physicochimie			
Apport Nutritif	Zone Eutrophe	Zone Eutrophe	Zone Oligotrophe
Cond Éclairement	limitante en hiver	limitante en hiver	Non limitante
Remise en suspension	Vive eau, vent	constant (marré et morpho bassin)	Période hivernale
Turbidité	un peu	moyen	fort
Biologie			
Période productive	Mars-juin + Septembre	Mars-octobre	Nov-Avril
Risque HABs	oui	oui	non
Eutrophisation	oui	non	non

En effet les trois zones remplissent les conditions optimales de croissance phytoplanctonique. La dynamique fluviale et/ou le brassage des trois zones favorisent l'accumulation et la remise en suspension des sels nutritifs. Des taux de nutriments suffisants sont donc observés même en Méditerranée. Et les conditions météorologiques correspondent toutes à un milieu tempéré. De ce fait, les variations saisonnières du phytoplancton semblent assez similaires. Toutefois, il sera possible de noter des différences à des échelles plus précises, notamment au niveau communautaire ou encore des périodes de production (Section 1.3.1.3, 1.3.2.3, 1.3.3.3). Il faut aussi s'attendre à des différences au niveau de l'amplitude des blooms car même si les facteurs sont similaires, des différences au niveau des taux seront observées. Toutes ces différences sont dues à des variations des périodes d'éclairement et de stratification optimale, ou encore à des effets combinés des différents facteurs énumérés précédemment.

Ainsi, l'hypothèse suivante est émise : bien que la répartition géographique et les écosystèmes de chaque site d'études soient différents, des schémas communs pourront être identifiés par rapport à une dynamique des systèmes tempérés, mais une certaine variabilité de ces schémas, qui fait la particularité de chaque site, pourrait être observée à des échelles plus fines, par exemple au niveau des périodes et des nombres d'occurrences. De ce fait lors de la comparaison de ces différents

écosystèmes, il est important et intéressant de prendre en compte à la fois l'aspect global (au niveau des conditions et des principaux schémas) et l'aspect spécifique (au niveau des réponses HF).

Nous espérons donc retrouver ces deux aspects via les analyses et outils numérique, c'est-à-dire :

1. déterminer et cibler des événements communs liés aux forçages géomorphologiques (affluents, cycle biogéochimique, ...) ou encore liés à un biotope et biocénose semblables (variation saisonnière des blooms, ...)
2. mettre en évidence des événements particuliers propres à chaque site d'études liés à des caractéristiques physiologiques de certains taxons, à des assemblages taxonomiques différents ou encore à des facteurs locaux.

De plus, ces deux aspects vont permettre de considérer l'approche numérique à différentes échelles. L'aspect global va permettre une approche générique et faciliter l'inter-comparaison. L'aspect spécifique va, quant à lui, aider à évaluer le niveau de détail du modèle de classification non supervisée.

1.5 Conclusions

Le phytoplancton se compose d'organismes pélagiques, entraînés passivement par les mouvements d'eau. Par définition, il est donc très dépendant du milieu dans lequel il se trouve. Pour étudier la dynamique de ces organismes, il est donc primordial de prendre en compte les différentes pressions environnementales influençant sa variabilité et les différentes échelles spatio-temporelles associées. Ainsi, la synthèse des caractéristiques hydrologiques et biogéochimiques donne une vue d'ensemble des facteurs de pression propres à chaque zone et donc des points communs et/ou des différences de chacune. Elle permet d'identifier les principaux forçages et leurs impacts sur l'écosystème. Mais ce bilan révèle aussi toute la complexité d'étudier les variations (des pressions et des réponses) dans son ensemble, surtout en zone côtière où les facteurs sont multiples, changeants, inter-connectés et agissent de manière simultanée et collective.

Dans ce contexte, les méthodes d'analyses multivariées et multi-échelles sont essentielles à la compréhension du fonctionnement des écosystèmes marins (et notamment à la dynamique phytoplanctonique). De plus en plus utilisés et complexes, les systèmes d'apprentissages par les données sont donc privilégiés. La nature et la quantité des données fournies demeurent un point crucial à leur efficacité. Il conviendra de démontrer leur pertinence pour répondre à la définition et comparaison des schémas du phytoplancton de chaque zone.

Chapitre 2

Matériels et états de l'art méthodologique

2.1 Réseaux d'observations et système instrumenté

La surveillance des eaux est un enjeu capital pour mieux appréhender le milieu et son fonctionnement. Elle répond à des questions sur des sujets majeurs comme : l'évaluation et la gestion de la qualité des eaux, les suivis et la compréhension des écosystèmes marins, l'étude à long terme des cycles environnementaux, la protection et la gestion durable de l'environnement.

L'Ifremer opère à l'échelle du littoral métropolitain de nombreux réseaux de surveillance tels que les réseaux d'échantillonnages nationaux BF, c'est-à-dire avec un échantillonnage conventionnel de fréquence mensuelle à hebdomadaire, comme par exemple sur la Côte d'Opale : le Réseau d'observation et de Surveillance du Phytoplancton et de l'hydrologie dans les eaux littorales (REPHY) ; auxquels s'ajoutent des réseaux de surveillance régionaux complémentaires comme par exemple pour le REPHY : le Suivi Régional des Nutriments (SRN). Des réseaux de surveillance HF sont aussi mis en œuvre comme les stations de MAREL.

En complément de ces mesures *in-situ* sont mises en place des méthodes de mesures "indirectes" comme les mesures par satellite (*e.g.* concentrations en sels nutritifs, en chlorophylle a (Chl_a), ...) ou encore via des modèles numériques. Dans cette partie sont présentées les différentes sources de données exploitées pour répondre à la problématique de la thèse sur l'évolution de la dynamique du phytoplanctonique en zone côtière.

2.1.1 MAREL

2.1.1.1 Présentation du réseau

Le réseau MAREL est un réseau de surveillance automatisé HF du milieu marin côtier. Il est constitué de stations instrumentées autonomes qui effectuent des mesures *in-situ* HF et assurent la transmission des données en temps *quasi* réel. Il à été mis en place par l'Ifremer en 1992 pour répondre à un besoin de développer des systèmes de surveillance automatisée de l'environnement pour décrire les effets directs et indirects des activités humaines sur le milieu marin. Depuis peu, dans le cadre des projets de l'IR-ILICO, le réseau MAREL s'est structuré en un réseau national multi-organismes nommé COAST-HF.

Toutes les stations ont pour objectif commun la mesure à haute fréquence et de manière automatique des paramètres physico-chimiques essentiels de l'eau de mer ainsi que de quelques autres indicateurs caractéristiques. Ainsi, chaque station est équipée pour répondre à cet objectif

et a été adaptée en fonction des contraintes environnementales et des problématiques propres à chaque zone d'étude. Les stations de mesures sont donc équipées de différents dispositifs d'ancrage (Bouées, tubes supports, pontons flottants) et d'un panel de capteurs en fonction de la zone d'étude et des contraintes associées. Ainsi, ce réseau permet l'acquisition HF d'une dizaine de variables physico-chimiques, biologiques ou météorologiques.

Dans cette étude, trois systèmes instrumentés sont étudiés : le système MAREL-Carnot en Manche (Boulogne-sur-mer, en fonction depuis 2004), le système MAREL-Iroise en mer d'Iroise (Brest, depuis 2000), le système MesuRho en Méditerranées (Marseille, depuis 2009).

2.1.1.2 Stratégies d'échantillonnage

Station MAREL-Carnot : La station multi-capteur MAREL-Carnot est une station fixe constituée d'un tube (15 mètres de long et pesant 12 tonnes) contenant un flotteur sur lequel sont disposés les différents capteurs (Figure 2.1 A) [Schmitt et Lefebvre 2016]. Ce flotteur permet de suivre les oscillations de marées afin que les capteurs soient immergés à 1,5 mètre de profondeur en permanence, quel que soit le marnage.

Initialement, la station pompait l'eau en sub-surface et la redistribuait aux différents capteurs via la chambre de passage (Figure 2.1 B) avec une chloration de celle-ci lorsqu'il n'y avait pas de cycle de mesure. La chloration de l'eau de mer par électrolyse protégeait les capteurs contre l'encrassement biologique (*biofulling*). La station était donc équipée d'un automate de contrôle de mesure, d'une pompe de circulation pour la sonde, d'un débitmètre pour le contrôle de la pompe, d'un chlorateur pour la production de chlore par électrolyse, d'une sonde multi-paramètres (Type CTD) et d'un analyseur de nutriments Systea (N, P, Si).

Depuis 2014, afin de réaliser des mesures *in-situ*, le système de pompage a été remplacé par une sonde immergée multi-paramètres Mp6-NKE (Figure 2.1 C).

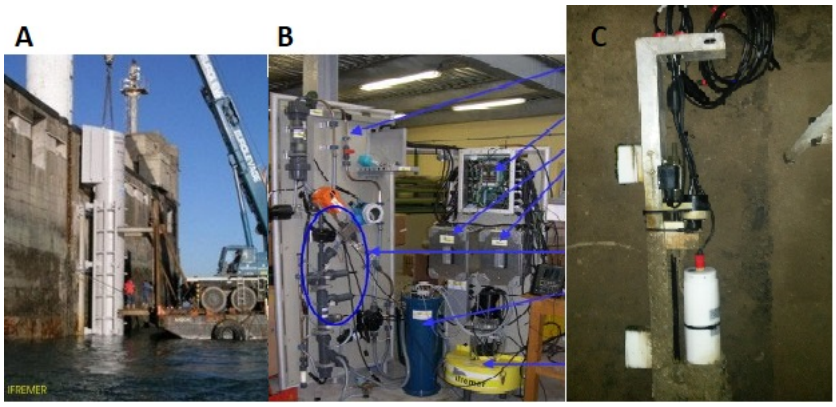


Figure 2.1 – Station MAREL-Carnot A : tube support de la station, B : ancien système de mesures hydraulique, C : Nouveau système de mesure (sonde immergée).

Actuellement, la station permet de mesurer 6 paramètres toutes les 20 minutes. En plus de ces mesures, certains paramètres tels que la salinité sont calculés. Les années comprises entre 2004 et 2014 intègrent le plus de capteurs, la base de données HF cumule 19 paramètres (Tableau 2.1). Cette base est accessible via la plateforme Coriolis depuis 2014 et elle est référencée dans le système *SEANOE (Sea scientific Open Data Edition)* depuis 2015 [Lefebvre 2015].

TABLEAU 2.1 – Descriptif des capteurs et des variables mesurées sur la plateforme MAREL-Carnot. dt : Fréquences d'échantillonage.

Paramètres atmosphériques	**Unités**	**Capteurs**	**dt**	**Maintenance**
PAR (surface)	$\mu mol\,m^{-2}\,s^{-1}$	NKE-MP6 : LI 190 SA	20 min	2004 - auj.
Paramètres océanique	**Unités**	**Capteurs**	**dt**	Maintenance
Température eau	°C	NKE-MP6 : Aandreaa	20 min	2004 - auj.
Conductivité (Salinité)	$mS\,cm^{-1}$ (PSU)	NKE-MP6 : WTW TeTracon 325	20 min	2004 - auj.
Profondeur (Hauteur d'eau)	m	Marégraphe : SHOM	20 min	2004 - auj.
Turbidité	NTU	NKE-MP6 : Seapoint	20 min	2004 - auj.
Oxygène dissous	$mg\,l^{-1}$	NKE-MP6 : Aandreaa	20 min	2004 - auj.
Fluorescense	FFU	NKE-MP6 : Seapoint Chl	20 min	2004 - auj.
PH		PONCPC-EH-10	20 min	2004 - auj.
Nitrate	$\mu mol\,l^{-1}$	SYSTEA	12h	2004 - 2014
Phosphate	$\mu mol\,l^{-1}$	SYSTEA	12h	2004 - 2014
Silicate	$\mu mol\,l^{-1}$	SYSTEA	12h	2004 - 2014

Station MAREL-Iroise : La station MAREL-Iroise est une bouée fixe qui se compose d'un flotteur hauturier de 4,5 mètres de diamètre et qui pèse 10 tonnes (Figure 2.2)[SCHMITT et LEFEBVRE 2016]. Ce flotteur permet la mise en place d'un module de mesures immergé et d'un module émergé. Le module immergé est équipé d'une sonde multi-paramètres (Type MP5-NKE) et d'un capteur CARbon Interface OCean-Atmosphere (*CARIOCA*). Le module émergé contient le capteur d'irradiance et un système automatisé de contrôle des mesures qui renvoie les données à une station informatique terrestre (Centre Ifremer-Brest) par le biais d'un antenne GSN/GPRS.

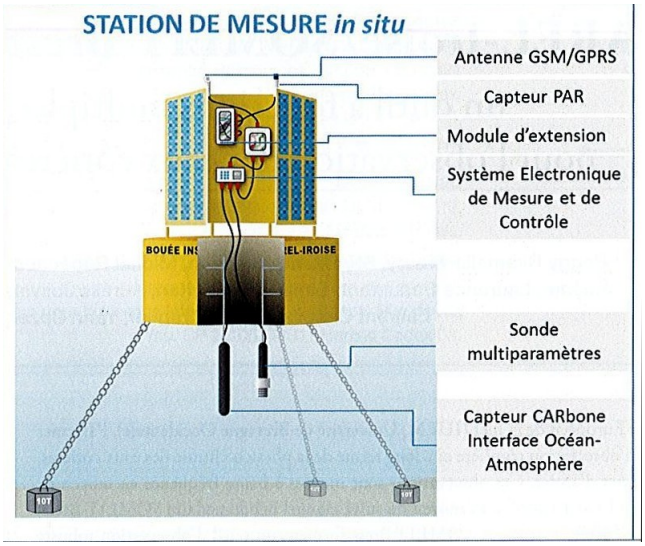


FIGURE 2.2 – Schéma de la plateforme MAREL-Iroise tiré de SCHMITT et LEFEBVRE 2016.

.

Déployée depuis 2000, la bouée MAREL-Iroise mesure ainsi 8 paramètres à HF, toutes les heures pour le dioxyde de Carbone et toutes les 20 minutes pour les autres paramètres (Tableau 2.2)

Tableau 2.2 – Descriptif des capteurs et des variables mesurées sur la plateforme MAREL-Iroise. dt : Fréquences d'échantillonage.

Paramètres Atmosphériques	**Unité**	**Capteurs**	**dt**	**Maintenance**
PAR (surface)	$\mu mol\, m^{-2}\, s^{-1}$	SerBird PAR-Sensor	20 min	2000 - auj.
Paramètres océaniques	**Unités**	**Capteurs**	**dt**	**Maintenance**
Température	°C	NKE-MP5		2000 - auj.
Conductivité (Salinité)	$mS\, cm^{-1}$ (PSU)	NKE-MP5	20 min	2000 - auj.
Oxygène dissous	$mg\, l^{-1}$	NKE-MP5	20 min	2000 - auj.
Turbidité	NTU	NKE-MP5	20 min	2000 - auj.
Fluorescence	FFU	NKE-MP5	20 min	2000 - auj.
CO_2 dissous	ppm	CARIOCA	20 min	2000 - auj.

Station MesuRho : La station MesuRho est une bouée de balisage rigide à flotteur Immergé (BFI) constituée d'une partie aérienne et d'une partie immergée (Figure 2.3)[Schmitt et Lefebvre 2016]. La partie aérienne assure l'alimentation et la liaison, ainsi que les mesures météorologiques. La partie immergée mesure, quant à elle, les différents paramètres physico-chimiques et biologiques en deux points de mesure en sub-surface (2-3 mètres) et au fond (18-19 mètres).

La station est équipée pour la partie aérienne : de panneaux solaires (pour l'alimentation), d'un automate ABIN (pour la liaison radiodiophonique), d'une station météorologique et d'un capteur de *PAR*. La partie immergée se compose de deux sondes multi-paramètres (type SMATCH-NKE), d'un *Acoustique doppler Courent Profiler (ADCP)*, d'un capteur de nitrate, d'une Sonde Température Pression Salinité (STPS) et d'une station benthique dotée d'une micro électrode à oxygène et de capteurs pour les études sur la reminéralisation du sédiment [Toussaint et al. 2014].

Mise en service en 2009, la station mesurait 6 paramètres atmosphériques et 8 paramètres océaniques (Tableau 2.3). Depuis 2011, la station benthique mesure l'oxygène dissous dans les sédiments et a été complétée d'un capteur de nitrate ainsi que de la sonde STPS pour mesurer la profondeur, la température et la conductivité à partir de laquelle est calculée la salinité. Pour des raisons de redondance avec un autre réseau de mesures quotidiennes sur la zone, le capteur de nitrate et la sonde STPS ont été retirés en 2013 [Schmitt et Lefebvre 2016]. Actuellement, la station mesure 6 paramètres atmosphériques et 9 paramètres océaniques (Tableau 2.3). Tous les paramètres sont mesurés à une fréquence de 30 minutes en moyenne.

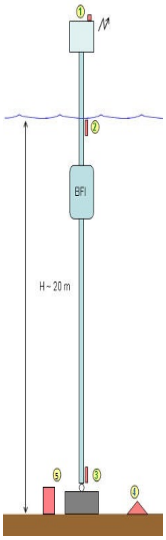


FIGURE 2.3 – Schéma de la platforme MesuRho. 1) Station météorologique, mesure d'irradiance. 2) Sonde multi-paramètres sub-surface (température, salinité, oxygène dissous, turbidité, Chlorophylle a). 3) Sonde multi-paramètres fond. 4) Courantomètre doppler (ADCP), mesure du profil de courant.

TABLEAU 2.3 – Descriptif des capteurs et des variables mesurées sur la plateforme MesuRho. dt : Fréquences d'échantillonage, Prof. : Profondeur de mesure.

Paramètres atmosphériques	**Unité**	**Prof.**	**Capteurs**	**dt**	**Maintenance**
Vitesse / Direction du vent	$m\,s^{-1}$ / °	10 m	Waisala WXT510	30 min	2009 - auj.
Pression atmosphérique	hPa	10 m	Waisala WXT510	30 min	2009 - auj.
Taux de précipitation	$mm\,j^{-1}$	10 m	Waisala WXT510	30 min	2009 - auj.
Température de l'air	°C	10 m	Waisala WXT510	30 min	2009 - auj.
Humidité relative	%	10 m	Waisala WXT510	30 min	2009 - auj.
PAR (surface)	$\mu mol\,m^{-2}\,s^{-1}$	10 m	Skye Quantum	30 min	2009 - auj.
Paramètres océaniques	**Unité**	**Prof.**	**Capteurs**	**dt**	**Maintenance**
Température	°C	3 m	NKE-SMATCH : VTW	30 min	2009 - auj.
Température	°C	20 m	RDI-600kHz	30 min	2009 - auj.
Conductivité (Salinité)	PSU	3 m - 20 m	NKE-SMATCH : VTW	30 min	2009 - auj.
Turbidité	NTU	3 m - 20 m	NKE-SMATCH : Seapoint	30 min	2009 - auj.
Oxygène dissous	$mg\,l^{-1}$	3 m	NKE-SMATCH : Anderaa	30 min	2009 - auj.
Oxygène dissous dans le sédiments (Station benthique)	$mg\,l^{-1}$	20 m	Micro-électrode à oxygène	20 min	2011 - auj.
Fluorescence Chl-a	$mg\,m^{-3}$	3 m - 20 m	NKE-SMATCH : Turner	30 min	2009 - auj.
Nitrate	$mg\,l^{-1}$	3 m	ISUS	30 min	2011- 2013
Courants	$m\,s^{-1}$	0,5 m - 20 m	RDI-600kHz	30 min	2009 - auj.
Pression	$m\,s^{-1}$	3 m - 20 m	NKE-SMATCH	30 min	2009 - auj.
Pression	bar	20 m	RDI-600kHz	30 min	2009 - auj.
Vagues à 20 m	m		RDI-600kHz	30 min	2009 - auj.

2.1.2 REPHY

2.1.2.1 Présentation du réseau

Le REPHY est un réseau national mis en place par l'Ifremer depuis 1984. Il a pour objectifs :

- d'estimer l'abondance et la composition taxonomique du phytoplancton des eaux côtières et lagunaires, afin de décrire la dynamique spatio-temporelle et de recenser les événements tels que les eaux colorées (ex. des marées rouges), les efflorescences exceptionnelles et les proliférations d'espèces toxiques ou nuisibles.
- de surveiller et d'alerter en cas d'évènements particuliers. Il détecte et suit plus particulièrement les espèces produisant des toxines destinées à la consommation et représentant un danger pour les consommateurs.

L'ensemble de l'échantillonnage est opéré par les 9 Laboratoires Environnement Ressources

(LERs) de l'Ifremer implantés sur douze sites le long du littoral français métropolitain. Ainsi, chaque laboratoire réalise l'échantillonnage, la conservation, le dénombrement phytoplanctonique, mais aussi les analyses physico-chimiques, la saisie des résultats dans la base de données nationales Quadrige2 ainsi que leurs valorisations et la coordination régionale.

En Manche Mer du Nord, l'échantillonnage est réalisé au niveau de trois écosystèmes caractéristiques de la manche orientale et de la baie de Somme : la radiale de Dunkerque, de Boulogne-sur-Mer et de la baie de Somme. Chaque radiale est construite selon un gradient côté-large.

Dans notre cas, c'est le point côtier de la radiale de Boulogne-sur-Mer (station BL1) qui est utilisé (Figure 2.4). C'est le point de prélèvement le plus proche de la station MAREL-Carnot.

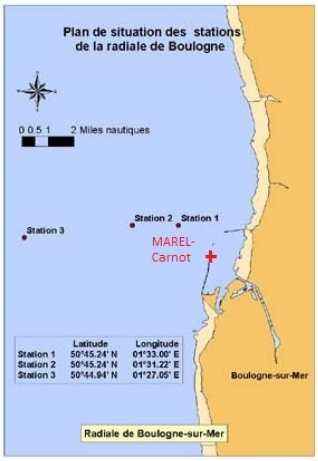


FIGURE 2.4 – Cartes des points de prélèvement de la radiale de Boulogne-sur-Mer.

2.1.2.2 Stratégies d'échantillonnage

REPHY est un réseau d'échantillonnage BF : de fréquences mensuelles à hebdomadaires. Les prélèvements se font généralement mensuellement. La fréquence d'échantillonnage devient bimensuelle entre les mois de mars et juin, période de forte production et de prolifération d'espèces toxiques comme *Phaeocystis globosa*.

Dans le cadre du REPHY, les observations phytoplanctoniques se font suivant trois protocoles : la flore totale (FLORTOT), la flore indicatrice (FLORIND) et la flore toxique (FLORPAR) (décrit dans le cahier des procédures du REPHY. Version 1. 2017). Seules les données de FLORTOT sont utilisé pour les analyses. C'est donc le seul protocole qui est décrit ci-dessous.

FLORTOT correspond à l'identification et le dénombrement de toutes les espèces phytoplanctoniques pouvant être identifiées dans les conditions d'observation, c'est-à-dire globalement toutes les espèces dont la taille est supérieure à 20 µm et celles dont la taille est inférieure mais qui sont en chaînes. Les mesures de FLORTOT sont obligatoirement accompagnées des mesures physico-chimiques suivantes : la salinité (PSU)), la température (°C), la turbidité (NTU), l'ammonium (NH_4, $\mu mol\,l^{-1}$), le nitrate (NO^{3-}, $\mu mol\,l^{-1}$), le nitrite (NO^{2-}, $\mu mol\,l^{-1}$), le phosphate (PO_4^{3-}, $\mu mol\,l^{-1}$), le silicate (SiO_4, $\mu mol\,l^{-1}$), les matières en suspension ($g\,l^{-1}$), la matière organique particulaire ($g\,l^{-1}$), la chlorophylle a ($\mu g\,l^{-1}$) et phéopigments.

Le protocole de prélèvement d'eau pour estimer l'abondance et la composition taxonomique suit le Cahier des Procédures du REPHY. Version 1. 2017 et le Manuel d'observation et de dénombrement du phytoplancton marin [Grossel, H. 2016].

Les principales étapes sont les suivantes :

— **Le prélèvement :** Il s'effectue en sub-surface (-1 m). Pour la Manche, il s'effectue de préférence en dehors de la zone d'estran, à pleine mer plus ou moins 2 heures.

— **L'échantillonnage :** Les échantillons d'un litre sont fixés sur le terrain avec une solution de lugol qui varie entre 1 et 10 ml en fonction de la densité algale. Les échantillons fixés sont conservés aux frais à l'abri de la lumière. Puis ils sont analysés dans les 4 semaines maximum qui suivent le prélèvement.

— **L'identification :** L'analyste identifie, à l'aide d'un microscope inversé [Utermöhl 1958], les cellules de taille supérieure à 20 µm et celles dont la taille est inférieure, mais qui sont en chaîne. Les espèces plus petites sont dénombrées seulement quand elles concernent des taxons nuisibles comme *Phaeocystis*, ou potentiellement toxiques comme *Chysochromulina*. Les dinoflagellés nanoplanctoniques et les cyanophycées sont aussi enregistrés.
Cette identification s'effectue à l'aide des manuels de taxinomie des microalgues [Moestrup, O. et al. 2009] et s'appuie sur le référentiel *World Register of Marine Species (WoRMS)* [WoRMS 2020]. L'identification doit se faire au plus bas niveau taxonomique (espèces ou genre), sinon, à un niveau taxonomique supérieur (genre, famille).

Le REPHY peut être complété par des réseaux de surveillance régionaux tels que le SRN. La dynamique et l'évolution de toutes les variables sont étudiées chaque année et retranscrites dans le rapport de mise en œuvre des réseaux REPHY [REPHY-SEANOE 2019] et SRN [Clabaut et al. 2019].

2.2 Les campagnes en mer

Plusieurs campagnes en mer ont été mises en place pour des stratégies de surveillance de la qualité marine ou des stocks de poissons. Celles-ci peuvent être récurrentes ou ponctuelles, sur des périodes clefs et offrent des séries de données souvent très complètes, car de faibles durées. Nous présentons ici uniquement celles utilisées dans ce manuscrit.

2.2.1 Présentation de la campagne DYPHYMA

La campagne de mesures continues du phytoplancton : DYPHYMA fait partie d'une des trois campagnes communes réalisées dans le cadre du projet de Développement d'un système d'observation DYnamique pour la détermination de la qualité des eaux MArines, basée sur l'analyse du PHYtoplancton (DYMAPHY, 2010-2014). Son objectif principal était d'améliorer l'évaluation de la qualité des eaux marines à travers l'étude du phytoplancton et de paramètres environnementaux associés [DYPHYMA 2012].

Lors de la campagne DYPHYMA réalisée sur le navire océanographique « Côtes de la Manche », des mesures à haute résolution spatiale et temporelle dans les eaux de surface ont été effectuées sur les deux côtés de la Manche orientale toutes les 1 à 10 minutes à partir d'un système « Pocket FerryBox ». L'abondance et la diversité des assemblages de phytoplancton ont été déterminées *in vivo* et *in situ* par le couplage de la fluorimétrie spectrale (Fluoroprobe, *Algae Online Analyser* (AOA)) et la cytométrie en flux (CytoSense) [*DYMAPHY Project* 2020]. Des prélèvements complémentaires sont réalisés pour des études en laboratoire.

2.2.2 Présentation de la campagne CGFS

La campagne CGFS (Channel Ground Fish Survey) [Coppin 1988] s'intègre dans le programme européen de suivi des ressources halieutiques, qui permet d'obtenir un ensemble de données relatives aux stocks exploités (maturité, structure en taille/âge, indices de recrutement). La série temporelle initiée en 1988 est utilisée chaque année par les groupes européens d'évaluation des stocks qui déduisent l'état de santé des principales espèces commerciales. Réalisée sur le N/O Thalassa, la campagne CGFS permet également un échantillonnage et une meilleure connaissance de l'ensemble de l'écosystème, répondant à la fois aux demandes de suivi des écosystèmes (DCSMM) et à la mise en place d'une approche écosystémique des pêches au niveau communautaire. Ainsi, les caractéristiques physico-chimiques de l'eau, les communautés de phytoplancton et zooplancton, l'abondance d'œufs de poissons, la composition spécifique des communautés nectoniques sont mesurées et analysées tout au long de la campagne [Coppin et Travers -Trolet 2017].

Depuis 2017, un FerryBox est installé sur le N/O Thalassa. Il a été utilisé pour la première fois lors des campagnes halieutiques CGFS 2017 (06/10/2017-23/10/2017) et plus tard pour IBTS 2018 (15/01/2018-12/02/2018). Il est utilisé dans les campagnes halieutiques en tant que support aux suivis et à la surveillance actuelle. Il apporte une information complémentaire haute fréquence et spatialisée.

2.2.3 Stratégies d'échantillonnage : (Pocket) FerryBox

Le FerryBox et le Pocket FerryBox sont des appareils de mesures HF autonomes. Ces dispositifs sont mis en place sur des navires de recherche ou des navires d'opportunité afin de compléter les données *in-situ* BF. Le Pocket FerryBox [Schroeder et al. 2008] est plus petit qu'un FerryBox mais possède les mêmes fonctionnalités. Sa petite taille facilite son installation et permet de le déployer sur des navires dont l'espace est plus restreint. Ils sont quasi-autonomes et une grande partie des fonctionnalités peuvent se faire à distance sauf pour le nettoyage des filtres d'eau de mer. Ce nettoyage doit se faire quotidiennement pour éviter le colmatage et la baisse de performance [Lefebvre et Devreker 2019a].

Ils mesurent toutes les minutes une dizaine de paramètres tels que la température et la salinité, le pH, la turbidité, la concentration en oxygène et des classes spectrales de phytoplancton (via le module AOA). Il est possible d'ajouter des capteurs supplémentaires qui mesurent la pression du CO_2, la concentration en nutriments, ...Toutes les informations de maintenance et les caractéristiques des sondes sont détaillées dans le guide technique rédigé par Crenan 2019.

Le FerryBox est un dispositif particulièrement intéressant car, via le fluorimètre spectral (module AOA), il mesure en continu et en temps réel la fluorescence chlorophyllienne des micro-algues. Bien que moins précise, l'approche spectrale offre l'avantage de produire rapidement des résultats contrairement à la méthode conventionnelle par microscopie. Elle permet toutefois d'avoir une approche taxonomique préliminaire en caractérisant 4 groupements spectraux : les vertes (Chlorophycées), les bleues-vertes (Cyanoph ycées), les rouges (Rhodophytes) et les jaune-marron contenant entre autres les Cryptophytes, les Bacillariophytes ou Diatomées, les Haptophytes, les Prymnésiophycées telles *Phaeocystis sp.*, les coccolithophoridés et les Dynophytes (dinoflagellés).

L'AOA se base sur les propriétés de fluorescence chlorophyllienne du phytoplancton pour différencier les groupements [Beutler, Wiltshire et al. 2002]. En effet, chaque classe d'algues a sa propre « empreinte » liée à sa pigmentation, qui va correspondre à un schéma spécifique en fonction de sa réponse aux différentes longueurs d'ondes d'excitation (450, 525, 570, 590 et 610 nm) (Figure 2.5) [Lefebvre et Devreker 2019a]. Le spectre de fluorescence du mélange d'algues est collecté par un photomultiplicateur. Ensuite, une procédure arithmétique statistique [Beutler, Wiltshire et al. 2002 ; Ruser, A. et al. 1999] permet de distinguer les 4 groupements d'algues.

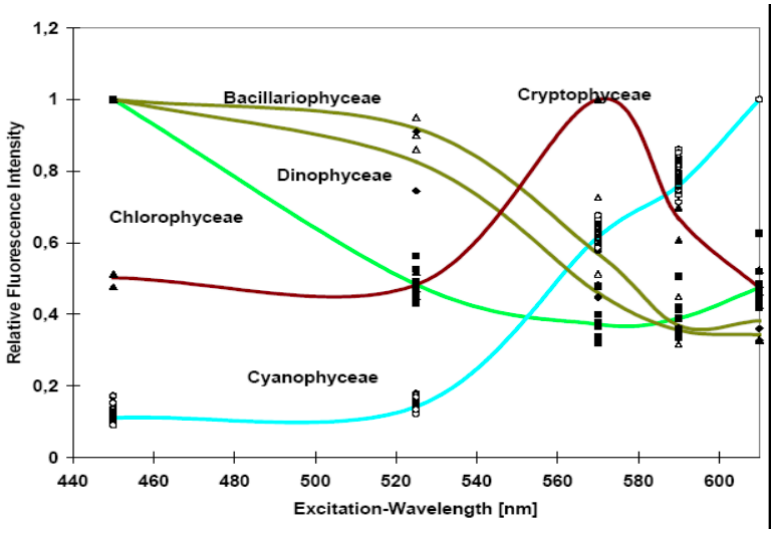


FIGURE 2.5 – Intensité en Fluorescence des différentes empreintes des groupes phytoplanctoniques en fonction de la longueur d'onde ©BBE.

2.3 Traitement statistique : Analyse descriptive des données

Afin d'établir une typologie des classes mises en évidence dans cette étude, il est primordial de définir les caractéristiques de chacune. Il est donc important d'utiliser des outils de description et d'analyse. Avant tout, les outils de description sont essentiels afin de mieux comprendre : les facteurs de contrôle, condition d'initiation et de terminaison, les variations de l'amplitude, les potentielles périodicités (Section 2.3.1). Ensuite, relier ces états à des schémas de fonctionnement des efflorescences phytoplanctoniques, à des stratégies de vie ou encore des assemblages, nécessite de disposer d'indicateur de biodiversité (Section 2.3.2). Enfin, une vision d'ensemble est recherchée dans cette étude, il est donc indispensable de se servir d'outils d'analyses multivariées (Section 2.3.3)

2.3.1 Analyses uni-variées du signal

Nous développons ici uniquement les approches couramment utilisées dans la littérature et adaptées à nos signaux et reprises dans nos analyses. En effet, la présence d'une quantité importante de données manquantes sur de longues périodes peut fausser certaines analyses comme la décomposition tendance/saison/cycle/résidu.

2.3.1.1 Décomposition modale empirique (EMD)

La représentation temporelle des signaux ne montre pas toujours les caractéristiques périodiques relatives au signal. Les décompositions spectrales peuvent mettre en évidence des périodicités. La décomposition modale empirique ou *Empirical Mode Decomposition (EMD)* décompose un signal en une somme de modes, sans quitter le domaine temporel.

L'EMD est une méthode de décomposition adaptative d'analyse de données [HUANG et al. 1998]. Elle a été choisie dans notre cas d'étude, car elle est adaptée aux données de processus non linéaires et non stationnaires comme c'est le cas généralement pour les signaux biologiques. Au lieu d'analyser le signal dans une base fixe comme avec Fourier, EMD construit au fur et à

mesure les fonctions de base appelées Fonction Mode Intrinsèque ou *Intrinsic Mode Function (IMF)* (Définition 1).

Définition 1 *Une IMF est une fonction oscillante de moyenne nulle, c'est-à-dire une fonction :*

1. *dont le nombre d'extrema et le nombre de passages par zéro diffèrent d'au plus un.*
2. *dont la moyenne entre les enveloppes supérieures et inférieures, formées à partir des maxima et minima locaux, au sens de la définition précédente, est nulle en tout point.*

L'EMD suppose que tout signal réel (s) se décompose en une moyenne locale (m) et une composante plus oscillante (d). On a ainsi :

$$s = d_1 + m \tag{2.1}$$

l'EMD définit s en une somme finie de modes oscillant d_k soit les IMFs.

$$s = \sum_{k=1}^{N} d_k + r \tag{2.2}$$

r étant le résidu, soit l'oscillation de plus basse fréquence (comprenant plus ou moins 3 extrema).

Ainsi l'EMD produit différentes échelles/modes de la série temporelle d'origine ayant un sens physique. Toutefois, cette méthode est sensible aux données manquantes et à cela s'ajoute le problème de mélange des modes.

2.3.2 Analyses uni-variées de la biodiversité

2.3.2.1 Unité Taxonomique

Dans le cadre des campagnes REPHY un grand nombre d'espèces phytoplanctoniques sont dénombrées à partir des analyses microscopiques (Section 2.1.2). En Manche orientale au point de mesure BL1 (Section 2.1.2) 199 taxons sont identifiés. Selon la même procédure dans Karasiewicz et al. 2018, il a été choisi de regrouper ces 199 taxons en 96 unités taxonomiques opérationnelles (en anglais *Operational Taxonomic Unit* (FLORPAR) (Annexe A). Une FLORPAR est un regroupement d'individus phylogénétiquement proches. Dans le cas présent, les unités taxonomiques ont été définies comme un groupe d'espèces qui appartiennent au même genre suivant leurs similitudes morphologiques. Cette classification correspond en grande partie à la nomenclature qui était utilisée avant que des études génétiques les séparent en groupes distincts. De plus nous avons pris soin que vérifier que les espèces regroupées aient bien des périodes d'apparitions identiques ou successives. Ainsi même si cette identification est moins précise elle est suffisamment pertinente. L'utilisation des unités taxonomiques permet de réduire considérablement le nombre de taxons ce qui facilite les analyses. Elle permet aussi de réduire le biais d'identification lié aux changements de taxonomistes et à leurs différences de formations et assure une homogénéité au niveau du traitement des données.

2.3.2.2 Abondance

L'analyse de la biodiversité a été réalisée à partir de la notion fondamentale de l'abondance. L'abondance d'une espèce (S) est le nombre total de cette espèce. Pour cette étude, les assemblages phytoplanctoniques de chacune des classes sont définis à partir des critères d'abondance relative et d'abondance cumulée.

L'abondance relative représente le nombre d'individus par unité d'espace pour une espèce donnée par rapport au nombre total d'individus toutes espèces confondues. L'abondance cumulée, ou abondance sommée, est la somme des abondances de plusieurs espèces ou groupes d'espèces.

Ainsi, les espèces dominantes (Esp. Dom.) sont définies comme toutes les espèces dont l'abondance relative cumulée est supérieure ou égale à 0,95 % de l'abondance totale (Abd. 95 %).

2.3.2.3 indice de Shanon

Un autre critère utilisé dans cette étude pour caractériser la diversité dans chacune des classes est l'indice de Shannon (H'). H' permet de mesurer la diversité spécifique du milieu [Shannon 1948]. Il donne des informations sur le nombre d'espèces de ce milieu (richesse spécifique) et sur la répartition des individus au sein de ces espèces (équitabilité spécifique).

$$H' = -\sum_{i=1}^{S} p_i * \log_2 p_i \tag{2.3}$$

Avec

i une espèce du milieu d'étude,

p_i la proportion d'une espèce i par rapport au nombre total d'espèces (S) dans le milieu d'étude (ou richesse spécifique du milieu)

Il va permettre de quantifier la complexité des assemblages phytoplanctoniques et son hétérogénéité. En effet plus l'indice sera faible et plus la diversité ou le nombre d'espèces dominantes sera faible.

2.3.2.4 Diagramme de Margalef

Margalef [Margalef 1978] a développé un modèle de succession d'espèces, connu sous le nom de "mandala", qui rend compte de l'équilibre entre les forces physiques et les forces nutritionnelles (Figure 2.6a). Les gradients de turbulence et les concentrations en sels nutritifs sont utilisés pour construire et définir la répartition des différents groupes phytoplanctoniques.

Margalef (1978) met en évidence une évolution entre les groupes ayant une stratégie r et K : les stratèges r (e.g. Diatomées) sont des espèces, avec une reproduction rapide, favorisées par des environnements instables (comme une turbulence et des concentrations en éléments nutritifs relativement élevés), alors que les organismes de stratégies K (e.g. Dinoflagellés) ont une reproduction plus lente et dominent les environnements plus stables.

Dans la continuité de Margalef, Reynolds établit un schéma de succession de stratégies de vie C-R-S pour le phytoplancton d'eau douce [Reynolds 1995], qui sera appliqué plus tard aux assemblages marins [Smayda et Reynolds 2001] (Figure 2.6b).

Le schéma C-R-S définit des phases de transition entre les espèces colonialistes-invasives (C), tolérantes au stress (S) et rudérales (R). Les espèces avec la stratégie C sont caractérisées par des organismes de petite taille, à fort taux de croissance, un rapport surface/volume élevé (S/V).Ces espèces sont favorisées par des milieux avec un mélange stable et riche en sels nutritifs.

Les espèces à stratégie S sont des organismes de grandes tailles avec un rapport S/V faible, à croissance lente et à forte préférence pour la lumière. Ils sont présents lorsque le mélange de la colonne d'eau et le taux en sels nutritifs sont faibles. Ces conditions sont favorables aux organismes ayant des moyens d'acquisitions d'éléments nutritifs alternatifs, par exemple la phagotrophie, les migrations verticales, des méthodes de fixation de l'azote plus efficaces.

Les espèces de stratégie R sont des organismes de grandes tailles (généralement de cellules allongées) et qui en dépit de leurs dimensions ont un rapport S/V élevé. Ces espèces sont peu tolérantes à la lumière et adaptées à des environnements bien mélangés.

Au cours des dernières décennies, plusieurs études ont enrichi cette théorie et apporté de nouveaux éléments. Par exemple, Cullen et al. fournit une revisite qui inclut le picoplancton et met l'accent sur les caractéristiques d'adaptations du phytoplancton et leurs relations avec le réseau trophique (Figure 2.7a) [J. J. Cullen et al. 2002 ; J. Cullen et al. 2007]. Tandis qu'Allen et Polimene (2011) ont suggéré l'ajout d'un gradient d'activité réactive de l'oxygène qui fournit un lien important entre la physiologie et l'écologie [Allen et Polimene 2011]. Pour aller encore plus loin, Gilbert (2016) établit un nouveau mandala défini par douze traits de réponse, d'effets ou de caractéristiques environnementales, liées aux différents types fonctionnels du phytoplancton [Glibert 2016] (Figure 2.7b). Cette étude fournit un schéma conceptuel qui intègre de nouveaux facteurs comme la lumière, les changements anthropiques et les rapports de nutriments. Elle met en évidence les liens entre des traits fonctionnels et les conditions environnementales.

Ainsi, dans le même esprit que les travaux présentés précédemment, nous proposons un nouveau diagramme d'interprétation (Figure 2.8).

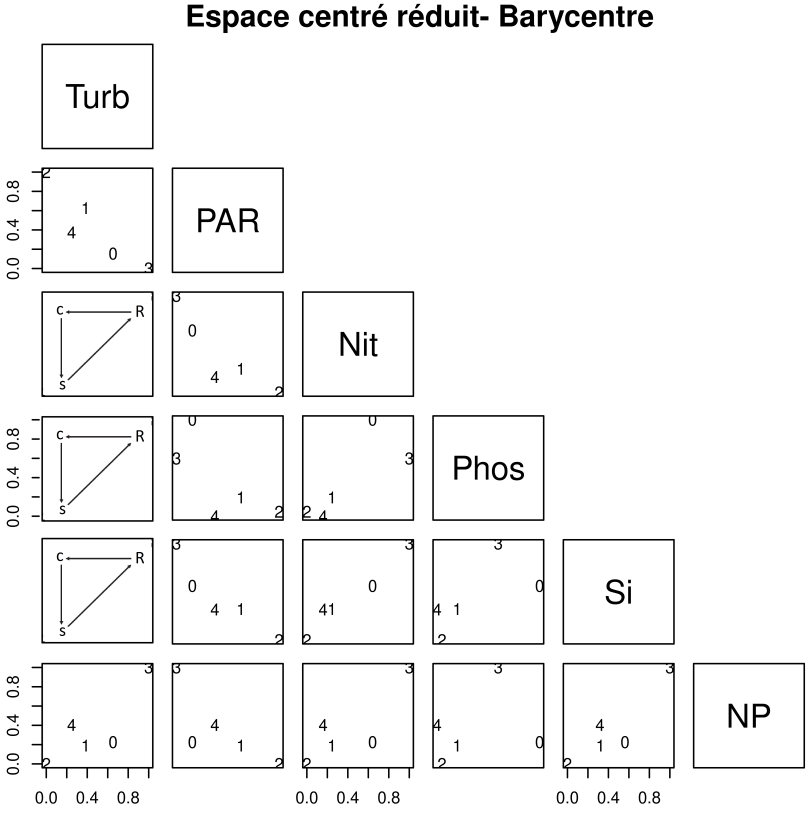
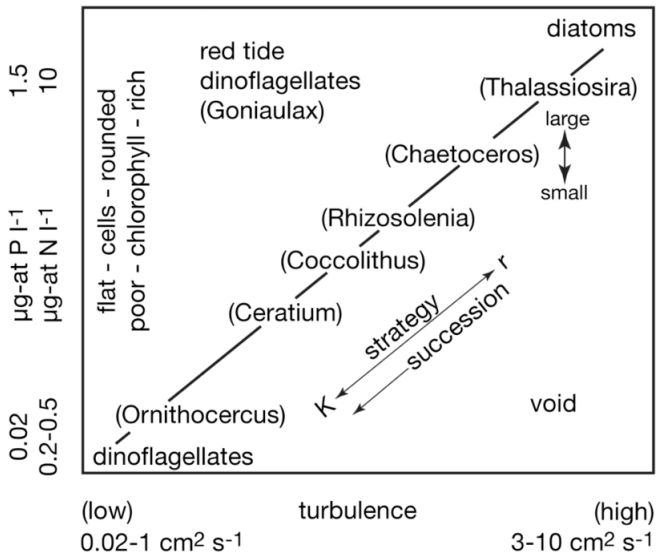


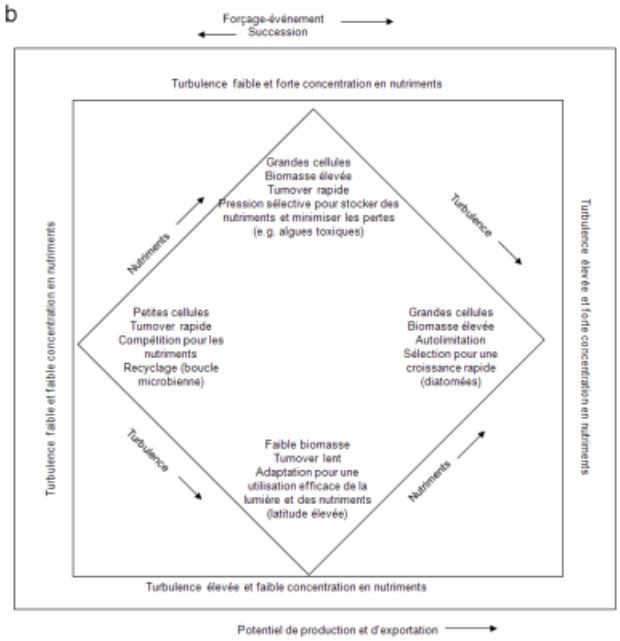
Figure 2.8 – Diagramme de Margalef étendu à 6 variables -Turbidité (Turb), PAR, Nitrate (Nit), Phosphate (Phos), Silicate (Sil), Rapport Nitrate sur Phosphate (NP) proposé pour l'interprétation des schémas de fonctionnement. Exemple pour 4 classes numéroté de 1 à 4 ; 0 correspond au données non classées. Pour



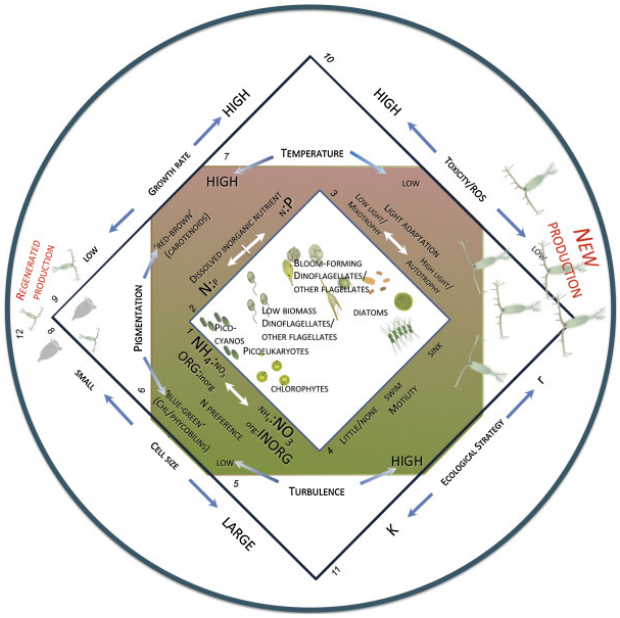
(a) "Mandala" de Margalef 1978

(b) "intaglio" de Reynolds, 2001

Figure 2.6 – Modèles d'organisation du phytoplancton. a) Schémas de succession des groupes phytoplanctoniques ("Mandala") suivant les conditions variables des facteurs environnementaux (d'après Margalef 1978). b) Schéma de succession saisonnière des stratégies de vie (C-R-S) pour le phytoplancton suivant la disponibilité en sels nutritifs et le niveau de mélange dans la colonne d'eau. Ce schéma est issu de Phlips et al. 2006 d'après Smayda et Reynolds 2001.



(a) Mandala de J. J. Cullen et al. 2002



(b) Mandala de Glibert 2016

Figure 2.7 – Modèles d'organisation du phytoplancton de Margalef revisité. a) Modèle de J. J. Cullen et al. 2002 mettant en évidence les caractéristiques et adaptations des assemblages du phytoplancton. b) Modèle de Glibert 2016 affichant les types fonctionnels du phytoplancton via 12 axes (représentés par les petits nombres dans le coin de chaque axe).

Ce diagramme, Figure 2.8, permet de visualiser la disponibilité en sels nutritifs, la quantité de lumière dans le milieu et les degrés de mélange en fonction des 6 variables explicatives : Turbidité (Turb), *PAR*, Nitrate (Nit), Phosphate (Phos), Silicate (Si) et le rapport N/P (NP). Ici, nous proposons une visualisation normée où chaque classe est représentée par le barycentre normalisé de l'ensemble des points qui le compose.

2.3.3 Analyses multivariées

2.3.3.1 Matrice de corrélation de Paerson

Des matrices de corrélation sont calculées afin de définir les covariations entre les variables et les classes. Le coefficient de Pearson (r) est choisi ici pour sa normalisation (Equation 2.4). Ainsi, nous sommes assurés que les résultats ne varient pas suivant les unités de mesure choisies.

$$cor_{X,Y} = r_{X,Y} = \frac{cov_{X,Y}}{\sqrt{S_X^2.S_Y^2}} = fraccov_{X,Y}\sqrt{S_X^2}.\sqrt{S_Y^2} = \frac{cov_{X,Y}}{S_X.S_Y} \tag{2.4}$$

Où S est la variance de la variable et cov la covariance entre deux variables

$$var_X = S_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} \tag{2.5}$$

$$cov_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{X}).(y_i - \bar{Y})}{n-1} \tag{2.6}$$

2.3.3.2 Analyse en composante principale (ACP)

L'analyse en composantes principales (ACP) est appliquée sur toutes les bases de données afin de simplifier la description de celles-ci.

L'ACP est une analyse factorielle multivariée qui permet de décrire le jeu de données par plusieurs facteurs ou composantes principales (Axes de l'ACP). Ces axes factoriels sont définis à partir de la variance du jeu de données. Chaque axe de $\mathbb{R}^n$ correspond à une combinaison linéaire des variables originelles et ils sont orthogonaux 2 à 2. Ils sont construits de sorte à maximiser la variance des données. Les cercles de corrélations représentent les projections des variables sur les plans formés par ces axes. Ainsi, les variables projetées de façons orthogonales seront non corrélées, dans des directions opposées seront corrélées négativement et dans la même direction seront corrélées positivement (Figure 2.9).

l'ACP réduit donc les dimensions d'une base de données multivariée $N \times M$ à $P \leq M$ composantes principales et permet une visualisation graphique en perdant le moins possible d'information. Chaque axe, composante, représente une part d'inertie expliquée des données. P est alors défini par l'expert selon un seuil minimal à atteindre de cette part cumulée. Ainsi des processus généraux peuvent être identifiés, ils régissent la répartition des variables qui sont plus ou moins corrélées entre elles.

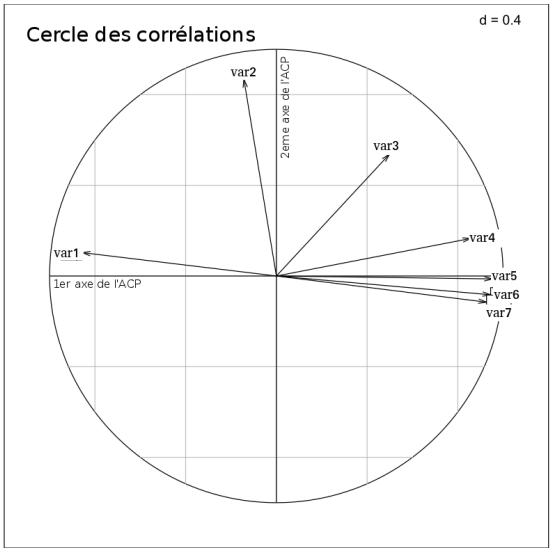


FIGURE 2.9 – Schéma du cercle de corrélation tracé pour le premier plan factoriel de l'ACP.

2.4 Méthodes d'apprentissage pour la segmentation et la prédiction dans les séries temporelles

L'un des objectifs majeurs de cette thèse est la caractérisation de schémas fonctionnels des efflorescences (facteurs de contrôle, conditions d'initiation, de terminaison et de contrôle de l'amplitude) dans les séries de données pour aider à la compréhension et à l'interprétation du fonctionnement des écosystèmes. Ainsi, nous cherchons des outils capables de séparer les données en différents groupes sans aucune connaissance *a priori* de leurs nombres, leurs formes ou leurs distributions.

Il existe de nombreuses méthodes de segmentation. Elles peuvent être soit génératives et le découpage sera alors basé sur des probabilités conditionnelles et des règles d'appartenance, soit discriminantes et il sera déterminé à partir de fonctions/règles de séparation. Ainsi, elles sont souvent différenciées par la manière dont elles découpent les bases de données.

Pour la segmentation des séries temporelles, on distingue, dans la littérature, différents moyens de segmenter comme le traitement des fenêtres (fenêtres temporelles ou région spatiale) [VAN HOAN et al. 2017 ; LÄNGKVIST et al. 2014], par des modèles génératifs sur la base d'algorithme espérance-maximisation (*Expectation Maximization* (EM)) [EMONET et al. 2014 ; DIAS et al. 2015 ; POISSON-CAILLAULT et LEFEBVRE 2017], soit par des coupes temporelles ou géométriques comme points de rupture [JAMES et MATTESON 2015], TSINASLANIDIS et KUGIUMTZIS 2014 ou les méthodes de classification.

Les trois premières approches ont besoin d'hypothèses sur la distribution des données et la taille du modèle. Or ces informations ne sont pas disponibles. De plus, ces méthodes ne peuvent pas être appliquées directement sur des bases de données volumineuses, multivariées, non linéaires et présentant une variabilité élevée comme dans notre cas d'étude. Nous nous sommes donc tournés vers un dernier type de méthode de segmentation : les méthodes de classification supervisée et non supervisée. Ces méthodes de classification offrent plusieurs avantages : premièrement, elles sont rarement contraintes par la taille des matrices de données et vont donc pouvoir traiter les volumineuses bases de données HF. Deuxièmement, ce sont des méthodes multi-dimensionnelles

ce qui va permettre de réaliser des études multivariées et donc, de prendre en considération les multiples impacts des facteurs environnementaux sur la distribution. Troisièmement, elles vont permettre le traitement de bases de données multi-échelles car elles sont capables de traiter des bases de données normalisées et donc adimensionnelles.

Une fois les schémas de fonctionnement identifiés, il est intéressant de pouvoir prédire ces schémas, afin d'adapter nos stratégies d'études, d'alertes en cas de dysfonctionnements ou d'évènements néfastes.

Dans la littérature, pour prédire le label d'une nouvelle observation, la comparaison à un modèle génératif ou l'introduction dans un modèle discriminant co-existe. Ce choix de modèle dépend fortement de la distribution, du volume des données et de la séparabilité des classes. Nous ne nous attarderons pas sur ces approches que nous utilisons et que nous n'avons pas étendues contrairement à la partie non supervisée où il paraît important de définir les approches intégrées à notre système global de segmentation.

2.4.1 Notations et concepts généraux

L'apprentissage artificiel (*Machine Learning*) a pour objectif de comprendre la structure des données et de les intégrer dans des modèles génératifs ou discriminants à partir de données. On différencie trois types de méthodes : le classement ou classification supervisée (*Classification*), la classification non supervisée (*Clustering*) et la régression (*Regression*) (Section 3).

Ainsi dans cette partie, nous utiliserons indifféremment le terme *clustering* ou classification non supervisée pour désigner le processus qui consiste à labelliser une donnée sans *a priori* sur son appartenance à un groupe donné par opposition à la classification supervisée basée sur un ensemble d'apprentissage (données, labels connus). Les anglicismes "Labellisation" et "label" (notés l_i pour le i^e label) seront employés comme termes analogues à étiquetage et étiquette. De même, les mots classe et *clusters* sont employés indifféremment pour désigner une structure découverte lors de la classification par opposition au concept de label ou état écologique désignant une classe étiquetée (labellisée après expertise) et/ou apprise.

Notations :
Nous désignerons pour la suite du document les notations suivantes

- X la matrice de données $N \times M$, avec N le nombre d'objets de la base et M le nombre de dimensions formant l'espace de description des objets ;
- $\vec{x_n} = \{x_{n1}\}, \ldots, x_{nM}$ pour $n \in [1, N]$ le n-ième objet de la base ;
- K le nombre de groupes identifiés après segmentation automatique par classification non supervisée ;
- $k \in 1, \ldots, K$ le numéro du k-ième groupe ;
- $cl_n \in 1, \ldots, K$ le numéro de classe de l'objet n ;
- l_n le label du l'objet n après labellisation des K groupes par un expert qui sera utilisé pour un apprentissage supervisé.

2.4.2 Méthodes usuelles de classification non-supervisé : labellisation des données par *clustering*

Un grand nombre de méthodes de classification non supervisées existant dans la littérature ; aucune méthode ne peut être considérée meilleure, elles dépendent fortement de l'application et des données. Chacune possède des caractéristiques propres qui lui permettent d'être adapté ou non à une problématique. Il est donc important de considérer chaque méthode en fonction du

cas d'étude. Les caractéristiques principales de chacune ont été résumées dans le tableau 2.4. Ce tableau va nous permettre de distinguer rapidement les différences de chaque méthode évoquée et d'avoir une vue globale pour définir les avantages et inconvénients de la méthode relatifs aux problématiques de l'étude.

Tableau 2.4 – Caractéristiques des méthodes des *clustering* et les définitions associées.

Caractéristiques	**Valeurs**	**Valeurs souhaitées**
Prise en compte du contexte	Supervisé ou non supervisé	non supervisé
Connaissances a priori	quantité d'information sur le problème à avoir	à minimiser
Méthode	Type de méthode	*
Présentation des résultats	Mise en relation des sorties	K classe
Complexité		*
Déterminisme	oui=Définition d'une (ou plusieurs) loi(s) nécessaire(s) à la segmentation	oui
incrémental	oui= classification incrémentielle traite les données un élément à la fois	oui
Hard vs soft	Hard : chaque objet appartient à un cluster ou non vs Soft : chaque objet appartient à chaque groupe à un certain degré	Hard
Tolérance au bruit	Bruit : signaux parasites qui viennent se superposer au signal dit utile, oui = Tolèrent	oui
Effet de chaîne	Des clusters proches mais distincts peuvent être fusionnés s'il existe une chaîne d'objets qui les relie	non
Tolérance aux clusters de tailles variées	oui = Tolèrent	oui
Tolérance aux clusters de densités variées	oui = Tolèrent	oui
Tolérance aux clusters concentriques	oui = concentrique (qui a un même centre)	oui
Tolérance aux clusters convexe et non linéairement séparable	oui = Tolèrent	oui

Dans cette thèse, le choix a été fait de développer notre propre algorithme (Chapitre 3) afin d'être aux plus proches de nos objectifs. Cet algorithme combine les concepts de trois méthodes de classification non supervisée usuelles : le *Hierarchical Clustering* (*HC*), la segmentation *K-means*, *Spectral Clustering* (*SC*). Elles ont été sélectionnées, car leurs caractéristiques semblaient importantes ou judicieuses à tester. Elles sont ainsi employées comme base des outils développés, mais aussi comme support de comparaison.

Dans cette section, il est décrit le principe général de chaque méthode. Ensuite, les atouts et les limites sont exposés. Enfin, les principales caractéristiques sont résumées (Tableau 2.4).

2.4.2.1 Clustering Hiérarchique

Le principe de la classification hiérarchique noté *Hierarchical clustering* (*HC*) en anglais (Algorithme 1, Package stats : :hclust) [Borcard et al. 2011] est de chercher à obtenir une hiérarchie, c'est-à-dire une collection de groupes d'observations en utilisant la relation d'ordre de finesse entre les partitions (Définition 2) et l'incertitude intra-classe. Les partitions sont définies en fonction d'une métrique telle qu'un indice de dissimilarité ou un critère d'agrégation. Cette méthode construit un arbre, qui regroupe au fur et à mesure des objets considérés comme similaires (Définition 3).

Soit un ensemble $X = \{x_i\}$ de M variables de N points. Il est possible de définir une hiérarchie sur X de deux manières :

Définition 2 *Une hiérarchie $H = \{h_i\}$ sur X est une chaîne de partitions de X : $P = \{P_1, \ldots, P_N\}$ dont la moins fine est $P_1 = \{X\}$ et la plus fine est $P_n = \{p_1, \ldots, p_r\}$.*

Une chaîne de partitions étant définie comme un ensemble de partitions $\{p_1, \ldots, p_r\}$ contenant une ou plusieurs éléments de la hiérarchie H soit $p_i = \{h_i\}$.

Définition 3 *Une hiérarchie $H = \{h_i\}$ sur X est un sous-ensemble des parties de X tel que :*
- *Pour tout élément x de $X, \{x\} \in H$;*
- *Pour tout couple d'éléments h, h' de H avec $h \neq h'$, on a :*
  - *Soit $h_i \cap h_j = \emptyset$,*
  - *Soit $h_i \cap h_j \neq \emptyset$, alors soit $h_i \subset h_j$, soit $h_j \subset h_i$.*

Deux méthodes existent pour définir les sous-ensembles ($\{h_i\}$) lors de la construction d'une hiérarchie (Figure 2.10) :

- La méthode ascendante : crée une partie en regroupant deux parties existantes. Il débute avec autant de clusters que d'objets initiaux et fusionne successivement les clusters les plus similaires.
- La méthode descendante : divise au contraire une partie existante pour en faire deux nouvelles. Il débute avec un seul cluster puis divise successivement le cluster pour obtenir la séparation la plus distincte possible.

Que ce soit avec l'approche ascendante ou descendante, elles ont besoin pour définir un critère d'agrégation ou de division de connaître la distance ($S(X) = \Delta(x, y)$) entre les points ou les centroïdes de 2 clusters. Il existe différentes manières de considérer une distance entre 2 clusters :

- *Single-link*, lien minimum choisit de maximiser la similarité entre les objets des différents sous-ensembles ($\{h_i\}$) soit minimiser la distance inter-cluster. Le problème principal de cette méthode est que cela peut créer un effet de "chaîne" c'est-à-dire que des clusters proches, mais distincts peuvent être fusionnés s'il existe une chaîne d'objets qui les relie.
- *Complete-link*, lien maximum est l'opposé du single link. Il minimise la similarité entre les objets des sous-ensembles ($\{h_i\}$) et donc maximise la distance inter-cluster. Cette méthode permet, ainsi, d'éviter l'effet de "chaîne".
- *Average-link*, lien moyen est une méthode intermédiaire. Le principe est de moyenner des distances entre toutes les paires de points possibles d'un sous-ensemble (h_i) à l'autre (h_j).
- *Centroïde-link*, lien centroïdal considère la distance entre les centroïdes de deux clusters.
- *Ward indexe*, indice de Ward [Ward 1963] tient compte de la variance des classes, en plus de la distance euclidienne entre les centres de gravité. Ce qui permet de limiter l'effet de

chaîne. Tout comme le lien minimum, il minimise la distance inter-cluster.

Il est possible d'indicer de manière monotone la hiérarchie sur son axe vertical (Définition 4). La construction d'une hiérarchie indicée sur un ensemble permet d'obtenir le partitionnement de la chaîne (P) en "coupant" la hiérarchie pour une valeur de l'indice. L'indice est choisi en fonction d'un critère de dissimilarité qui permet de s'assurer que le partitionnement choisi est complètement satisfaisant. C'est ce partitionnement qui constituera le vecteur de clusters ($C = \{c_1, \ldots, c_i\}$).

Définition 4 *Une hiérarchie indicée est une hiérarchie H sur un ensemble fini pour laquelle on associe une suite de nombre réel r_i. Une hiérarchie indicée est monotone si pour deux éléments h_i et h_{i+1} consécutifs dans H, avec h_i plus fin que h_{i+1}, on a $r_i \leq r_{i+1}$*

HC est une méthode classiquement utilisée en biologie, écologie. Toutefois, c'est une méthode coûteuse en temps de calcul puisque les distances entre toutes les paires d'objets possibles doivent être calculées. De plus, c'est une méthode qui ne peut pas être utilisée de manière incrémentale et qui est influencée par l'effet de "chaîne". De plus, dans l'espace euclidien, elle requiert pour être optimale de faire l'hypothèse que les données soient non-convexes et linéairement séparables ce qui n'est pas garanti avec les données physico-chimiques et biologiques.

Algorithme 1 Clustering hiérarchique ascendant

Require: S matrice de similarité de X et X ensemble de $M \times N$ objets.
Soit la table de distance $T_D = S(X)$
while T_D a plus d'une colonne **do**
Choisir les deux sous-ensembles h_i, h_j de X tel que $S(h_i, h_j)$ est le plus petit nombre réel dans T_D
Supprimer h_j de la table, remplacer h_i par $h_i \cup h_j$
Re-definir T_D soit Calculer les mesures de similarité S entre $h_i \cup h_j$ et les autres éléments de la table
end while
return les h_i sous ensemble obtenues, et le vecteur de cluster ($C = \{c_i\}$)

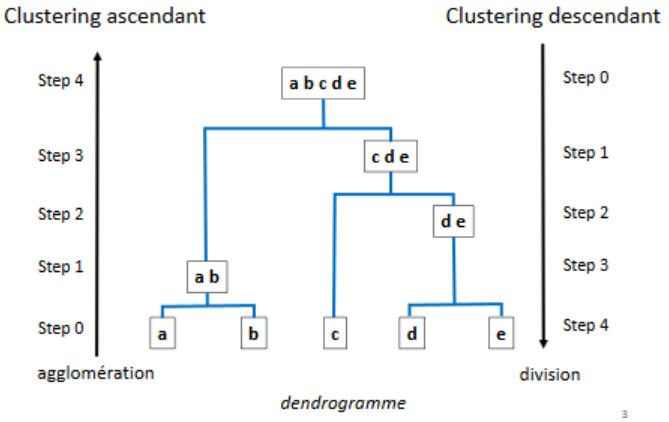


FIGURE 2.10 – Illustration du principe de segmentation par la méthode hiérarchique et du résultat obtenu appelé dendogramme.

Caractéristiques	**Valeurs**
Prise en compte du contexte	Non supervisé
Connaissances a priori	Métrique et critère d'agrégation
Méthode	basé sur la connectivité
Présentation des résultats	hiérarchique
Complexité	O($M * N^2$)
Déterminisme	oui
Incrémental	non
Hard vs soft	Hard
Tolérance au bruit	non
Tolérance l'effet de chaine	en fonction de la méthode de liens choisie
Tolérance aux clusters de tailles variées	oui
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	oui
Tolérance aux clusters convexe et non linéairement séparable	non dans l'espace euclidien

TABLEAU 2.5 – Tableau des caractéristiques : Hierarchical Clustering. En vert les caractéristiques souhaitées pour l'étude, en rouge les non souhaitées.

2.4.2.2 K-means

L'algorithme des K-moyennes ou *K-means* [HARTIGAN et WONG 1979] cherche à retrouver la partition des données en K-clusters ce qui minimise la variance au sein de chaque cluster. Il regroupe les points en fonction de leurs distances (Euclidienne) par rapport aux centres (Définition 5). L'idée principale est de construire un vecteur de prototypes, à partir de l'ensemble des observations de M variable de N points (notées $X = \{x_i\}$), tout en conservant l'information HF (Algorithme 2, Package stats : :kmeans).

Définition 5 *La méthode K-means consiste à partitionner un ensemble de données* $X = \{x_i\}$ *pour* $i \in [1, N]$ *en* K *cluster* $C = \{c_1, ..., c_K\}$ *selon un critère de minimisation des distances intra-clusters* J.

$$J(X, G) = \sum_{k=1}^{K} \sum_{i, x(i) \in c_k}^{N} \|x(i) - \mu_k\|^2 \tag{2.7}$$

Avec :

$\mu_k = \sum_{i, x(i) \in c_k} \frac{x(i)}{card(c_k)}$ le barycentre du cluster c_k ;

$X = \{x(i)\}$ l'ensemble des observations ;

K le nombre de clusters ;

$C = \{C_k\}$ l'ensemble des clusters.

Pour construire les clusters, la procédure suivante est appliquée (Figure 2.11) :

1. Initialisation des K-centres
2. Affectation de chacun des points X à son centre le plus proche (au sens de la distance euclidienne).

3. Calcul des centres de gravité des K-clusters soit ré-estimation des K-centres
4. Calcul du critère de répartition J, retour à 2) si une affectation a été modifiée. Sinon passage à l'étape 5).
5. Partitionnement fixé : retourne $C = \{c_1, ..., c_K\}$ clusters

Les étapes 3 et 4 de cette procédure peuvent être effectuées de manière incrémentale ou non. C'est-à-dire qu'il est possible de mettre à jour seulement les clusters concernés chaque fois qu'un changement d'assignation est effectué ou de les mettre à jour simultanément après toutes les assignations aux clusters. L'initialisation des centres est un point sensible de la méthode. En effet, celle-ci a tendance à converger vers des solutions optimales locales en fonction de la sélection initiale. La solution proposée est d'itérer la méthode. Ainsi, il est possible de faire varier les centres initiaux de manière aléatoire et de sélectionner un résultat optimisé moins dépendant des optimums locaux. La spécification du nombre de clusters (K) est aussi un inconvénient évident de la méthode. En effet, ce critère n'est pas toujours connu et n'est pas toujours simple à définir. Une des solutions courantes, est de réaliser plusieurs partitionnements avec des nombres de clusters différents, puis via des tests statistiques chercher un compromis entre le nombre de classes et la variance, qui permettent de sélectionner le partitionnement le plus approprié [Hamerly et Elkan 2003].

L'algorithme de quantification vectorielle *K-means* (équation 2.7) reste bien adapté aux problématiques de segmentation de données et est populaire en classification [Jain 2010].

Algorithme 2 Partitionnement K-moyennes : K-means

Require: k nombres de cluster, X
  Initialisation des K centres
  while Mouvement des centres lors de l'affectation **do**
    Affectation de chaque point X à son centre le plus proche.
    Ré-estimation des K centres en fonction des clusters (C)
  end while
  return les K centres obtenus, et le vecteur de cluster ($C = \{c_i\}$ et $c_i \in \{1, \ldots, K\}$)

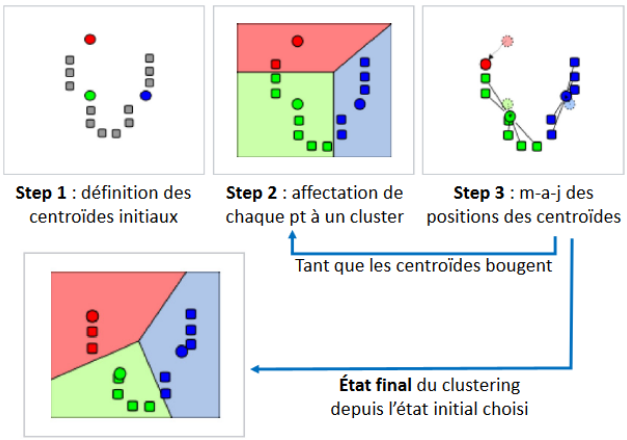


Figure 2.11 – Schéma des différentes étapes de la méthode *K-means* tiré de Reghunath 2017, chaque couleur correspondant à l'assignation à un cluster.

Caractéristiques	**Valeurs**
Prise en compte du contexte	Non supervisé
Connaissances a priori	Nombre de cluster K
Méthode	basée sur la distance (Euclidienne)
Présentation des résultats	K centroïdes
Complexité	O($M * N * K$)
Déterminisme	non
Incrémental	oui
Hard vs soft	Hard mais une version soft existe : fuzzy-kmeans (cmeans) [Dunn 1973 et Bezdek 1981]
Tolérance au bruit	non
Effet de chaine	oui
Tolérance aux clusters de tailles variées	non
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	non
Tolérance aux clusters convexe et non linéairement séparable	non

Tableau 2.6 – Tableau des caractéristiques : K-means. En vert les caractéristiques souhaitées pour l'étude, en rouge les non souhaitées.

2.4.2.3 Spectral clustering

Le principe des méthodes spectrales consiste à projeter les objets dans un nouvel espace de variance maximum (espace spectral) et d'ensuite y appliquer une méthode de partitionnement (ex : *K-means*) pour déterminer les clusters (Définition 6) [Verma et Meila 2003]. La transformation des données d'entrée dans un nouvel espace permet de réaliser les étapes de partitionnement dans un espace où les différences entre clusters sont plus importantes que dans l'espace initial. Ce principe s'appuie sur la théorie de Vapnik 2013 : dans les espaces de grandes dimensions, la probabilité d'obtenir une géométrie séparable des données augmente. Ainsi, elle facilite le partitionnement des données en les rendant plus linéairement séparables et permet de s'affranchir de la structure complexe non convexe et non-linéairement séparable des données (Figure 2.12).

Définition 6 *Soit* $\{x_i\}_{i=1,\ldots,m}$ *avec* $x_i \in in\mathbb{R}^d$ *et* S *une matrice de similarité entre toutes les paires* (x_i,x_j)*. Les méthodes spectrales de catégorisation cherchent à partitionner les points en groupe (cluster) tels que la similarité inter-groupe soit maximisée, tandis que la similarité intragroupe sera minimisée.*

Cette méthode estissue de la théorie des graphes qui s'intéresse aux rapports entre le spectre (vecteurs propres et valeurs propres) d'une matrice de relation entre des données. L'ensemble des observations peut être représenté par un graphe valué où chaque donnée définit un nœud et, chaque élément de la matrice de relation définit la valeur d'un arc inter-nœuds. Cet arc de connexion entre deux nœuds (x_i,x_j) peut être modélisé par une mesure de similarité $W = \{w_{ij} = w(x_i, y_i)\}$ qui représente la proximité entre les observations afin d'obtenir des groupes de points (Classes). Deux points sont considérés comme similaires si la valeur de la similarité $w_{ij} (ou w_{ji})$ est proche de 1.

A partir de cette matrice pondérée des arcs du graphe de données W, il est possible de définir et de calculer plusieurs caractéristiques du graphe : comme D la matrice diagonale des degrés avec $d_{ii} = \sum_j w_{ij}$ et $d_{ij,i\neq j} = 0$, puis de construire la matrice Laplacienne (L) utilisé dans le critère de

couper du graphe. De cette matrice L, le spectre du graphe est extrait : ses valeurs propres et vecteurs propres qui représenteront le nouvel espace de partitionnement.

Il existe plusieurs façons de calculer L ce qui a donné naissance à de nombreuses variantes en fonction du critère choisi. Les algorithmes de *SC* les plus courants sont les suivants [Luxburg 2007] :

La classification spectrale non-normalisée :

Le *SC* non-normalisé (Algorithme 3,Package anocvi : :spectralClustering) est basé sur le calcul d'un Laplacien non-normalisé (L) défini comme : $L = D - W$ avec D matrice diagonale et W matrice de poids.

Algorithme 3 Clustering spectral non-normalisé

Require: Matrice de similarité $S \in \mathbb{R}^{N \times N}$, K le nombre de classes
Construire le graphe de similitude avec W=f(S) sa matrice d'adjacence pondérée.
Calculer le Laplacien non normalisé L
Extraire les K premiers vecteurs propres $\{v_1, \ldots, v_K\}$ de L
Former $V \in \mathbb{R}^{N \times K}$ la matrice des vecteurs propres $\{v_1, \ldots, v_k\}$
Pour $i = 1, \ldots, n$, avec $y_i \in \mathbb{R}^K$ le vecteur correspondant à la i-ème ligne de V
Partitionner les points $(y_i)_{i=1,\ldots,N} \in \mathbb{R}^k$ avec l'algorithme $K - means$
return le vecteur de cluster $C = \{c_1, \ldots, c_N\}$

La matrice W peut être directement la matrice de similarité d'entrée ou une matrice creusée en annulant les arcs trop éloignés. Dans chaque algorithme, les étapes différentes sont surlignées pour les mettre en évidence au lecteur.

La classification spectral normalisé :

Il existe deux versions différentes de classification spectrale normalisée qui dépendent du Laplacien normalisé utilisé ; les différences sont grisées dans chaque algorithme :

— Shi et Malik.,2000 : *Bi-parted Spectral Clustering (Bi-SC)* (Algorithme 4,Package kknn : :specClust) [Shi et Malik 2000] utilisent les vecteurs propres généralisés de L (non-normalisé) ce qui revient à utiliser une matrice normalisée L_{rw} (matrice liée à la marche aléatoire "*randum walk*") telle que $L_{rw} = D^{-1}L$. C'est pourquoi elle est qualifiée de clustering spectral normalisé [Luxburg 2007]. Le bi-partitionnement est défini en fonction des signes du second vecteur propre dominant (de la matrice U) contenus dans la matrice des vecteurs propres U par ordre croissant de ses valeurs propores.

Algorithme 4 Clustering spectral Shi et Malik.,2000

Require: Matrice de similarité $S \in \mathbb{R}^{N \times N}$, K le nombre de classes
Construire le graphe de similitude avec W sa matrice d'adjacence pondérée.
Calculer le Laplacien non normalisé L
Calculer les K premiers vecteurs propres généralisés $\{v_1, \ldots, v_K\}$ avec $Lv = \lambda Dv$
Soit $V \in \mathbb{R}^{N*K}$ la matrice des vecteurs propres $\{v_1, \ldots, v_K\}$
Pour $i = 1, \ldots, N$, avec $y_i \in \mathbb{R}^K$ le vecteur correspondant à la i ligne de V
Partitionner les points $(y_i)_{i=1,\ldots,N} \in \mathbb{R}^k$ avec l'algorithme $K - means$
return le vecteur de cluster $C = \{c_1, \ldots, c_N\}$

— Ng et al., 2001 : *NJW-Spectral Clustering (NJW-SC)* (Algorithme 5, Package uHMM : :KpartitionNJW) [Ng et al. 2001], utilise également un Laplacien normalisé, mais cette fois la matrice L_{sym} (matrice de symétrique) avec $L_{sym} = D^{\frac{1}{2}} L D^{\frac{1}{2}}$ qui est utilisée au lieu de L_{rw}.

Algorithme 5 Clustering spectral Ng, Jordan and Weiss.,2001

Require: Matrice de similarité $S \in \mathbb{R}^{N \times N}$, K nombre de classes
Construire le graphe de similitude avec W sa matrice d'adjacence pondérée.
Calculer le Laplacien normalisé L_{sym}
Calculer les K premier vecteurs propres $\{v_1, \ldots, v_k\}$ de L_{sym}
A partir de la matrice $T \in \mathbb{R}^{N*K}$ de V en normalisant les lignes à la norme 1 avec $t_{ij} = \frac{v_{ij}}{(\sum_k v_{ij}^2)^{\frac{1}{2}}}$
Soit $V \in \mathbb{R}^{N*K}$ la matrice des vecteurs propres $\{v_1, \ldots, v_k\}$
Pour $i = 1, \ldots, N$, avec $y_i \in \mathbb{R}^K$ le vecteur correspondant à la i-ème ligne de T
Partitionner les points $(y_i)_{i=1,\ldots,N} \in \mathbb{R}^K$ avec l'algorithme $K - means$
return le vecteur de cluster $C = \{c_1, \ldots, c_N\}$

De plus, une variante de la méthode de Ng et al. 2001 a été proposée par Sanchez-Garcia, Fennelly et al. 2014 : Le *Hierarchical Spectral Clustering (H-SC)*. La particularité de cette méthode est qu'elle utilise une méthode de partitionnement hiérarchique plutôt que la méthode *K-means* (Algorithme 6). Cette méthode combine donc le principe de hiérarchie et de la théorie des graphes. Le passage dans l'espace spectral permettant d'utiliser les propriétés de la méthode de *clustering* hiérarchique de manière optimale.

Algorithme 6 Clustering spectral hiérarchique Garcia et al, 2014

Require: Matrice de similarité $S \in \mathbb{R}^{N \times N}$, K nombre de classes
Construire le graphe de similitude avec W sa matrice d'adjacence pondérée.
Calculer le Laplacien **normalisé** L_{sym}
Calculer les K premier vecteurs propres $\{v_1, \ldots, v_k\}$ de L_{sym}
A partir de la matrice $T \in \mathbb{R}^{N \times K}$ de V en normalisant les lignes à la norme 1 avec $t_{ij} = \frac{v_{ij}}{(\sum_k v_{ij}^2)^{\frac{1}{2}}}$
Soit $V \in \mathbb{R}^{N \times K}$ la matrice des vecteurs propres $\{v_1, \ldots, v_K\}$
Pour $i = 1, \ldots, N$, avec $y_i \in \mathbb{R}^K$ le vecteur correspondant à la i-ème ligne de T
Partitionner les points $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^K$ avec l'algorithme de Classification Hiérarchique
return le vecteur de cluster $C = \{c_1, \ldots, c_N\}$

Ainsi la classification spectrale peut être envisagé comme un problème de coupe de graphe en 4 grandes étapes :

1. Construction du graphe de données
2. Construction de la matrice de similarités
3. Construction de la matrice Laplacienne et extraction des valeurs et vecteurs propres associés
4. Recherche de groupes dans l'espace spectral

La méthode cherche à partitionner les données en K classes selon un critère de coupe (N_{Cut}) définie en fonction de la mesures de similarité W du graphe de données. N_{Cut} est choisi en

fonction de l'application : coupe intra-cluster, coupe inter-cluster, coupe normalisée, selon le cardinal ou le volume des clusters obtenus [Shi et Malik 2000]. Ce critère de coupe est basé sur la minimisation des variances au sein des clusters et la maximisation des variances entre les clusters. Il utilise les distances entre points dans l'espace spectral obtenu à partir du critère de coupe. L'idée est d'employer les premiers vecteurs propres d'une matrice dérivée des distances entre les points pour représenter les données.

N_{Cut} peut être réécrit de manière matricielle sous la forme d'un quotient de Rayleigh [Shi et Malik 2000](équation 2.8) faisant intervenir :

$W = \{w_{ij} = w(x_i, x_j)\}$ matrice de similarité qui doit être symétrique semi-définie positive (matrice de Gram).

D la matrice diagonale des degrés avec $d_{ii} = \sum_j w_{ij}$ et $d_{ij,i\neq j} = 0$. d_{ii} correspond à la somme en ligne de W, soit le volume de chaque nœud (dit aussi degré) ;

Z la matrice des vecteurs propres ;

F la matrice indicatrice.

$$NCut(X,G) = \frac{F^T(D-W)F}{F^TDF} = \frac{D^{-\frac{1}{2}}Z^T(D-W)ZD^{-\frac{1}{2}}}{Z^TZ} \tag{2.8}$$

Les v vecteurs propres sont donc extraits de la matrice Laplacienne L_{N_g} de nos données ($L_{N_g} = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ ordonnés selon leurs valeurs propres croissantes).

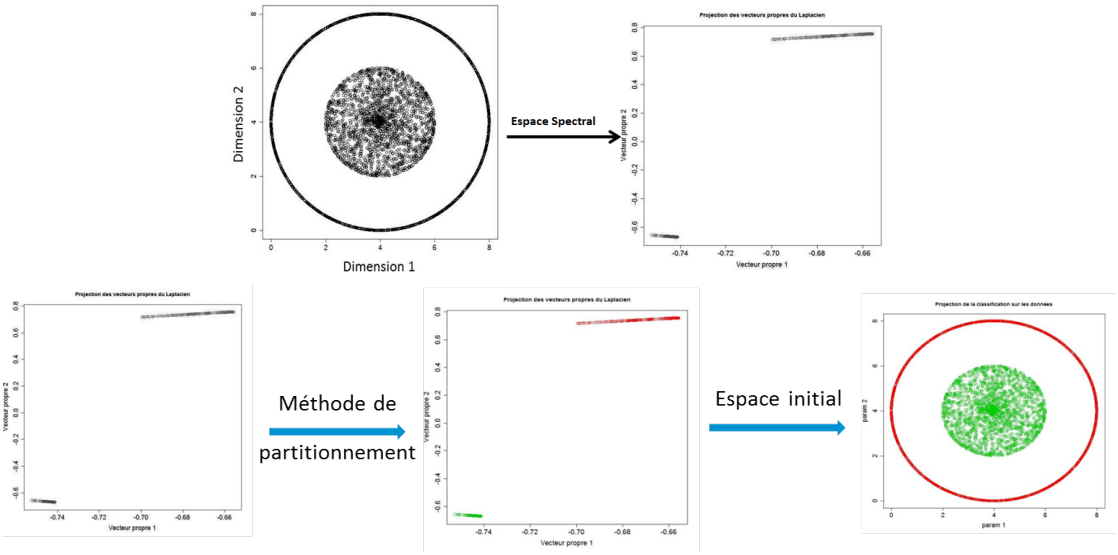


Figure 2.12 – Schéma des différentes étapes de la méthode *SC*. Les couleurs correspondant à la projection des clusters (l_i) sur les points x_i dans l'espace des vecteurs propres puis dans l'espace initial.

Caractéristiques	Valeur
Prise en compte du contexte	Non supervisé
Connaissances a priori	Matrice de similarité
Méthode	basée des matrices de similarité
Présentation des résultats	K cluster
Complexité	O(M*N*K)
Déterminisme	non
Incrémental	oui
Hard vs soft	Hard (adaptable Spectral c-means)
Tolérance au bruit	non
Effet de chaine	oui
Tolérance aux clusters de tailles variées	oui
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	oui
Tolérance aux clusters convexe et non linéairement séparable	oui

TABLEAU 2.7 – Tableau des caractéristiques : Spectral clustering. En vert les caractéristiques souhaitées pour l'étude, en rouge les non souhaitées.

Choix pour l'étude :

Dans notre cas, nous calculons la matrice de similarité de ZELNIK-MANOR et PERONA 2004, qui est définie à partir d'un noyau gaussien normalisé par des dispersions locales. Celle-ci permet de rendre compte des distances entre les observations mais aussi des relations de voisinage. Ainsi, cette fonction diminue de manière exponentielle lorsque les distances sont importantes ce qui a pour conséquence de creuser la matrice de similarité et d'accentuer les différences entre les observations et ce qui offre l'avantage en terme de segmentation.

De plus nous utilisons des graphes entièrement connectés et nous générons donc un espace des vecteurs propres (ou espace spectral) de la taille de la matrice de données et non de dimension des composantes principales, comme par exemple dans le cas des analyse en composante principale (ACP). Le fait d'utiliser un espace de grande dimension (pour rappel déjà creusé par le calcul de la matrice de similarité par la méthode des noyau gaussien) augmente la probabilité que les données soient davantage linéairement séparables.

Fast Spectral clustering :

De plus, la grande taille de ces bases de données HF ralentit fortement le traitement par classification spectrale. En effet, lorsque le volume de données est grand (supérieur à 20 000 points - *i.e.* pour une machine de calcul standard actuel en 2018 : x86 64 bits avec 8Go RAM), comme dans notre cas, les calculs sont limités [SHINDLER et al. 2011]. C'est pourquoi dans des travaux précédents ; [ROUSSEEUW et al. 2015b], une variante de la méthode de NG et al. 2001 a été développée appelée *Fast NJW-Spectral Clustering (Fast NJW-SC)* (Package R uHMM :vFastSpectralNJW).

Dans cette variante, une étape supplémentaire a été ajoutée : le sous-échantillonnage des données. Cette étape est appliquée avant la phase de classification. Lors de cette phase, un sous-échantillonnage optimal de la base est opéré par le biais d'une quantification vectorielle réalisée par la fonction *K-means*. Les centres finaux obtenus seront utilisés pour constituer un nouvel ensemble réduit de données. Ainsi, le nombre optimal de clusters K est incrémenté jusqu'à

ce que le pourcentage de variance expliquée ou le nombre K_{max} de prototypes retenus soit atteint (par défaut la variance expliquée =97% et K_{max}=2000).

Ainsi, le processus itératif suivant est appliqué :

1. Initialisation de $K = 2$ centres ;
2. Affectation de chacun des points X à son centre le plus proche (Distance Euclidienne) ;
3. Ré-estimation des centres ;
4. Calcul du critère de répartition J et de la variance expliquée, retour à 2) si l'un des critères n'est pas respecté avec $K = K + 1$.
5. Construction de l'ensemble réduit à partir des K centres

De cette version initiale, a été proposé de remplacer les K centres par une sélection aléatoire de N observations réelles appartenant à chaque centre de manière à représenter plus fidèlement la densité liée à chaque cluster [Poisson-Caillault 2020]. La propagation des numéros de cluster à l'ensemble des données X est réalisée par un algorithme des k-plus proches voisins à ces représentants issus de cette quantification vectorielle adaptée. C'est cette version optimisée qui est utilisée dans cette thèse.

2.4.3 Méthodes usuelles de classification supervisée : prédiction des classes par apprentissage

Dans le but de prédire les classes (cl) définies à partir des méthodes de classification non supervisée - *clustering*, plusieurs méthodes d'apprentissage supervisé sont testées.

Les méthodes de classification supervisée se déclinent en deux grandes familles, celles dites approches discriminantes et génératives. Elles diffèrent selon la manière d'estimer le label l en fonction des entrées x. Un modèle discriminant apprend la distribution de probabilité conditionnelle $p(l|x)$, il construit alors un ensemble de frontières qui permet de séparer les classes. Un modèle génératif apprend la distribution de probabilité conjointe $p(x,l)$, il nécessite selon les règles bayésiennes de connaître la distribution d'une classe $p(l)$. Dans nos applications, la première solution est donc privilégiée n'ayant aucune information suffisante sur la distribution réelle de chaque classe.

Plusieurs classifieurs discriminants classiques seront employés : des forêt aléatoires (*Random Forest (RF)*), des séparateur à vaste marge (*Support vector machine (SVM)*), dans leur version primaire ou des versions profondes pour assurer la séparabilité des classes [Vapnik 1995].

Un dernier algorithme n'appartenant pas aux deux familles précédentes est utilisé celui des *k plus proches voisins (k-ppv)*. En effet, celui-ci affecte le label à une observation en fonction des labels des k plus courtes distances aux données de la base dite d'apprentissage c'est-à-dire aux couples (x,l) connus. Il est largement utilisé pour sa facilité de mise en oeuvre, son interprétation aisée et ses performances non éloignées des techniques de Machine Learning les plus poussées [Chatzigeorgakidis et al. 2018].

Dans cette section, il est décrit le principe général de chaque méthode. Ensuite, les atouts et les limites sont exposés. Enfin, les principales caractéristiques sont résumées (Tableau 2.4). Les concepts sont nombreux, c'est pourquoi un rappel sur le principes de chaque méthode est fait dans les parties 4.2.2 et 3.3.4.

2.4.3.1 k-plus proche voisin

La méthode k-plus proche voisin (k-ppv) [Cover et Hart 1967], ou *k-Nearest Neighbors (KNN)* en anglais, est un algorithme de classification supervisée basé sur une instance. En effet,

il ne comprend pas de phase d'apprentissage en tant que telle, car il construit des hypothèses directement à partir des instances du jeux de données labellisée composé de n couples d'observations (x_i, l_i). C'est pourquoi il est qualifiée de « paresseux » (*lazy learning*).Le principe de l'algorithme (K-ppv) est de calculer la distance entre un point non labellisé (observation) et le jeu de données labellisées pour attribuer à l'observation inconnue la classe majoritaire de ses k-plus proches voisins (Définition 7, algorithme 7).

Définition 7 *Soit* $X = \{(x_i, l_i) | i \in [1..N],\ x_i \in \mathbb{R}^M,\ l_i \in \mathbb{N}\}$, *un ensemble d'apprentissage de* N *points de dimension* M *où* l_i *est la classe de labellisation de* x_i. *Et* k *le nombre de voisins à considérer.* $(x_1, l_1), (x_N, l_N)$ *est un ré-arrangement des données d'apprentissage tel que les* k *premiers couples soient les* k *plus proches voisins de* x *respectant la métrique c'est-à-dire* $\|x_1 - x\| \leq \cdots \leq \|x_k - x\| \cdots \leq \|x_N - x\|$. *Le* k*-voisinage de* x *noté* $V_k(x)$ *est l'ensemble des* k *premiers points de cette arrangement.*

Ainsi, la classification d'un point p est déterminée à l'aide d'une mesure de distance, d'un nombre k fixé et d'une règle de décision basée sur un vote majoritaire parmi les k labels du voisinage $V_k(p)$ c'est-à-dire l'ensemble des points situés à une distance d_{max} de p. Pour la distance, plusieurs choix sont possible parmi lesquels on retrouve la distance Euclidienne, la distance de Manhattan, la distance de Minlowski de paramètre q, la distance maximale, la distance de Camberra. En fonction du nombre de points k choisi pour définir le voisinage, on choisira d_{max} comme la distance entre p et le k-ème point le plus proche (Figure 2.13).

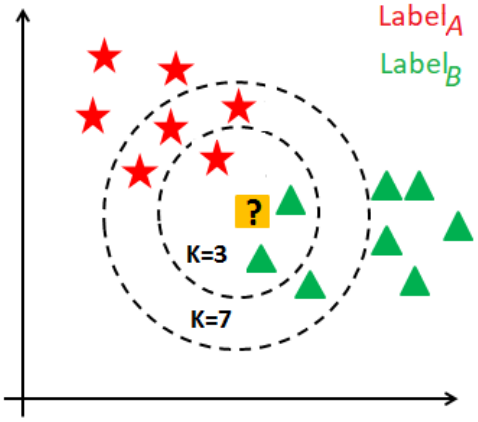


FIGURE 2.13 – Illustration du principe de classification par la méthode k-plus proches voisin

Cette méthode simple est facile a mettre en œuvre. Elle ne nécessite pas la construction de modèle ou de faire des hypothèses supplémentaires, ni l'ajustement d'un grand nombre de paramètres. Toutefois, il est évident que cette méthode est dépendante du nombre de K choisi. Si l'on choisit un nombre de k voisin faible, un modèle local sera construit en revanche choisir un k élevé permet de réduire le bruit en lissant les résultats, et de diminuer le risque de sur-apprentissage mais la frontière entre les classes risque d'être moins nette. L'estimation du modèle peux devenir mauvaise si le ratio entre le nombre de voisin et la densité des points du jeu de données n'est pas équilibrée. Il est important de trouver un compromis entre le nombre de points et le nombre de voisins choisi. De plus, l'algorithme ralentit considérablement à mesure que le nombre d'observations et/ou de variables dépendantes/indépendantes augmente. En effet, elle recalcule toutes les distances pour chaque point inconnu. Une des solutions proposées pour

optimiser ce temps est un découpage du plan en cellules (couramment nommée zone de Voronoi). Le calcul du diagramme de Voronoi permet le découpage de l'espace en cellules, chacune associée à un point. Chaque cellule est créée à partir d'un exemple d'apprentissage et forme l'ensemble des points les plus proches de l'exemple correspondant (Figure 2.14).

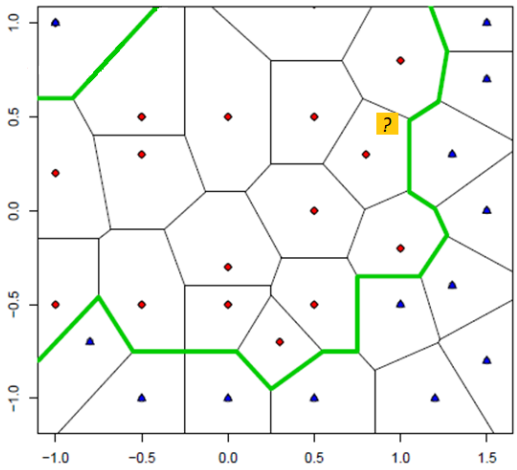


FIGURE 2.14 – Illustration du principe de k-plus proches voisin avec l'application des cellules de Voronoi sur une jeu de donnée a deux dimension. En noir, la limite de la cellule pour chaque point et en vert, la limite définie par les classes rouge et bleue.

Algorithme 7 K plus proche voisin

Require: p un nouveau point, k nombre de voisins, X la base labellisée
Pour tout couple $\{x_i, l_i\}$ de l'ensemble des points d'apprentissage X
Calculer les distances $d(x_i, p)$
Compter le nombre d'occurrences de chaque label $l_{i,\ldots,k}$ pour les k plus proches voisins de p
Attribuer à x la classe majoritaire l
return l le label de p

Caractéristiques	**Valeur**
Prise en compte du contexte	supervise
Connaissances a priori	base labéllisée
Méthode	basée sur la distance
Présentation des résultats	définis la classe (cl) du point x ; pas de modèle
Complexité	$N \times \log(N)$
Déterminisme	non
Incrémental	oui
Hard vs soft	Hard
Tolérance au bruit	non
Effet de chaine	oui
Tolérance aux clusters de tailles variées	oui
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	oui
Tolérance aux clusters convexe et non linéairement séparable	oui

TABLEAU 2.8 – Tableau des caractéristiques : K plus proches voisin (K-ppv).

2.4.3.2 Random Forest

La méthode Random Forest [BREIMAN 2001] est fondée sur les arbres de décision qui réalisent la classification d'un objet par une suite de tests en fonction des attributs qui le décrivent. C'est un schéma/graphe représentant les résultats possibles d'une série de choix interconnectés. Il emploie pour cela une représentation hiérarchique descendantes sous forme de séquences de décisions, chaque test représente un nœud.

Il existe 3 types de nœuds :

- nœud racine : l'accès à l'arbre se fait par ce nœud (symbole carré Figure 2.15)
- nœuds internes : les nœuds qui ont des descendants (ou enfants), qui sont à leur tour des nœuds (symbole rond Figure 2.15)
- nœuds terminaux (ou feuilles) : nœuds qui n'ont pas de descendant. (symbole triangle Figure 2.15)

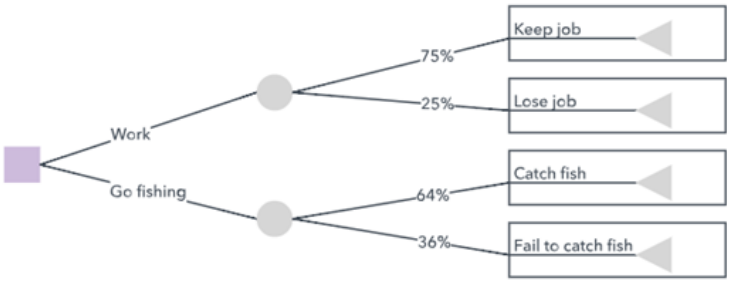


FIGURE 2.15 – Illustration de la construction d'un arbre de décision

La construction d'un arbre de décision (unique) est fortement basée sur l'expertise qui lie l'objet que l'on veut classer à ces attributs. La construction d'une «forêt d'arbres» (donc,

plusieurs arbres de décision), de manière aléatoire (le *Random Forest*) va permettre de sélectionner automatiquement un arbre de décision (performant et généraliste) qui s'appuie seulement sur l'ensemble d'apprentissage. En effet, cette méthode, qualifiée d'« *ensemble method* » combine une multitude de modèles «faibles» pour créer un modèle robuste.

Le nombre d'arbres possible croît exponentiellement avec le nombre d'attributs, il n'est pas possible de tester exhaustivement tous les arbres. Il faut donc définir le nombre d'arbres et le nombre de couches de ces arbres (Définition 8 et 9)

Définition 8 *Généralement le nombre moyen d'arbres C_n à $n > 0$ noeuds est défini par le n-ième nombre de Catalan (les nombres de Catalan forment une suite d'entiers naturels)*

$$C_n = \frac{1}{1+n}\binom{2n}{n}$$

Définition 9 *Le nombre de nœud détermine le partitionnement de l'espace des exemples. Plus il est important, plus la classification sera développer. A partir de 10 niveaux la procédure est considérée comme de l'apprentissage profond.*

La méthode Random Forest utilise donc un grand nombre d'arbres de décision construits chacun avec un sous-échantillon différent de l'ensemble d'apprentissage, et pour chaque construction d'arbre. La décision à un nœud est fait en fonction d'un sous-ensemble de variables $(f_1, \ldots, f_n)$ tirées au hasard. Ensuite, l'ensemble des arbres de décision produits sont utilisés pour faire la prédiction, avec un vote à la majorité (pour de la classification, variable prédite de type facteur), ou une moyenne (pour de la régression, variable prédite de type numérique) (Figure 2.16).

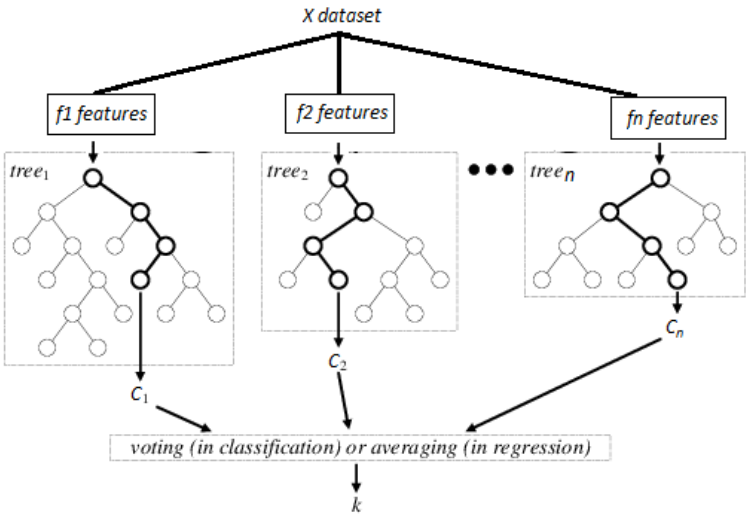


FIGURE 2.16 – Illustration de la construction d'un arbre de décision

La méthode de segmentation, des arbres et des forêts d'arbres, par succession de choix facilite l'interopérabilité de l'arbre et la manipulation de données « symboliques ». De plus, c'est une classification très efficace en particulier sur des entrées de grande dimension, peu sensible à des étendues des variables différentes.

Toutefois, pour N observations de M dimensions, le temps d'apprentissage peut être long avec une représentation complète, la complexité est alors égale à O($M \times N \times \log(N)$). Il est de même pour une architecture réduite à nA arbres de profondeurs d et une sélection de m<M variables, la complexité étant de $0(nA \times m \times d \times N)$. De plus, Les valeurs extrêmes sont souvent mal estimées dans le cas de régression et cette méthode est sensible aux bruits et aux points aberrants. De plus la stratégie d'élagage est souvent délicate. En effet, un arbre de décision sera efficace, s'il respecte le principe de « simplicité » de l'arbre et de proximité du risque empirique et du risque réel.

Algorithme 8 Construction récursive d'un arbre de décision : construction du nœud X

Require: le couple $X = \{(x_i, l_i)\}$ de l'ensemble des points d'apprentissage
if Pour tout x ∈ à une même classe
Créer une feuille portant le nom de cette classe
else
Choisir le meilleur attribut pour créer un nœud
le test associée à ce nœud sépare X en deux partie noté x_g, x_d
return l label de x

Caractéristiques	**Valeur**
Prise en compte du contexte	supervisé
Connaissances a priori	Base labellisée (entrainement et test), nombre nœud/arbres
Méthode	Méthode itératif de partitionnement récursif basé sur une succession de questions
Présentation des résultats	modèle de classification et label de x
Complexité	max : $(M \times N \times \log(N))$
Déterminisme	oui
Incrémental	oui
Hard vs soft	Soft
Tolérance au bruit	oui
Effet de chaine	non
Tolérance aux clusters de tailles variées	oui
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	oui
Tolérance aux clusters convexe et non linéairement séparable	oui

Tableau 2.9 – Tableau des caractéristiques : Random Forest (RF).

2.4.3.3 Séparateur à vaste marge

La méthode *SVM* Ben-Hur et Weston 2010 est un classifieur discriminant basé sur la notion de frontières entre deux classes. L'apprentissage d'un SVM linéaire revient à déterminer l'équation de cette frontière modélisée par un hyperplan optimal (équation 10). L'hyperplan optimal en rouge sur la Figure 2.17 est obtenu à partir de points supports (observations labellisées entourées) et un critère de minimisation du risque empirique défini par l'équation 11 caractérisant ainsi la marge maximale acceptée. Tout point dans la zone des deux hyperplans bleus a un risque d'être mal classé.

Définition 10 *L'équation de l'hyperplan s'exprime sous la forme :*

$$h(x) = \langle w, x \rangle + w_0 = \sum_{i=0}^{d} w_i x_i = 0$$

Avec w le vecteur directeur (orthogonal à l'hyperplan)

Définition 11 *L'hyperplan optimal est définie par :*

$$\arg\max_{w,w_0}(\min\{\|x - x_i\| : x \in \mathbb{R}^d, (w^T x + w_0) = 0, i = 1, \ldots, m\})$$

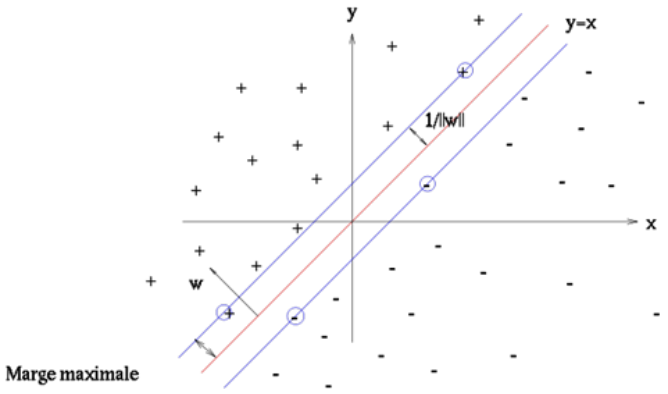


FIGURE 2.17 – Illustration de la définition de hyperplan optimal séparant les points de deux classes appliqué par la méthode de séparateur à vaste marge (SVM)

Les définitions précédentes supposent que les exemples d'apprentissages sont linéairement séparables ce qui est rarement le cas dans nos jeux de données. Une des méthodes proposées est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension, dans lequel il est probable qu'il existe une séparation linéaire (l'espace de re-description) : c'est la technique de *kernel trick* (ou astuce du noyau) . En effet, l'utilisation des fonctions noyaux permet de traiter des problèmes de discrimination non linéaire, et de reformuler le problème de classement comme un problème d'optimisation quadratique. Ainsi, selon la métrique choisie (fonction noyau), l'hyperplan peut devenir non linéaire. De plus, la fonction noyau a l'avantage de ne pas nécessiter de connaissance explicite de la transformation à appliquer.

Caractéristiques	**Valeur**
Prise en compte du contexte	supervise
Connaissances a priori	Base labellisée (entrainement et test), information relative à la fonction noyau choisie, nombre de classes
Méthode	Classification binaire basée sur la distance à un hyperplan
Présentation des résultats	modèle de classification et label de x
Complexité	entre O(N^2) et O(N^3)
Déterminisme	oui
Incrémental	oui
Hard vs soft	Soft
Tolérance au bruit	Initialement non mais oui avec l'utilisation d'une fonction de coût de substitution
Effet de chaine	non
Tolérance aux clusters de tailles variées	oui
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	non
Tolérance aux clusters convexes et non linéairement séparables	Initialement non mais oui avec la technique de Kernel trick (astuce des noyaux)

TABLEAU 2.10 – Tableau des caractéristiques : Séparateur à Vaste Marge (SVM).

2.4.4 Validation

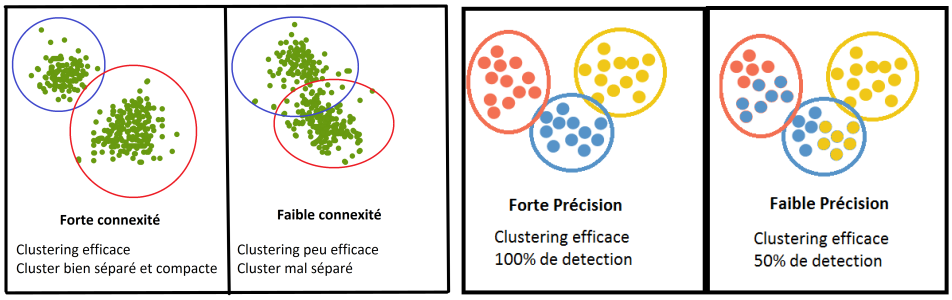
2.4.4.1 Classification

Afin d'évaluer les performances de classification, cinq indicateurs [ZHAO 2012] sont calculés : l'indice de Rand ajusté (ARI), l'indice de Dunn, le score Silhouette et la précision de classification (issus du tableau de confusion).

- L'indice de Rand ajusté (ARI) est la version corrigée du hasard de l'indice de rand (RI) disponible sur le package fossile : modifiedRandIndex [VAVREK 2011]. RI mesure le nombre d'accords entre deux partitions (mêmes éléments de classification et mêmes paires de séparation) par rapport au nombre total de paires. Son principe est de mesurer la consistance (le taux d'accord) entre deux partitions. L'indice Rand ne sera jamais réellement nul. ARI peut générer des valeurs négatives si l'indice est inférieur à l'indice attendu.

- L'indice Dunn (Package clValid : :dunn [BROCK et al. 2008]) est le rapport de la plus petite distance entre les observations ne faisant pas partie du même groupe et de la plus grande distance intra-groupe. L'indice Dunn a une valeur comprise entre zéro et l'infini et doit être maximisé. Il reflète la séparabilité des données et des groupes. Un indice de Dunn proche de zéro indique des données très liées, tandis qu'un score élevé indique des données facilement séparables.

- La Silhouette (Package cluster : :silhouette [MAECHLER et al. 2018]) est la mesure de la similitude d'un objet avec son propre cluster (cohésion) par rapport à d'autres clusters (séparation). Une grande silhouette (presque 1) est très bien séparée, tandis qu'une petite silhouette (environ 0) signifie que les observations sont connectées entre deux groupes.



(a) Illustration de la notion de connexité

(b) Illustration de la notion précision

FIGURE 2.18 – Illustration de la notion de connexité (a) et de la précision (b). Les colonnes de gauche représentant un indice fort et les colonnes de droite un indice faible. Les cercles sont les clusters "hypotétiques", c'est-à-dire, crées lors du partitionnement. Pour le graphique de a) les couleurs des points représentent les clusters "vrais", c'est-à-dire définis par exemple par un expert.

- Le taux de précision (Package MLmetric : :Accuracy) est le pourcentage de classes bien reconnues selon une vérité terrain donnée (labels). C'est le rapport entre l'observation correctement classée et l'ensemble des observations. Les tables de confusion entre les partitions et sa labellisation et un calcul d'assignation par des règles de vote à la majorité sont utilisés pour indiquer si nos événements attendus sont détectés. Le principe du vote à la majorité suivant est appliqué pour affecter chaque classe à un label.

Ainsi, les indexes RI, ARI, Silhouette, dont les principales caractéristiques sont résumées dans le tableau 2.11, permettent d'avoir une information sur la connexité des groupes (Figure 2.18a). La précision et les tableaux de confusion donnent une indication sur le taux de détections de chaque événement (Figure 2.18b).

2.4.4.2 Apprentissage - Prédiction

Pour évaluer l'efficacité des modèles de prédiction, 4 indices de performance sont calculés. Le calcul de ces indices est basé sur les valeurs issues du tableau de confusion.

Les tableaux de confusion sont souvent utilisés pour décrire la performance d'un modèle de classification supervisée et d'apprentissage. Ce tableau confronte les résultats issus de l'apprentissage et les résultats "vrai" soit les résultats attendus. De ce tableau peuvent être extraites 4 valeurs (Figure 2.19) :

- Vrais positifs (VP) sont les valeurs positives correctement prédites, ce qui signifie que la valeur de la classe vraie et la classe prédite sont identiques et définis les valeurs correctement déclarées comme appartenant à la bonne classe. Par exemple sur la figure 2.19 : la classe vraie est oui et que la valeur de la classe prédite est également oui.

- Vrais négatifs (VN) sont les valeurs négatives correctement prédites, ce qui signifie que la valeur de la classe vraie et de la classe prédite sont identiques, mais cette fois ce sont les valeurs correctement déclarées comme n'appartenant pas à la classe. Par exemple sur la figure 2.19 : la classe vraie est non et que la valeur de la classe prédite est également non.

- Faux positifs (FP) sont les valeurs positives mal prédites, ce qui signifie que la classe vraie et en contradiction avec la classe prédite. Dans le cas positif, c'est la classe prédite qui est

TABLEAU 2.11 – Tableau des caractéristiques des indices de comparaisons.

Index	Definiton	Valeurs	Package
Rand index RI	mesure de similarité entre deux partitions d'un ensemble	0 : mauvais regroupement 1 :Bon regroupement	fossile : :RandIndex
Ajusted Rand index ARI	version corrigée du hasard de l'indice Rand	-inf : mauvais regroupement 1 : Bon regroupement	fossile : :modifiedRandIndex
Dunn	rapport entre la plus petite distance entre les observations ne se trouvant pas dans le même groupe et la plus grande distance intra-groupe.	0 : mauvais regroupement +inf : quand maximisé très bien regroupé	clValid : :dunn
Silhouette	mesure de la similitude d'un objet avec son propre cluster (cohésion) par rapport aux autres clusters (séparation)	-1 : mauvais regroupement 0 : l'observation se situe entre deux groupes 1 : Bon regroupement	cluster : :silhouette
Précision	pourcentage de classe bien reconnues	0 : 0 % de reconnaissance 1 : 100 % de reconnaissance	MLmetric : :Accuracy

déclarée comme appartenant à une classe alors qu'elle ne l'était pas. Par exemple sur la figure 2.19 : la classe vraie est non et que la valeur de la classe prédite est également oui.

- Faux négatifs (FN) sont les valeurs négatives mal prédites, ce qui signifie que la classe vraie et en contradiction avec la classe prédite. Mais dans le cas négatif, c'est la classe prédite qui est déclarée comme n'appartenant pas à une classe alors qu'elle l'était. Par exemple sur la figure 2.19 : la classe vraie est oui et que la valeur de la classe prédite est également non.

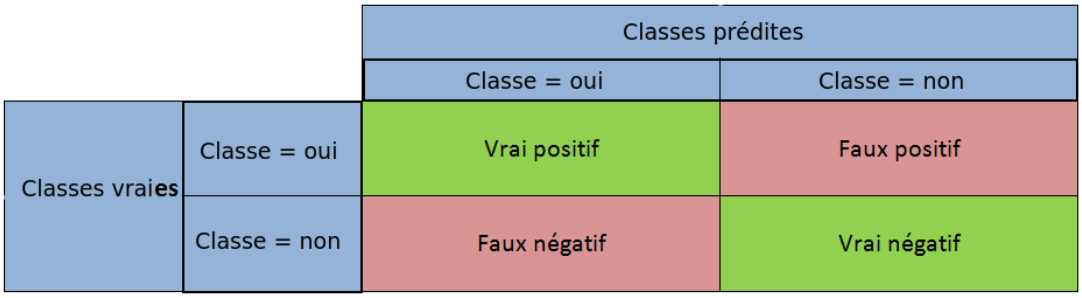


FIGURE 2.19 – Représentation des différentes informations fournies par un tableau de confusion.

Ensuite, nous calculons :

- *Accuracy* (précision) est le rapport entre l'observation correctement prédite et l'ensemble des observations. Plus la précision est proche de 1 plus notre modèle est précis. Mais cela ne signifie pas que c'est le meilleur. Pour être vraiment juste, il faut avoir des ensembles de données symétriques où les valeurs des faux positifs et des faux négatifs sont presque identiques. D'autres indices sont nécessaires pour compléter.
  $Précision = \frac{TP+TN}{TP+FP+FN+TN}$

- *Precision* (précision) est le rapport entre les observations positives correctement prédites et le total des observations positives prédites. Un résultat proche de 1 est un bon résultat. $Précision = \frac{VP/}{VP+FP}$
- *Recall* (Rappel ou sensibilité) est le rapport entre les observations positives correctement prédites et toutes les observations de la classe réelle. Il répond à la question combien de classes ont bien été prédites. Un *Recall* supérieure à 0,5 est considéré comme un bon score. $Rappel = \frac{VP}{VP+FN}$
- *F1 Score* est la moyenne pondérée de la précision et du rappel. Par conséquent, ce score tient compte à la fois des faux positifs et des faux négatifs. Le F1 est généralement plus utile que la précision, surtout pour une distribution de classe inégale. $ScoreF1 = \frac{2*(Rappel*Précision)}{(Rappel+Précision)}$

Chapitre 3

Développement d'une méthode de *clustering* innovante pour la détection d'évènements dans les séries temporelles

3.1 Introduction

Pour répondre aux objectifs de cette thèse formulés dans l'introduction générale (section 3), nous cherchons un moyen d'établir une typologie de la dynamique phytoplanctonique à partir de signaux multivariées HF et sans connaissance *a priori* sur les données. Définir cette typologie nécessite d'isoler dans les observations une suite d'états environnementaux distincts. Cependant l'identification des états environnementaux dans ces systèmes dynamiques non linéaires est une tâche difficile. En effet, ces états sont très variables en termes de distributions, de formes, de durées, de périodes d'apparitions. De plus, les bases de données écologiques sont souvent dépourvues de méta-données sur les proliférations algales ou les évènements caractéristiques et elles sont très rarement labellisées. Ainsi, des outils numériques optimisés capables de segmenter les données en différentes classes sans aucune information *a priori* sont nécessaires. La caractérisation, la labellisation des bases de données et la mise à disposition des méta-données sont essentielles pour développer notre état de compréhension sur le fonctionnement des écosystèmes. La mise à disposition de bases labellisées pourrait permettre la comparaison avec d'autres écosystèmes et donc d'étendre les recherches.

Afin de répondre au mieux à ces contraintes, nous avons opté pour une approche semi-supervisée couplant une phase de classification non supervisée, une labellisation par expertise et une phase d'apprentissage supervisée via des méthodes de *Machine learning* (Introduction générale section 3). La phase de classification est un point clef de cette approche. Elle va permettre l'identification de classes qui seront ensuite reliées par un expert à l'aide d'informations/capteurs complémentaires à des états environnementaux (Figure 3.1).

Dans ce chapitre, une nouvelle méthode de classification non supervisée, appelée Multi-level Spectral Clustering (*M-SC*) est proposée pour effectuer cette classification. Cette nouvelle méthode basée sur une approche spectrale est composée d'une architecture multi-couche afin d'extraire la quantité maximale d'informations des données. En effet, le choix d'une architecture profonde a

été fait pour gérer l'information à plusieurs échelles et donc permettre la caractérisation aussi bien des schémas généraux que des événements extrêmes. Cette classification automatique vise à améliorer la définition et la compréhension des variations dans les séries temporelles multivariées. Elle permet ainsi d'alléger la phase de la labellisation inhérente aux phases d'interprétations et au système de prédiction tout en améliorant la définition active de ces événements nouveaux ou rares.

Tout d'abord nous détaillons donc le principe de la nouvelle méthode *M-SC* et ses caractéristiques ainsi que les étapes de pré-traitements et les optimisations algorithmiques réalisées au cours de cette thèse (Section 3.2). Le pré-traitement est une phase essentielle à réaliser avant toute classification. Ici, un protocole en quatre étapes est établi. Il permet de s'assurer que la base de données est au format "idéal" avant de l'utiliser comme base pour la classification. De même, au cours du développement de la méthode, des algorithmes pré-existants ont été améliorés afin de rendre optimale la classification.

Puis, les performances de l'approche *M-SC* sont évaluées (Section 3.3.3). Pour cela l'algorithme *M-SC* est comparé à plusieurs méthodes de classification non supervisées dont le concept est détaillé Chapitre 2, section 2.4.2. Un panel de méthodes est donc sélectionné en fonction de leurs caractéristiques : des approches directes (*e.g.* *K-means*, *SC*, ...) ou des approches hiérarchiques comme (*e.g.* *HC*, le *H-SC*, *Bi-SC*, ...) (Section 3.3.3.1).

Enfin, les réponses de la méthode sont analysées sur un premier cas concret en Manche Orientale. D'autres études répondant à des problématiques identiques aux nôtres ont déjà été publiées sur cette zone [Rousseeuw et al. 2015b, Caillault-Poisson et Lefebvre 2017, Phan et al. 2017)]. Dans la continuité de ces travaux, la station MAREL-Carnot a donc été définie comme site test. Dans ce sens, il est considéré comme le site de référence tout au long de l'étude.

La section 3.4 présente le jeu de données, la calibration et les sorties de la classification multivariée. Les informations multivariées issues de la classification permettent d'avoir une vue d'ensemble sur les facteurs de contrôle liés à chaque classe. À ces informations, une analyse taxonomique BF est associée. Sur la base des facteurs d'influence (hydrodynamiques et biogéochimiques) et les variations phytoplanctoniques, détaillés chapitre 1 section 1.2.2, les classes sont reliées à des états environnementaux et labellisées comme telles.

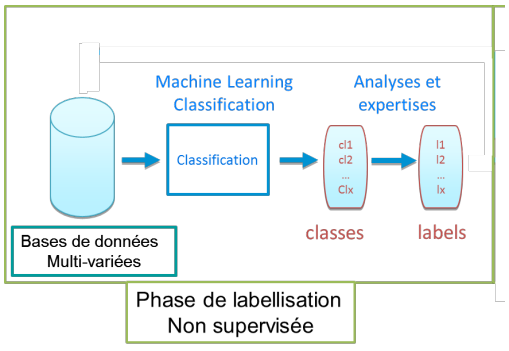


Figure 3.1 – Schématisation de la phase de classification de l'approche semi-supervisée.

3.2 Nouvelle approche non supervisée : La Classification spectrale multi-niveau

Cette section présente une nouvelle méthode développée pendant cette thèse [Grassi, Poisson-Caillault et Lefebvre 2019] (publication en annexe F). Dans la première partie sont définis le concept, le principe de l'algorithme et les caractéristiques principales de la méthode. La deuxième partie décrit le protocole de pré-traitement. Enfin, la troisième partie fait un point sur les problématiques rencontrées et les solutions et développements apportés.

3.2.1 Présentation de la méthode : Multi-level spectral clustering (M-SC)

M-SC est un algorithme de classification non supervisée. Basé sur une classification spectrale profonde (Algorithme 9, 10, il permet une segmentation implicite multi-niveau d'une série temporelle aux épisodes singuliers ou usuels. Le principe est de combiner une méthode spectrale et une architecture profonde. Les caractéristiques des méthodes spectrales, définies dans la section 2.4.2.3, répondent en grande partie à la problématique du manque de connaissances *a priori* sur les données puisque ce sont des méthodes non supervisées. De plus, elle permet de s'affranchir de la forme complexe de nos données. Ensuite, le choix d'une architecture profonde a été fait pour gérer l'information en partant des changements globaux jusqu'aux événements spécifiques, extrêmes, intermittents ou rares. D'un point de vue algorithmique, la méthode *M-SC* combine les points forts des méthodes spectrales et hiérarchiques (Tableau 3.1) afin d'optimiser la définition des classes dans les bases de données HF pour aller vers de la détection des évènements extrêmes.

Le principe consiste à appliquer récursivement une classification spectrale normalisée semblable à *NJW-SC* [Ng et al. 2001] (Rappel de la méthode section 2.4.2.3) (Figure 3.2, Algorithme 10), avec des valeurs de K variables et estimées automatiquement pour chaque niveau (Algorithme 9). Les classes du niveau supérieur $(Niv+1)$ correspondent aux sous-états du niveau précédent (Niv). Ainsi, elle segmente les séries temporelles multivariées X (M signaux de N échantillons de temps) de manière non supervisée sur plusieurs niveaux.

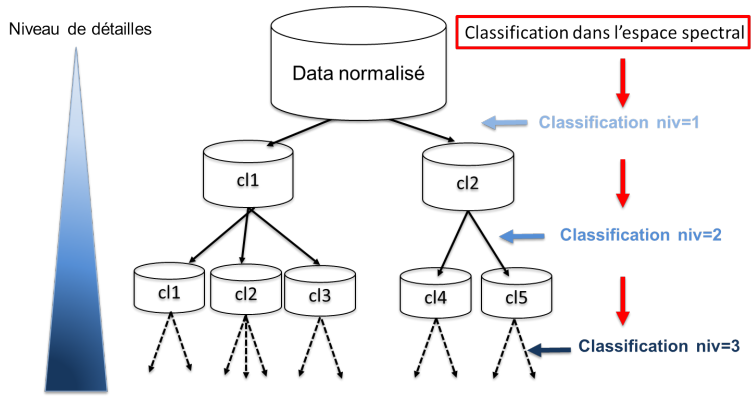


Figure 3.2 – Représentation schématique du processus de classification spectrale profonde : M-SC.

Pour faciliter la séparation des données, la méthode spectrale travaille dans l'espace spectral ou espace des vecteurs propres. Séparer les données revient à maximiser un critère de coupe ici choisi normalisé. Ce critère de coupe normalisé suit le principe de la classification *NJW-SC*. Il fait apparaître une matrice Laplacienne L issue de la matrice de similarité S qu'il est possible de résoudre par un système de valeurs propres standards. Les valeurs propres e et vecteurs propres v

sont alors extraits de la matrice Laplacienne L de S selon la coupe choisie. Ici, le critère de coupe normalisée est défini pour maximiser les valeurs propres. Ces vecteurs propres v représentent le nouvel espace de caractéristiques (espace spectral) et ils constituent l'entrée de l'algorithme de classification non supervisée.

Ainsi, notre méthode de classification spectrale est proche de l'architecture *NJW-SC* [Ng et al. 2001]. Pour plus de robustesse dans l'algorithme *M-SC*, la fonction *K-means* de *NJW-SC* est remplacée par sa version robuste *Partition Around Medoids (PAM)*, également appelée *K-médoide* qui est préférée pour les grandes bases de données afin d'éviter un calcul du centre de gravité et de conserver un représentant existant dans les observations (Algorithme 9). De manière générale, c'est cette version plus robuste, que nous appelons Spectral-PAM (Algorithme 9), qui sera donc appliquée pour segmenter chaque série temporelle. Lorsque la taille du jeu de données est grande, une version "Fast" [Rousseeuw et al. 2015b] (Section 2.4.2.3) est utilisée.

Au premier niveau, l'entrée de *M-SC* correspond à l'ensemble des N points $X = \{x_n \in \mathbb{R}^M, n \in [1,.,N]\}$, puis au niveau suivant chaque entrée est réduite aux x_n points de chaque classe du niveau précédent. Par conséquent, la sortie *M-SC* est une matrice Cl classes où $cl[n,niv]$ correspond au numéro de classe de l'observation x_n au niveau Niv. La profondeur maximale du nombre de niveaux ($NivMax$) de *M-SC* dépend de l'utilisation prévue et du standard d'interprétation requis et doit être définie en entrée de l'algorithme (Définition 12).

Définition 12 *Soit pour un ensemble $X = \{x_{nm}, n \in [1,\dots,N], m \in [1,\dots,M]\}$ de M séries temporelles de N échantillons de temps, l'entrée M-SC est la matrice de Similarité $S(X)$ obtenue à partir de l'ensemble des observations X. À chaque niveau Niv et à chaque classe cl, l'algorithme est réitéré sur les observations concernées, alors la matrice de Similarité est recalculée sur l'ensemble $X' \subset X$.*

À chaque niveau (Algorithme 9), v et e sont extraits de la matrice Laplacienne normalisée L pour réaliser la classification. De même, pour chaque niveau et pour chaque sous-état le nombre de classes K est déterminé automatiquement, selon un critère (crit) dépendant de ces valeurs propres e (Définition 13) (Détails section 3.2.3.1). Il est important de noter que ce n'est pas une séparation binaire comme c'est le cas pour d'autres méthodes tel que *Bi-SC* de Shi et Malik 2000. Ce n'est pas non plus une segmentation hiérarchique comme dans l'algorithme *H-SC* Sanchez-Garcia, Fennelly, Norris et al. 2014.

Définition 13 *Dans l'algorithme M-SC, le nombre de classes (K)(Algorithme 10) choisies pour la classification sera défini par le critère du gap (GAP) ou par une sélection des valeurs propres dominantes Principal EigenValues (PEV) (Section 3.2.3.1). Seules les valeurs propres dominantes ($e > 0,9$) sont considérées comme assurant une bonne séparabilité des classes [Ng et al. 2001]. Ainsi, avec le critère GAP, le nombre de classes K est calculé à partir de l'écart maximal entre deux valeurs propres dominantes successives e. Avec le critère PEV, le nombre de classes K est calculé à partir du nombre de valeurs propres égales à 1.*

De plus, une classe est coupée au niveau suivant uniquement si ces observations ne forment pas une cohésion/connexité optimale. Contrairement aux approches hiérarchiques descendantes qui subdivisent le jeu de données jusqu'à ce que chaque feuille corresponde à une observation, cette cohésion est définie par un critère de silhouette (sil.min) (définition de silhouette section 2.4.4). A chaque niveau et pour chaque classe un coefficient de silhouette est calculé dans l'espace initial ou initial réduit si l'on est en Fast-spectral. Si cette silhouette est supérieure au critère sil.min, la classe est considérée comme assez connexe et ne sera plus subdivisée par l'algorithme de classification dans les niveaux plus profonds ($Niv+1$) (Définition 14).

Définition 14 *Pour chaque sortie de l'algorithme M-SC $cl[n,niv]$ est réalisé un calcul de silhouette sil. Tant que cette silhouette est inférieure au critère d'arrêt sil.min alors la classification continue. Dans le cas contraire, la segmentation est stoppée et la classe est conservée dans les niveaux plus profonds.*

Algorithme 9 Classification Spectral-PAM avec détermination de K pour 1 niveau

Require: $S(X)$ a $N \times n$ matrice Gram , Kmax maximum de clusters acceptés
Variable : W, D, L, e, v
$W = S$ similarité avec pondération
 # Calcul du Laplacian
$\forall i, w_{ii} = 0$
Construction de $D = d_{ij}$ la matrice diagonale $n \times n$
$d_{ii} = \sum_j w_{ij}$ et $d_{ij,i \neq j} = 0$
$L = D^{-1/2} W D^{-1/2}$ la matrice Laplacienne
{e,v}=eigen(L,Kmax) extraction des valeurs et des vecteurs propres
 # Calcul du nombre de clusters K
if GAP **then**
$K = \underset{i-1}{\text{argmax}} \ (e_i - e_{i-1}),\ e_i > 0.9,\ i \geq 2$
end if
if PEV **then**
$K = sum(e_i = 1),\ e_i > 0.9,\ i \geq 2$
end if
 # Classification dans l'espace des vecteurs propres
Sélection des K plus grands vecteurs propres v de L ;
Former $V = [v_1 v_2 \dots v_K]$ matrice $N \times K$
Former Y matrice de normalisation des lignes de V.
cl = PAM (Y,K)
return cl

Algorithme 10 Multi-Level Spectral Clustering

Require: X, NivMax, Kmax, sil.min, $S(X)$ a $N \times n$ matrice Gram
Variables : W, cl, sil, clusterToCut=1, niv=1, stop=false, groups, k
$W = S$ similarité avec pondération
 # Initialisation
cl[n,niv]=1 matrice $N \times NivMax$
 # Clustering par niveau
while (stop !=false) **do**
for k in clusterToCut **do**
Calcul de la similarité W de $X' = \{x_n \in X | \text{cl[n,niv]=k}\}$
groups=Spectral-PAM(W, Kmax=card(X'))
$\forall n | \text{cl[n,niv]=k}$, cl[n,niv+1]= groups+card(clusterToCut)+1
end for
 # calcul de la silhouette des nouvelles sous-classes
clusterToCut={} # initialisation : vecteur nulle
for k $\in$ unique(cl[n,niv+1]) with n $\in$ [1,. . . ,N] **do**
$X' = \{x_n | cl(n, niv+1) = k\}$
sil=silhouette(X') # means of point silhouette
if sil<sil.min **then**
clusterToCut= {clusterToCut,k} # compteur des k classes à découper
end if
end for
stop= ((niv+1)$\geq$NivMax) | (card(clusterToCut)==0)
niv=niv+1
end while
return cl matrix of $N \times$niv

Caractéristiques	**Valeur**
Prise en compte du contexte	Non supervisé
Connaissances a priori	Matrice de similarité
Méthode	basée des matrices de similarité
Présentation des résultats	K classes + hiérarchique
Complexité	O(M*N*K)
Déterminisme	non
Incrémental	oui
Hard vs soft	Hard
Tolérance au bruit	non
Effet de chaine	oui
Tolérance aux clusters de tailles variées	oui
Tolérance aux clusters de densités variées	oui
Tolérance aux clusters concentriques	oui
Tolérance aux clusters convexes et non linéairement séparables	oui

TABLEAU 3.1 – Tableau des caractéristiques : Multi-level Spectral clustering. En vert les caractéristiques satisfaisantes ; En rouge les caractéristiques non satisfaisantes.

3.2.2 Protocole de pré-traitement

Il existe quelques étapes essentielles de pré-traitement des données avant toutes étapes de classifications. Ces étapes sont nécessaires pour avoir une échelle de temps régulière et identique, pour limiter les valeurs hors gamme, le nombre de données manquantes ou le bruit dans la série. Dans notre cas, elles sont aux nombres de quatre : l'alignement, la correction, la complétion et la normalisation des données (Figure 3.3). L'étape de correction est divisée en 2 sous-étapes : la correction des valeurs aberrantes et la correction du bruit. SCHMITT et LEFEBVRE 2016 ont démontré que les séries MAREL n'étaient ni gaussiennes ni bruitées quelle que soit l'échelle de temps retenue. Ainsi l'étape de correction de bruit n'est pas réalisée dans ce cas d'étude, mais c'est un point à vérifier avant d'utiliser les méthodes spectrales. Pour les autres étapes, un protocole automatique a été mis en place. Les modalités de chacune des étapes sont présentées dans les sections suivantes.

3.2.2.1 Alignement temporel

Pour faciliter l'étude de la dynamique temporelle, le choix a été fait d'utiliser un intervalle de temps identique entre chaque mesure. Cela simplifie les comparaisons entre les variables, les calculs d'indices et des statistiques temporelles relatives à chaque état (duré, date de début et de fin, ...). En effet, les mesures des différents capteurs ne se font pas simultanément, ce qui crée un décalage temporel pouvant aller de quelques secondes à quelques minutes ou heures. De plus, les séries possèdent dans certains cas des réplicats. Ainsi les bases de données sont toutes synchronisées sur un intervalle de temps moyen.

Le protocole d'alignement temporel commence par calculer les pas de temps moyens de la série temporelle pour chaque paramètre. Ensuite, l'intervalle moyen le plus petit est sélectionné. En fonction de ce pas de temps une variable temps idéale (régulière/sans réplicat) est calculée. En cas de réplicats dans l'intervalle de ce pas de temps moyen, l'utilisateur choisira de retenir une

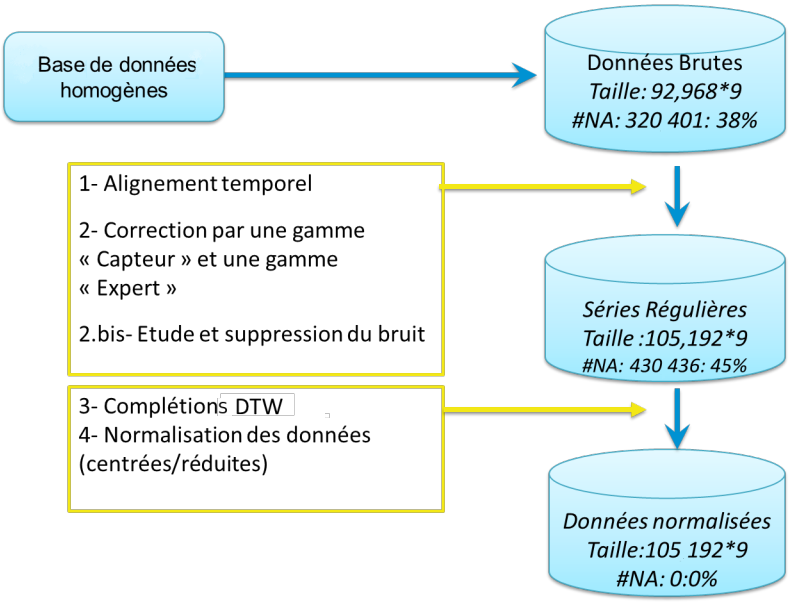


FIGURE 3.3 – Représentation schématique du protocole de pré-traitement. Taille correspondant à la dimension de la matrice de données sur la période choisie et #NA aux nombres de données manquantes.

valeur "idéale" en fonction de son application ou de la variable, comme le minimum, le maximum la moyenne entre les réplicats. Ici, la valeur max a été retenue. Elle nous assure de ne pas lisser les données, ce qui est important pour l'identification des évènements extrêmes.

La routine développée permet alors le calcul automatique du pas de temps et de fournir une série régulière. Elle permet un alignement à 1, 10, 20, 30 minutes, 1 heure, 1 et 2 jours, 1 semaine, 1 mois et 1 an.

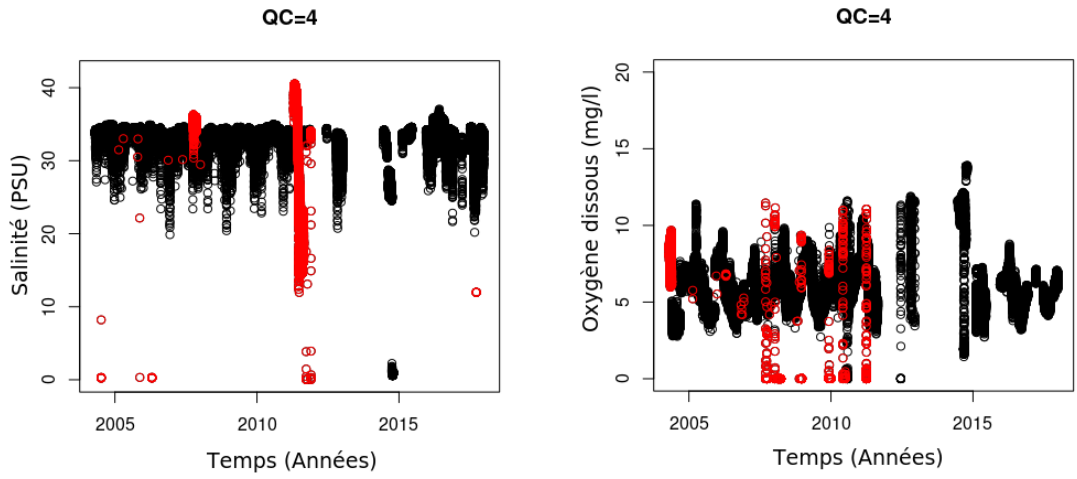
3.2.2.2 Correction de gamme et code qualité

Sur la plateforme Coriolis (2.1.1), l'ensemble des données brutes non qualifiées et le code qualité (ou *Quality Code (QC)*) (Tableau 3.2) associé sont accessibles.

TABLEAU 3.2 – Code qualité (QC) associé à chaque donnée et sa signification.

Code qualité	Signification
0	Pas de code qualité
1	Bonne
2	Probalement Bonne
3	Fausse mais correction possible
4	Fausse
5	Modifier
6	//
7	Valeur nominales
8	Valeurs interpolées
9	données manquantes

Le QC est la seule information dont nous disposons sur la qualité de chaque mesure. Ce code est normalement attribué après que les procédures de contrôle de qualité soient effectuées. Lors de cette procédure les données sont vérifiées automatiquement, puis une validation ou des modifications peuvent être réalisées par un expert. Pour notre approche nous souhaitons rester au plus proches des données brutes ce qui nous assure de contrôler la phase de modification et de suppression des données aberrantes ou hors gamme. C'est pourquoi, nous avons choisi de considérer pour nos travaux l'ensemble des données et écarter uniquement les données dont le niveau de qualité est égal à 4 (Données Fausses). La figure 3.4a est un exemple de correction via le $QC = 4$ sur le signal de Salinité. Toutefois, la procédure QC n'est pas toujours fiable, les experts ne peuvent pas vérifier toutes les mesures et, une valeur fausse peut être identifiée sous un autre code comme le $QC = 2$ ou 3. Par exemple, les mesures d'oxygène égales à 0 (Figure 3.4b) vers 2012 semblent fausses mais ne correspondent pas au $QC = 4$. De plus, certaines bases de données ne possèdent pas de QC. Nous souhaitions donc un moyen d'écarter de manière automatique une grande partie des possibles valeurs aberrantes sans utiliser le QC. Pour cela, le protocole par correction de gammes "capteur" et "expert" suivant est mis en place.



(a) Salinité (PSU)

(b) Oxygène ($\mathrm{mg\,l^{-1}}$)

FIGURE 3.4 – Séries temporelles MAREL-Carnot entre 2004 et 2017, A) de la salinité PSU et B) de la concentration en oxygène $\mathrm{mg\,l^{-1}}$. Il est représenté en noir : le signal initial est en rouge : les mesures identifiées par le code qualité QC=4.

La gamme "capteur" est un intervalle de valeurs dites correctes et définies sur les notices d'appareils de mesure, la gamme "expert" est un intervalle de valeurs dites correctes et définies par un expert du domaine. Notons que la gamme "expert" étant définie comme plus précise que la gamme "capteur", celle-ci valide ou modifie par un intervalle plus strict le niveau de qualité de la gamme "capteur". Sur l'ensemble des paramètres, ne sont retenues que les valeurs comprises dans la gamme, alors que les valeurs hors gammes sont remplacées par des données manquantes (*not Available (NA)*).

Les gammes de MAREL-Carnot représentées (Tableau B.1) et les gammes MAREL-Iroise, MesuRho (Annexe B), sont établies à partir des fiches techniques pour la gamme "capteur" et à partir de l'analyse descriptive de chaque base de données pour la gamme "expert". L'analyse descriptive a permis de dégager les paramètres centraux et de dispersion de la série temporelle

(minimum, maximum, moyenne, médiane, écart-type), les tendances et les valeurs particulières (extrême, ou erreur) des données brutes.

Tableau 3.3 – Gamme "capteur" et "expert" pour le système de mesures MAREL-Carnot

Noms Paramètres	Coriolis	Unités	Gamme Capteur min	Gamme Capteur max	Gamme Expert min	Gamme Expert max
Température	TEMP.LEVEL1	°C	-5	35	0	30
Salinité	PSAL.LEVEL1	PSU	2	42	5	35
Oxygène dissous	//	$mg\,l^{-1}$	0	30	0	20
Oxygène dissous	DOX1.LEVEL1	$ml\,l^{-1}$	//	//	//	//
Saturation en Oxygène	OSAT.LEVEL1	%	0	150	0	150
Fluorescence	FLU3.LEVEL1	FFU	0	500	0,03	150
Turbidité	TUR4.LEVEL1	NTU	0	1500	0	150
PAR	LGH4.LEVEL1	$\mu mol\,m^{-2}\,s^{-1}$	//	//	0	2500
Direction du vent	WDIR.LEVEL0	°	0	360	0	360
Vitesse du vent	WSPD.LEVEL0	$m\,s^{-1}$	0	40	0	40
Nitrate	NTRZ.LEVEL1	$\mu mol\,l^{-1}$	0	100	0	100
Phosphate	PHOS.LEVEL1	$\mu mol\,l^{-1}$	0	100	0	10
Silicate	SLCA.LEVEL1	$\mu mol\,l^{-1}$	0	100	0	50
Niveau de la mer	SLEV.LEVEL1	m	//	//	0	20
Température de l'air	DRYT.LEVEL0	°C	-20	40	-20	40

3.2.2.3 Complétions

De manière générale, les fonctions de classification spectrale (Section 2.4.2.3) ne tolèrent pas les données manquantes (*NA*). Il en est de même pour la fonction *M-SC*. C'est pourquoi une étape de complétion est ajoutée au protocole de pré-traitement. Cette étape n'est pas obligatoire mais fortement conseillée, car un jeu avec trop de données manquantes (notées *NA-Not Avalaible*) peut considérablement modifier la distribution du signal et, donc les résultats. En effet, les données manquantes (*NA*) sont un problème récurrent dans les bases de données. L'information fournie est incomplète et les analyses sont donc irrégulières et moins fiables. Cette problématique est accentuée pour des séries temporelles HF : une série de données manquantes d'une journée dans des séries échantillonnées toutes les 20 minutes (comme dans notre cas) correspond à 72 *NA*, une semaine 504 *NA* consécutifs ... Ajouter à cela la variabilité et le bruit dus à la HF, la complétion s'en retrouve d'autant plus complexe. Le choix de la méthode a donc son importance. De plus, dans l'algorithme de *M-SC* (qui est une classification multivariée), toutes les lignes contenant au moins un *NA* sont considérées par défaut comme non utilisables. Ainsi, toutes ces lignes (*i.e.* ici une date (m) pour n variables) sont automatiquement supprimées de la matrice de données. Le nombre de *NA* dans le fichier limite donc considérablement la méthode. Pour notre cas d'étude (Basses de données MAREL-Carnot), 100 % des lignes du fichier contiennent au moins 1 *NA* par ligne, de ce fait 0 % du fichier est exploitable. La complétion est donc un point clef de notre protocole de pré-traitement.

Cette importante étape méthodologique a été réalisée en collaboration avec l'Université du Littoral Côte d'Opale (ULCO) et du Laboratoire d'Informatique Signal et Image de la Côte d'Opale (LISIC), dans le cadre de la thèse de T.T.H Phan [Phan et al. 2017]. Elle a travaillé sur une méthode de complétion particulière : la complétion élastique (en anglais *Dynamic Time Warping (DTW)*). Lors de son travail, deux packages R d'imputation monovariée : *theDTWBI* (R-CRAN:: DTWBI) et multivariée : *the Dynamic Time Warping Based Multivariate Imputation (DTWBI)* (R-CRAN:: DTWUMI) ont été développés. Ces travaux démontrent que cette méthode permet d'obtenir une complétion qui préserve la dynamique des signaux complexes dans des séries avec une quantité conséquente de données manquantes consécutives. Ces caractéristiques

répondant en grande partie à nos problématiques de complétions, cette méthode a donc été choisie pour notre protocole de pré-traitement.

Le principe de cette méthode est de rechercher des dynamiques similaires dans le signal pour ensuite compléter les données manquantes par le signal lui-même. Ainsi, pour chaque série de *NA* est définie une fenêtre de recherche de la taille de cette série (T_gap). Cette fenêtre correspond au signal avant ou après la série de données. Puis, une requête avant ou arrière, c'est-à-dire un balayage de la série, est réalisée sur le signal afin de trouver une correspondance. La recherche et la sélection d'une correspondance sont effectuées seulement si le critère de cross-corrélation est suffisant. Ce critère de cross-corrélation permet de s'assurer que la série possède une redondance suffisante pour se permettre de compléter la série. C'est aussi ce critère qui détermine en partie si une séquence est suffisamment similaire pour être sélectionnée pour la complétion.

Initialement, le package R : :DTWBI complète la série en utilisant seulement le signal avant ou arrière. Or il est possible qu'aucune correspondance ne soit trouvée. Dans le cadre de la thèse, la fonction a donc été modifiée pour effectuer des requêtes des deux côtés de la séquence manquante et ainsi augmenter le pourcentage de reconnaissance. La méthode de complétion exécute maintenant une première requête avant ou arrière en fonction de la série temporelle (N) la plus longue ($> N/2$). Puis, s'il ne trouve aucune correspondance avec la première requête (ex : avant), il va effectuer l'autre requête (ex ici : arrière) (Figure 3.5).

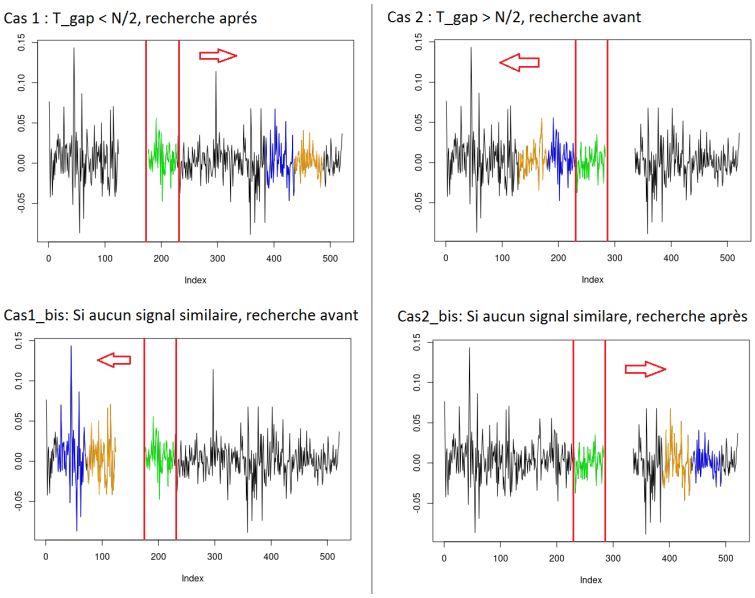


FIGURE 3.5 – Protocole de complétion des grandes séquences par DTW. En vert la requête, en orange le signal similaire et en bleu le signal qui sera imputé dans l'intervalle des données manquantes. Exemple présenté sur un signal tiré du package *DTWBI*. Cas 1 : requête arrière initiale ; Cas 1 bis : requête avant si aucune correspondance n'est trouvée au Cas 1. Cas 2 : requête avant initiale ; Cas 2 bis : requête arrière si aucune correspondance n'est trouvée au Cas 1.

La version initiale de T.T.H Phan (*the DTWBI* (R-CRAN:: DTWBI) est performante et permet de compléter 100 % du fichier MAREL-Carnot. Mais l'algorithme complète des séries de données manquantes de quelques jours à plus de 3 ans sans faire de distinctions. Or compte tenu des échelles de temps de la dynamique phytoplanctonique (Section 1.2.3, Figure 1.2) une complétion de plus de 3 ans de données est trop importante. C'est la raison pour laquelle, lors de

cette thèse une version contrainte de la fonction *DTWBI* initiale est développée.

Dans cette version deux paramètres (smallHole et acceptedHole) sont introduits. Ils permettent d'adapter la taille des fenêtres de complétion.

smallHole est le nombre de données manquantes (*NA*) consécutives que l'on considère comme petit, c'est-à-dire inférieur à la dynamique de changement du phénomène étudié (fenêtres à valeurs monotones).

acceptedHole est le nombre de données manquantes consécutives, que l'on considère comme trop grand en ce qui concerne le changement de la dynamique du processus, pour être complété.

Afin de définir au mieux ces deux paramètres une recherche d'optimale via une analyse de sensibilité est réalisée (Détails et figure(s) en annexe D). Cette étape a pour but de trouver un compromis entre (i) un nombre de données suffisamment important pour que la classification ait un sens et (ii) une taille de fenêtre suffisamment petite pour être cohérente avec le sujet d'étude. En effet, une semaine correspondant à 504 points le choix d'un smallHole trop grand pourrait lisser notre signal. De même, le choix d'un acceptedHole trop grand peut recréer des cycles complets de notre dynamique. Mais il faut noter que plus la séquence à compléter est grande, plus il est difficile de trouver un critère de cross-corrélation suffisant.

Six tests avec des tailles de fenêtres (smallHole et acceptedHole) différentes sont réalisés. Le choix de la taille des fenêtres est fait en fonction des périodes clefs de la dynamique (journalières, mensuelles, saisonnières). Pour cette étude, 10 variables (Oxygène-dissous, salinité, température, hauteur d'eau, PAR, turbidité, Nitrate, Phosphate, Silicate, fluorescences) sont sélectionnées. Comme pour le protocole de pré-traitement classique, le fichier est aligné et corrigé par les gammes "capteur" et "expert".

Les résultats majeurs sont résumés dans le tableau 3.4 et l'étude de sensibilité complète est présentée en Annexe D.

Tableau 3.4 – Récapitulatif des tests de sensibilité pour les paramètres de l'algorithme de complétion DTWBI. En gras la paramétrisation choisie.

tests	smallHole	acceptedHole	% données utilisables
Non complété	-	-	0,00
DTWBI	1 semaines	∞	100
1	1 jour	15 jours	55,38
2	1 jour	1 mois	55,43
3	1 jour	2 mois	55,43
4	1 jour	6 mois	55,43
5	1 semaine	1 mois	70,57
6	**3 jours**	**1 mois**	**66,70**

D'après le tableau 3.4, un acceptedHole d'un mois permet une augmentation des données utilisables de 55,38 % à 55,43 %. Si l'on augmente encore acceptedHole, le pourcentage de données utilisables reste constant. Il semble que la méthode *DTWBI* ne trouve plus de correspondance au-delà de cette limite de 1 mois. C'est donc un acceptedHole de 1 mois qui sera retenu. Toutefois, une augmentation de 0,5 % n'est pas suffisante. Un smallHole de 1 jour (soit 72 points) semble trop restrictif. Une augmentation de smallHole à 1 semaine (soit 504 points) et 3 jours (soit 216 points) augmente le taux de données utilisables. Cependant, il est souhaitable de limiter la taille des trous à compléter par moyenne mobile pour éviter les effets de palier. De ce fait, nous retenons la configuration 6 (smallHole = 3 jours et acceptedHole = 1 mois) comme solution optimale pour

notre cas d'étude.

Ces modifications permettent d'augmenter les chances de trouver une correspondance dans le signal et limite la complétion de trop grandes séries de *NA*. Mais les méthodes basées sur *DTW* nécessitent un temps et une capacité de calculs assez importants. Pour réduire ce temps de calcul, T.T.H Phan réalise dans ses travaux une première phase de complétion par moyenne mobile sur les jeux de données avant d'appliquer la fonction DTW sur les données. Mais dans le package, cette approche n'est pas proposée. Pour améliorer la méthode, une nouvelle version est proposée. Ainsi trois étapes de complétions sont définies :

1. la complétion des valeurs de NA-isolées ($x_{n,m} = NA$) se fait par moyenne des valeurs directement adjacentes.
2. la complétion des données manquantes consécutives ($x[n-T, n:m]$) se fait par moyenne mobile pondérée (soit une taille de trou T $\leq$ au critère smallHole),
3. la complétion des longues séquences de données manquantes se fait par *DTW* (soient toutes les séquences tel que smallHole $<$ T $<$ acceptedHole).

Ainsi une petite séquence (étape 2) correspondra à une séquence $\leq$ au critère smallHole et une séquence sera considérée comme grande (étape 3) lorsqu'elle sera $\geq$ au critère smallHole. Ces nouveaux paramètres permettent de définir la taille des fenêtres de complétions et donc de contraindre et d'adapter le protocole de complétion en fonction du problème. Elle offre la possibilité de modifier les phases de complétion par moyenne mobile ou par *DTWBI*, ce qui aura pour conséquence de modifier les taux de reconnaissance de la fonction et donc l'efficacité de complétion.

3.2.2.4 Normalisation

La normalisation a de nombreuses applications dans la fouille des données. Dans notre cas c'est un préalable important pour l'application des algorithmes de classifications non supervisées (chapitre 2 section 2.4.2). Elle va standardiser les données et va permettre d'obtenir des données indépendantes de l'unité ou de l'échelle choisie.

La normalisation va centrer les données en soustrayant aux données leurs moyennes empiriques (μ) et va réduire les données en divisant toutes les valeurs par l'écart type (σ).

$$x = \frac{X - \mu}{\sigma} \tag{3.1}$$

Une base de données centrée-réduite aura l'avantage d'égaliser le poids de chaque dimension (de chaque descripteur). Cela revient à réaliser un changement d'unité qui standardise le jeu de données sans modifier les profils de variations.

3.2.3 Nouveaux ajouts méthodologiques

Pour chacune des problématiques rencontrées, lors de la mise en place et de l'automatisation de cette méthode et de son protocole, les solutions et les développements suivants ont été apportés :

3.2.3.1 Définition automatique des K-clusters

L'estimation du nombre de groupes nommés K-classes est un point fondamental de notre approche non supervisée. Il existe plusieurs méthodes pour le déterminer :

— L'analyse des amplitudes des valeurs propres :

- technique des *PEV* : détermination du nombre de K par l'énumération des valeurs propres principales dominantes (c'est-à-dire égales ou proches de 1) ;
- technique du "GAP" : définition du nombre K, tel que l'écart entre les valeurs propres soit maximal ;

— L'analyse des vecteurs propres :
- par la recherche d'une base de vecteurs propres orthogonaux robustes [Kong et al. 2013 ; Zelnik-Manor et Perona 2004] ;
- par le processus itératif de sélection des K premiers vecteurs propres tel qu'aucune donnée dans l'espace spectral à K vecteurs soit proche de l'origine [Sanguinetti et al. 2005] ;
- par un processus d'*Expectation Maximization* (EM) sur un critère de vraisemblance basé sur la capacité de chaque vecteur propre à séparer les données [Xiang et Gong 2008].

C'est la technique du "GAP" qui a été retenue dans la thèse de Kevin Rousseuw [Rousseeuw et al. 2015a] (équation 3.2.3.1), car elle est l'une des plus simples à implémenter et la plus rapide.

Une autre option est ajoutée dans le nouveau code : la technique des *PEV* (équation 3.2.3.1, Algorithme 9). Cette méthode a l'avantage d'être plus restrictive et donc de limiter dans les couches profondes le nombre de classes. Le choix de l'une ou de l'autre permet plus de flexibilité en fonction du jeu de données. Cela permet d'adapter le calcul en fonction du nombre, de la taille et de la représentativité de chacun des groupes présumés. Un critère de coupe (crit) est ajouté pour permettre de choisir une de ses deux options en fonction de la problématique de recherche.

Dans le cas de crit égale à GAP :

$$K = \arg\max_i (e_{i+1} - e_i) \tag{3.2}$$

Dans le cas de crit égale à PEV :

$$K = card(e_i = 1) \tag{3.3}$$

3.2.3.2 Limitation de la sur-segmentation

Un critère d'arrêt (sil.min) est défini pour stopper la segmentation d'une classe lorsque celui-ci est bien isolé. En effet une classe peut être isolée dès les premiers niveaux. Il n'est donc pas toujours nécessaire d'aller vers des niveaux plus profonds. Sans ce critère, les niveaux plus profonds permettaient donc d'identifier les classes extrêmes, mais ils sur-segmentaient les classes plus génériques.

Ce critère est basé sur la notion de Silhouette. Pour rappel, le score Silhouette, R-CRAN:: cluster : silhouette [Maechler et al. 2018] est une mesure de la similarité entre un objet et sa propre classe (cohésion) par rapport aux autres classes (séparation). Une grande Silhouette (valeur proche de 1) représente une classe très bien groupée, tandis qu'une petite Silhouette (valeur proche de 0) signifie que la distance entre des observations de deux classes est faible et peut engendrer de la confusion.

Dans notre cas, si la Silhouette d'une classe est supérieure à sil.min, cela signifie que notre classe est suffisamment connexe (soit suffisamment représentatif) ; À ce moment, la segmentation est stoppée pour cette classe. Réciproquement, la segmentation continuera si la Silhouette d'une

classe est inférieure à sil.min (Algorithme 10). Afin d'être au plus proche des classes, le calcul de la silhouette s'effectue dans l'espace spectral dans la fonction de classification. Ce choix offre aussi l'avantage d'optimiser le temps de calcul. Bien entendu, la définition du critère est dépendante du cas d'étude et peut être ajustée. Dans notre cas le critère d'arrêt sil.min est fixé à 0,7 par défaut. De manière générale, c'est un bon compromis entre la sur-segmentation et la détection de classe de taille restreinte.

3.3 Analyse comparative et validation sur plusieurs jeux de données tests

Cette section une analyse comparative est réalisée [Grassi, Poisson-Caillault, Bigand et al. 2020] (publication en annexe F). Nous comparons les performances de notre méthode avec les méthodes usuelles présentes dans la littérature (Rappelées et détaillées Section 2.4.2 chapitre 2) sur différents jeux de données tests. Le but est d'évaluer la capacité de chaque méthode à proposer une classification efficace. Ainsi, il sera possible de replacer notre nouvelle méthode dans un contexte global et donc de confirmer ou d'infirmer son efficacité.

Cette section détaille l'ensemble des jeux de données labellisés. Ensuite le processus de validation : choix des méthodes pour la comparaison, le paramétrage de chacune de ces méthodes et le calcul des indices de performance sont décrits (Détails Section 2.4.2 chapitre 2). Puis, les résultats de chaque méthode sont comparés et commentés.

3.3.1 Présentation des jeux de données.

Afin de réaliser une analyse globale, des jeux de données spatiales et temporelles sont sélectionnés (Tableau 3.5). Chaque jeu de données est labellisé, c'est-à-dire, que chacun des points est rattaché à une classe connue. Cette classification sera considérée comme la réponse juste et sera donc notre point de comparaison. On l'appellera : le/les vrai(s) label(s) (*True labels*). Cette approche permet d'évaluer les performances de détection des classes via les jeux de données spatiales et la détection des tendances et des ruptures via les séries temporelles. Certains jeux de données sont des cas artificiels (art.) issus de bases de données de référence, d'autres sont des cas expérimentaux (exp.) découlant de bases de données *in-situ*. Chaque jeu de données est de dimension N observations $\times$ D variables. Les D variables sont constituées uniquement des variables contributives. Ainsi, les dimensions temporelles et spatiales ne sont pas incluses dans l'étape de regroupement.

Toutes les bases de données réparties spatialement proviennent de la plateforme des dépôts de bases de données de référence pour l'apprentissage : *UCI - Machine Learning Repository* [Dua et Graff 2017]. Deux jeux de données artificiels (*Aggregate* et *Compound*) et deux jeux de données expérimentaux (*Iris* et *Species*) sont choisis pour leurs caractéristiques géométriques. *Aggregate* est constitué de formes relativement simples, contrairement à *Compound* qui possède des formes bien séparées mais imbriquées. Ils ont respectivement six et sept classes et ont deux dimensions. *Iris* et *Species* sont des cas plus complexes avec un nombre de classes plus important. *Iris* est le cas le plus simple car il ne possède que trois classes (catégories de plantes) avec 50 objets (observations) par classe, alors que le jeu de données *Species* compte 100 classes (espèces) avec 16 observations par classe.

Les bases de données de séries temporelles sont constituées de données expérimentales issues de campagnes en mer et d'une base de données artificielles construite spécifiquement dans le cadre de cette thèse. La problématique de segmentation des séries temporelles est un point clef, il était donc

Tableau 3.5 – Caractéristiques des jeux de données : Appellation. L'appellation du jeu de données, Dom. le domaine de mesure soit exp.= expérimental, soit art. = artificiel, Dim. la dimension (N objets × M dimension), C le nombre de classes, Dist. le pourcentage de distribution de la plus petite classe (E = équirépartie). En gras : ensemble de données de séries chronologiques.

	Appellation.	Dom.	Dim.	C	Dist.
1	Aggregate [Dua et Graff 2017]	art.	788×2	7	4,31
2	Compound [Dua et Graff 2017]	art.	399×2	6	4
3	Iris [Dua et Graff 2017]	exp.	150×4	3	33,33 (E)
4	Species [Dua et Graff 2017]	exp.	$1,600 \times 64$	100	1,6 (E)
5	**Simulated** [Grassi, Poisson-Caillault et Lefebvre 2019]	art.	$1,000 \times 3$	4	3,2
6	**DYPHYMA leg1** [DYPHYMA 2012]	exp.	$2,032 \times 18$	3	12,20
7	**DYPHYMA leg2** [DYPHYMA 2012]	exp.	$3,285 \times 18$	3	11,96
8	**DYPHYMA leg3** [DYPHYMA 2012]	exp.	$5,599 \times 18$	3	7,30

plus judicieux de sélectionner des jeux de données similaires à notre cas d'étude. Ainsi, les bases de données expérimentales utilisées sont fournies par le programme DYPHYMA [DYPHYMA 2012](Section 2.2.1). Elles proviennent de trois périodes de campagnes (leg) différentes : leg1 (1 jour mi-avril), leg2 (3 jours fin avril) et leg3 (5 jours fin mai début juin). Pour les tests, toutes les variables physico-chimiques disponibles et les quatre concentrations d'algues provenant d'un fluorimètre spectral à longueur d'onde fixe (Algae Online Analyzer (AOA)) couplés au FerryBox sont sélectionnées (Section 2.2.3). Chacune de ces bases ont été labellisées par un expert dans le cadre de l'étude de [Lefebvre et Poisson-Caillault 2019] et les trois classes (indiquant la présence d'un assemblage d'au maximum 2, 3 ou 4 espèces) sont retenues.

Puis un jeu multivarié de données artificielles, que l'on appellera "*Simulated*" est élaboré pour représenter les différents types de signaux que l'on souhaite détecter dans les jeux de données *in-situ*. Le jeu de données simulé est donc construit à partir de 3 fonctions sinus $x_i(t) = a_i \times \sin(2\pi f_i t + \phi_i) + b_i(t)$ de 1000 points. Chaque paramètre $(a_i; f_i; \phi_i)$ est différent pour chaque signal. Le bruit ajouté $b_i(t)$ est généré à partir des distributions uniformes aléatoires avec quelques décalages pour certains points. Ces signaux forment notre signal général (gs) et représentent la tendance de nos données avec une variabilité saisonnière/inter-annuelle marquée. Ensuite trois événements courts sont introduits. Ainsi, deux fortes variations similaires aux événements extrêmes (ev1, ev2) sont imputées sur chaque signal $x_{i=1,2,3}$. Le dernier événement correspondant à un décalage en amplitude '*offset*' (ev3) est imputé seulement sur le signal x_3 afin de simuler une défaillance du capteur. Ainsi nous obtenons trois séries chronologiques libellées $X(t)$ qui représentent la dynamique potentielle d'une série biochimique *in-situ*.

3.3.2 Choix des indices de validation

Comme présenté dans le Chapitre 2 section 2.4.4, cinq indicateurs de performance sont sélectionnés afin d'évaluer l'efficacité de chaque méthode. Pour rappel, deux indicateurs du taux de détections sont retenus. Ils sont établis à partir du tableau de confusion entre la classification obtenue par la méthode (K) et les vrais labels (C) : (1) #Iso, le nombre de motifs bien isolés

après un vote majoritaire. Ainsi, un motif est considéré comme bien isolé s'il est représenté par plus de la moitié des observations positives réelles et, (2) la précision totale, définie à partir du tableau de confusion après vote majoritaire. Ensuite, trois scores conventionnels non supervisés (indice de connexité) sont ajoutés pour l'interprétation : index Rand ajusté, index Dunn, score Silhouette [Zhao 2012]. Ici, ils sont calculés dans l'espace brut quelles que soient les méthodes de clustering utilisées.

3.3.3 Validation de la méthode : Évaluation des performances de segmentation

3.3.3.1 Sélection des méthodes

De nombreuses approches de *clustering* peuvent être adaptées à la segmentation de données temporelles et/ou spatiales. Ces approches peuvent être distinguées en fonction de leurs façons de couper : par voie directe ou hiérarchique, et en fonction de leurs espaces de coupes : l'espace brut (à partir des données ou d'un noyau) ou l'espace spectral.

Afin d'effectuer la comparaison la plus intéressante possible, nous avons choisi des méthodes usuelles directes et hiérarchiques, spectrales ou non, qui font le lien avec notre méthode de *clustering*. En effet, *M-SC* combine les aspects hiérarchiques et spectraux. Le principe et les caractéristiques de chacune de ces méthodes sont détaillés dans la section 2.4.2 du chapitre 2.

Les approches directes sélectionnées sont :

- Les algorithmes *K-means* [Hartigan et Wong 1979] et *K-medoids* (aussi appelé *PAM*) qui partitionnent N observations en un nombre fixe K clusters en minimisant la variance au sein de chaque classe (Section 2.4.2, chapitre 2). Ils sont généralement efficaces dans le cas de partitionnement pour des classes convexes sans chevauchement et sont souvent utilisés pour la quantification vectorielle (réduction des données).

- Les approches de *clustering* spatial basées sur la densité *Density-Based Spatial Clustering (DBSCAN)* [Ester et al. 1996] dont l'algorithme s'appuie sur la densité estimée des classes pour effectuer le partitionnement. L'algorithme *DBSCAN* utilise deux critères de partitionnement : (1) deux points sont agglomérés dans la même classe s'ils sont dans un rayon de ϵ-distance et (2) la classe obtenue est enregistrée s'ils ont un nombre minimal de points (minPts). *DBSCAN* est capable de gérer les données convexes et les données "aberrantes" (en les éliminant du processus de partitionnement). *DBSCAN* présente donc un intérêt, par exemple, dans le traitement des données bruitées où certaines observations ne seront pas regroupées. Cependant, il a des difficultés à gérer les classes de densités différentes.

- La technique de classification spectrale (*SC*) qui est caractérisée par l'utilisation d'un nouvel espace (espace des valeurs propres) pour définir une segmentation. En effet, les méthodes spectrales ont pour point de départ un graphe de similarité pondérée W issue des données et d'un critère de coupure. À partir de W sont alors calculées, une matrice Laplacienne L, puis la décomposition en valeurs propres de celle-ci. L'étape de *clustering* se fait dans cet espace spectral à partir des K-premiers vecteurs propres. Il existe de nombreuses variantes (Section 2.4.2.3, chapitre 2) comme les *NJW-SC* [Ng et al. 2001] qui utilisent une matrice Laplacienne symétrique normalisée ($L_{NJW} = D^{-1/2}WD^{1/2}$; D la matrice de degrés de W) et l'algorithme *K-means* pour le partitionnement. Dans notre cas, nous avons dérivé *NJW-SC* avec le partitionnent *K-means* via un partitionnement *K*-

medoids (*PAM*). Pour simplifier, elles seront appelées respectivement SC-KM et SC-PAM. Elles offrent l'avantage de s'affranchir d'un grand nombre d'hypothèses sur la distribution des données. Ainsi ces méthodes sont souvent utilisées dans le cadre de classes à densité hétérogène ou faible.

Les approches hiérarchiques sélectionnées sont :

- Les techniques classiques de *clustering* hiérarchique (*HC*) [Borcard et al. 2011] (Section 2.4.2.1, chapitre 2) sont basées sur la proximité entre les observations dans l'espace initial. Pour les divisions nous avons choisi la méthode descendante, c'est-à-dire, chaque observation est d'abord affectée à une seule et même classe, puis il est divisé successivement pour obtenir la séparation la plus distincte possible.

- Les approches qui sont dites spectrales et hiérarchiques tels que *H-SC* et *Bi-SC*. *H-SC* est proposée par [Sanchez-Garcia, Fennelly, Norris et al. 2014]. Dans cet algorithme l'étape de partitionnement est basée sur *HC*, mais elle se fait dans l'espace spectral ($L_N JW$). *Bi-SC* [Shi et Malik 2000] conduit à un arbre binaire : à chaque niveau, chaque nœud est subdivisé en K = 2 classes selon le signe du deuxième vecteur propre du Laplacien $L_{Shi} = I - D^{-1/2} W D^{1/2} = I - L_{NJW}$. Cette contrainte de séparation en 2 groupes est bien adaptée quand une structure dominante est présente (comme le fond dans une image).

- *Hierarchical Density-Based Spatial Clustering* (*HDBSCAN*) est une extension hiérarchique de *DBSCAN*, algorithme où une dissimilarité basée sur le voisinage électronique est utilisée pour agréger les observations.

- Les analyses des points de rupture (*change-point*) sont également ajoutées pour le cas des séries temporelles. Nous retenons : *Divisive estimation* (*e.divisive*) et *Agglomerative estimation* (*e.agglo*) qui sont des approches de *clustering*. Ce sont également des approches hiérarchiques basées sur la distance énergétique (e) [James et Matteson 2015]. *e.divisive* définit les segments par une méthode de bissection binaire et un test de permutation. *e.agglo* crée des classes homogènes basées sur un *clustering* initial. Si aucun *clustering* initial n'est défini comme tel, chaque observation est affectée à son propre segment.

3.3.3.2 Traitement des données et paramétrages des méthodes

Dans un premier temps, tous les jeux de données présentés section 3.3.3.2 sont normalisés pour éviter l'impact de la variation d'échelle dans le processus de *clustering*. On obtient ainsi une base de jeux de données homogènes affranchie des différences liées aux unités de mesures et aux gammes de valeurs utilisées dans les données artificielles.

Dans la mesure du possible, toutes les méthodes sont appliquées avec leurs paramètres par défaut. Toutefois certains choix doivent être faits au niveau des méthodes de calcul du regroupement (nombres de voisins, matrice Laplacienne) et au niveau du critère d'arrêt du *clustering*. Ainsi, pour toutes les approches directes, le nombre de classes (K) est fixé suivant le nombre de vrais labels (C) pour chaque jeu de données, soit $K = C$. De plus, les méthodes spectrales directes (*M-SC*, *H-SC*) sont basées sur le Laplacien L_{NJW} et un nombre de points minimaux pour constituer une classe, ici fixé à 7. Pour faire ce choix, nous nous sommes basés sur [Zelnik-Manor et Perona 2004] qui ont adapté localement le noyau gaussien avec la distance du 7^e voisin dans le calcul de la similarité. Pour *M-SC*, une contrainte supplémentaire est ajoutée : nous avons fixé la silhouette minimum de chaque cluster (sil.min) à $0,7$. Pour les

méthodes hiérarchiques (*HC*, *H-SC*, *HDBSCAN*) et les méthodes divisives (*Bi-SC*, *M-SC*) les arbres sont coupés pour obtenir au moins C clusters. Pour *DBSCAN*, la détermination des classes $K = C$ nécessite aussi un voisinage. Dans ce cas, il est automatiquement déterminé par une estimation du UIK (*Unit Invariant Knee*) de la distance k moyenne du plus proche voisin. Enfin, pour la base de données *Species* (le nombre de classes est très important), seuls les résultats des 20 premières composantes principales sont retenus. Elles sont suffisantes pour obtenir une somme cumulée de la variance expliquée supérieure à 95 %.

3.3.3.3 Analyse des performances

Une comparaison avec les algorithmes : de partitionnement direct (*K-means*, *DBSCAN*), direct et spectral (*SC*, SC-KM, SC-PAM), hiérarchique (*HC*, *HDBSCAN*), spectral hiérarchique (*H-SC*, *Bi-SC*) et par point de rupture (*e.divisive*, *e.agglo*) est réalisée sur l'ensemble de jeux de données du tableau 3.5.

Le tableau 3.6 récapitule toutes les méthodes qui isolent correctement au moins 50 % des vrais labels. Elles sont ordonnées de la plus à la moins performante en fonction du nombre de classes bien isolées (#Iso), la précision totale (Tot.acc) et le nombre de classes décrites (K) (qui doit être minimisé pour réduire la tâche de labellisation). Les quatre premières lignes du tableau 3.6 sont les jeux spatiaux et les quatre autres sont les séries temporelles. Le résultats de chaque classification est représenté sous forme de graphique en Annexe (Annexe C)

Pour les jeux de données spatiaux, les méthodes spectrales (directes ou hiérarchiques) sont globalement les plus efficaces. Elles parviennent à identifier toutes les classes (#Iso=C) avec des scores élevés (Tableau 3.6), sauf pour *Species* où il est difficile d'identifier un algorithme efficace. Mais cela peut s'expliquer par la faible répartition des classes et par des classes fortement connectées (silhouette moyenne =0,1). Dans le cas simple, comme le jeu de données *aggregation* ce sont les méthodes spectrales hiérarchiques qui sont les plus performantes. En effet, la majorité des méthodes ont isolé les sept classes, mais les algorithmes *H-SC* et *M-SC* ont la plus grande précision totale (100 %). Il en est de même pour le jeu de données Iris où ce sont encore une fois les algorithmes spectraux hiérarchiques (*M-SC* et *Bi-SC*) qui sont les meilleurs. Lorsque les motifs deviennent plus complexes, comme pour le jeu de données *compound*, seul *M-SC* détecte toutes les classes. En effet, certains motifs du jeu de données *compound* sont imbriqués les uns dans les autres et il est difficile pour la plupart des méthodes de les différencier (Figure 3.6).

Pour la segmentation des séries temporelles, ce sont les méthodes hiérarchiques et par points de rupture qui sortent du lot, en particulier *M-SC*, *HDBSCAN* et *e.divisive*. Dans le cas simulé, trois fortes variations (ev1, ev2, ev3) correspondant à deux évènements extrêmes et un *offset* sont à identifier. Dans la plupart des cas, c'est l'*offset* qui est difficile à séparer du signal global pour les différentes méthodes (Figure 3.7). Seul *M-SC* et *e.divisive* y parviennent. Mais *e.divisive* ne semble pas adaptée à nos problématiques. En effet, même si elle détecte correctement les 3 évènements, elle a tendance à sur-segmenter ($C = 23$) (Figure 3.7). Elle a donc des scores (Tot.acc= $0,97$, Silhouette= $0,03$) plus élevés, mais *M-SC* avec seulement $C = 8$ classes détectées, a une meilleure connexité intra-classe avec des indices de Dunn et ARI égales à $0,007$ et $0,43$, respectivement. Dans le cas des données expérimentales, les conclusions émises pour le jeu de données simulé se rejoignent avec le leg1 et le leg2. Par contre pour le leg3, aucun des algorithmes n'isole les 3 classes. *M-SC* réussit à les isoler au niveau 3 avec $K = 102$ et une précision totale de 93 %. Ce nombre de classes à labelliser peut paraître excessif pour l'expert mais les autres méthodes propose un nombre de classes du même ordre de grandeur sans pour autant isoler les vrais labels.

Les résultats obtenus révèlent une bonne capacité de généralisation en seulement 3 niveaux et démontrent que la classification spectrale à plusieurs niveaux est donc une des méthodes les plus efficaces pour détecter des structures cohérentes aussi bien pour des données spatialisées ou les séries temporelles. En effet *M-SC* est une méthode qui répond bien lorsque la distribution des données est complexe. Contrairement aux méthodes directes comme *K-means* ou *HC* qui nécessitent des formes non convexes et des groupes linéairement séparables pour être optimales. De même, la segmentation par points de rupture (*e.divisive* et *e.agglo*) est performante lorsque les signaux ont une dynamique stationnaire séparée par des ruptures marquées. Ces ruptures ne sont pas si évidentes dans le cas des données marines où les événements extrêmes ne constituent pas une variation intensive avec le signal global en moyenne ou en variance. Enfin, *DBSCAN* et *HDBSCAN* considèrent souvent ces événements extrêmes comme une anomalie (un *outlier*).

Tout comme les autres méthodes, *M-SC* peut sur-segmenter ou ne pas détecter toutes les classes, mais elle offre la possibilité à l'expert de trouver un compromis entre la sur-segmentation et le nombre de classes voulues via le réglage du paramètre de la silhouette minimum (sil.min) (Section 3.2.3.2).

TABLEAU 3.6 – Indices de performance des différents algorithmes de classification non-supervisées pour chaque jeu de données. Chaque méthode est ordonnée en fonction du nombre de motifs bien isolés (#Iso). Puis les indices de performance : Indice Rand ajusté (ARI), indice Dunn et Silhouette (Sil.), précision totale (Tot.acc) et le nombre de clusters K sont décrits. La première ligne de chaque jeu (Labels Vrais) représente le score obtenu avec les vrais labels (soit le score maximum pour chaque indice). C : le nombre total de classes.. (0,00 est un résultat non nul mais $> 1.10^{-2}$ et en gras les algorithmes pour lesquels #Iso=C).

Jeux de données	Clustering	K	ARI	Dunn	Sil.	Tot.acc	#Iso
1 Aggregate	Vrais Labels	C =7	1.00	0,04	0,49	1.00	7
	H-SC	7	0,99	0,04	0,49	**1.00**	**7**
	M-SC	9	0,89	0,03	0,42	**1.00**	**7**
	SC-PAM	7	0,97	0,03	0,50	0,98	**7**
	SC-KM	7	0,96	0,03	0,50	0,98	**7**
	Bi-SC	8	0,88	0,02	0,42	0,96	6
	HC Ward.d2	7	0,80	0,04	0,45	0,95	6
	KM	7	0,73	0,04	0,49	0,90	5
	DBSCAN	5	0,81	0,11	0,41	0,83	5
	HDBSCAN	5	0,81	0,11	0,41	0,83	5
2 Compound	Vrais Labels	C =6	1.00	0,07	0,16	1.00	6
	M-SC	6	1,00	0,07	0,16	**1,00**	**6**
	KM	6	0,56	0,02	0,35	0,85	5
	Bi-SC	8	0,62	0,03	0,26	0,81	4
	SC-KM	6	0,45	0,03	0,29	0,74	4
3 Iris	Vrais Labels	C =3	1,00	0,06	0,50	1,00	3
	M-SC	3	1,00	0,06	0,50	**1,00**	**3**
	Bi-SC	8	0,72	0,06	0,27	**1,00**	**3**
	KM	3	0,62	0,04	0,51	0,83	**3**
	HC Ward.d2	3	0,61	0,07	0,50	0,83	**3**
	H-SC	3	0,45	0,05	0,53	0,67	2
	SC-KM	3	0,45	0,03	0,53	0,67	2
	SC-PAM	3	0,45	0,03	0,53	0,67	2
4 Species	Vrais Labels	C =100	1,00	0,11	0,10	1,00	100
	SC-KM	100	0,48	0,14	0,07	0,65	75
	SC-PAM	100	0,46	0,09	0,07	0,64	73
	KM	100	0,45	0,12	0,09	0,63	72
	H-SC	100	0,46	0,14	0,08	0,64	71
	HC Ward.d2	100	0,45	0,16	0,11	0,64	70
	M-SC	115	0,31	0,16	0,01	0,50	55
5 Simulated	Vrais Labels	$C = 4$	1,00	0,01	0,16	1,00	4
	e.divisive	23	0,39	0,00	0,03	0,97	**4**
	M-SC	8	0,43	0,007	0,28	0,94	**4**
	HDBSCAN	5	0,62	0,01	0,13	0,94	3
	e.agglo	9	0,45	0,00	-0,03	0,94	3
6 DYPHYMA leg1	Vrais Labels	$C = 3$	1,00	0,00	-0,02	1,00	3
	M-SC	32	0,66	0,00	-0,21	0,94	**3**
	HDBSCAN	42	0,68	0,00	-0,10	0,91	**3**
	e.divisive	42	0,49	0,00	-0,22	0,96	**3**
	e.agglo	10	0,48	0,00	-0,14	0,79	**3**
	KM	3	0,57	0,00	-0,04	0,84	2
	HC Ward.d2	3	0,57	0,00	-0,03	0,84	2
	SC-KM	3	0,53	0,00	-0,01	0,84	2
	SC-PAM	3	0,53	0,00	-0,01	0,84	2
	H-SC	3	0,53	0,00	-0,01	0,84	2
7 DYPHYMA leg2	Vrais Labels	$C = 3$	1,00	0,00	-0,03	1,00	3
	M-SC	53	0,51	0,00	-0,18	0,94	**3**
	e.divisive	55	0,55	0,00	-0,17	0,92	**3**
	HDBSCAN	62	0,48	0,00	-0,25	0,89	**3**
	HC Ward.d2	3	0,21	0,00	0,26	0,72	2
	KM	3	0,21	0,00	0,25	0,72	2
	e.agglo	3	0,20	0,00	0,28	0,72	2
	SC-PAM	3	0,11	0,01	0,12	0,64	2
	SC-KM	3	0,06	0,01	0,20	0,64	2
8 DYPHYMA leg3	Vrais Labels	$C = 3$	1,00	0,00	-0,02	1,00	3
	HC ward.d2	3	0,23	0,00	-0,00	0,80	2
	KM	3	0,21	0,00	0,01	0,79	2
	M-SC	4	0,41	0,00	-0,15	0,79	2

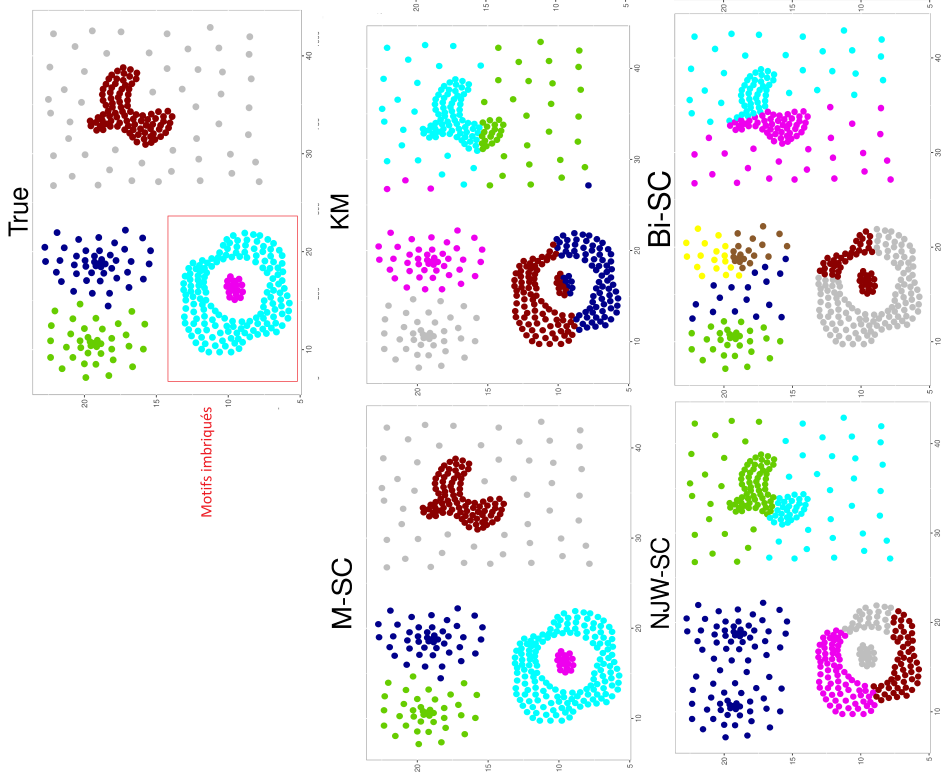


FIGURE 3.6 – Résultats des classifications sur le jeu test de données spatiales : "Compound". Le jeu de données "Compound" est un jeu labellisée : les couleurs de la figure True représentent les vrais labels C. Pour les autres graphiques, la couleur correspond aux classes C définie par chacune des méthodes de classifications les plus efficaces (ici : M-SC, H-SC, Bi-SC, SC-KM, KM (*K-means*), HC).

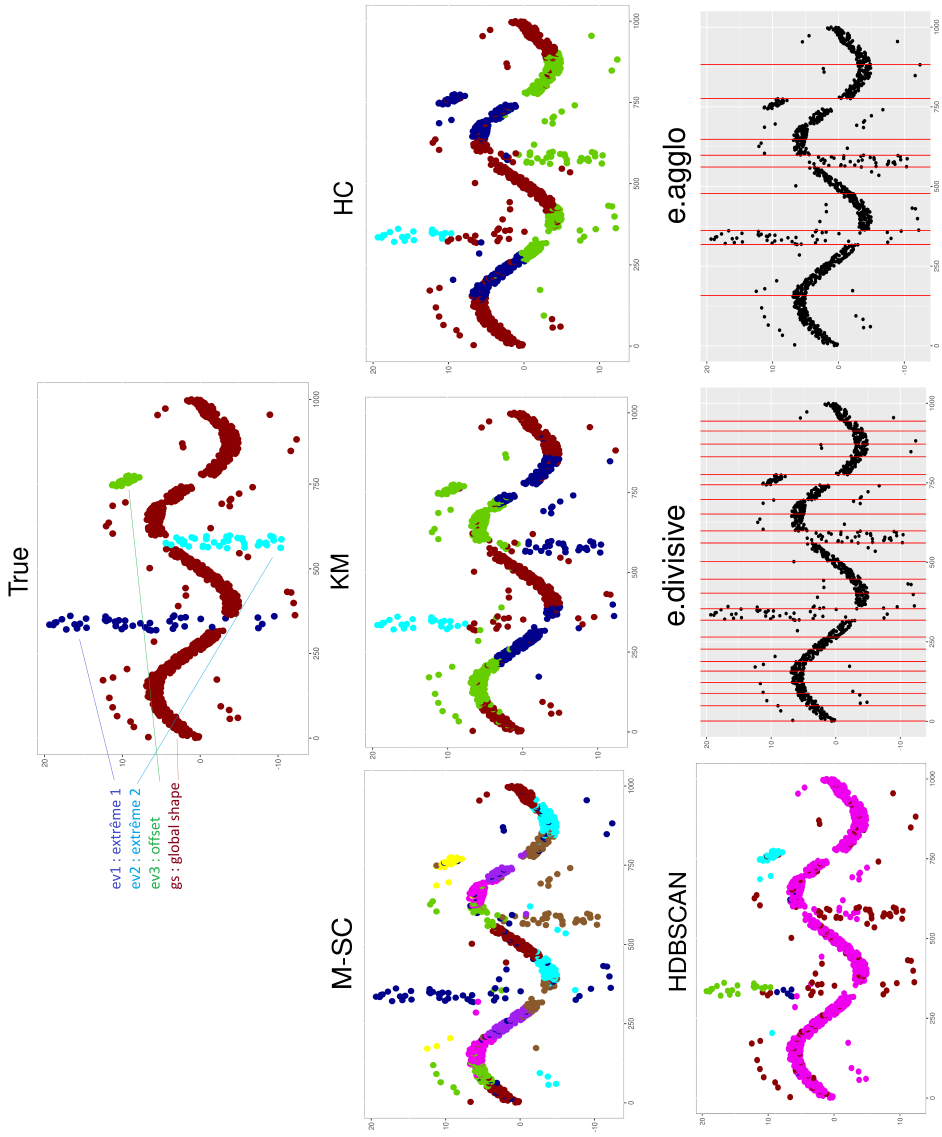


FIGURE 3.7 – Résultats des classifications sur le jeu test de données temporelles : "Simulated". Le jeu de données "Simulated" est un jeu labellisée : les couleurs de la figure True représentent les vrais labels C. Pour les autres graphiques, la couleur correspond aux classes C définie par chacune des méthodes de classifications les plus efficaces (ici : M-SC, H-SC, Bi-SC, SC-KM, KM (*K-means*), HC).

3.3.4 Validation de la méthode : Évaluation des capacités de labellisation

Les méthodes de classification reposent uniquement sur la géométrie des données pour assurer la segmentation. Au contraire, les méthodes de classification supervisées s'appuient sur les informations déjà existantes, les vrais labels notés de 1 à C. L'objectif de cette dernière partie est de tester leurs capacités à fournir une première classification et de le confronter à nos résultats obtenus avec *M-SC*. Nous comparons donc maintenant des techniques supervisées sur le cas qui est plus en lien avec nos problématiques de recherche : la détection de classes dans les séries temporelles.

3.3.4.1 Sélection des méthodes

Pour répondre à cette question, trois propositions de classifications supervisées sont explorées (détails des méthodes section de chapitre 2) :

- La classification par plus proches voisins, k-plus proches voisins (*k-ppv*), en anglais *k-Nearest Neighbors* (*k-Nearest Neighbors* (*k-nn*)), qui est fondée sur les rapports distances entre les points (*e.g.* Euclidienne. Il calcule la distance entre des données et des prototypes (données labellisées) pour attribuer à l'observation inconnue la classe de son plus proche voisin.

- L'algorithme *RF* qui est construit à partir d'arbres de décision. Ces arbres réalisent la classification d'un objet via une suite de tests sur les attributs qui le décrivent. Le *RF* est la construction d'une "forêt d'arbres" (donc, plusieurs arbres de décision), de manière aléatoire. Cette méthode est qualifiée de méthode d'ensemble : plusieurs modèles «faibles» sont combinés pour créer un modèle robuste.

- La classification par un réseau de neurones qui est basée sur un ensemble de neurones interconnectés par couche. À chaque neurone est associé un poids qui représente l'influence d'un neurone sur un autre. L'apprentissage des poids résulte d'un entraînement du réseau à effectuer des tâches de classification ou de régression à partir d'une base labellisée (données, décision). Le principe est basé sur la propagation d'information dans des unités de calculs élémentaires et de rétropropagation de l'erreur de décision pour corriger ses poids. Le perceptron linéaire est un réseau particulier de neurones avec une couche d'entrée et une couche de neurones de sortie. Ainsi, le classifieur est linéaire c'est-à-dire qu'il utilise des séparatrices linéaires (droites, hyperplans) pour sa classification. Lors que l'on multiplie les couches on obtient un perceptron multi-couche (en anglais *MultiLayer Perceptron* (*MLP*)). Cela permet d'obtenir une classification non linéaire.

3.3.4.2 Traitement des données et paramétrage des méthodes

Comme pour la comparaison entre les méthodes non supervisées (section 3.3.3), toutes les bases de données sont normalisées. Ensuite toutes les séries temporelles (de 5 à 8 dans le tableau 3.5) sont subdivisées en deux bases : une base d'apprentissage et une base test. La base d'apprentissage permet la construction du modèle de classification et la base test permet de vérifier les capacités du modèle. Deux configurations ont été testées. Dans le premier cas de figure, la base d'apprentissage représente 20 % du volume de chaque classe et donc la base test couvre 80 % du reste. Dans le second cas, les deux bases sont à 50 %. Quelle que soit la base, elle couvre tous les événements temporels $K = C$. Pour les quatre jeux de données, le nombre d'entrées I correspond aux nombres de variables (attributs) des jeux de données. Il est égal à 18 pour les campagnes DYPHYMA et à 3 pour le jeu *Simulated*. Le nombre de sorties pour *MLP* et *RF* est fixé en fonction de C. Pour *k-nn*, deux valeurs k sont choisies. $k = 1$ pour attribuer le label de l'observation à son plus proche voisin, puis $k = 7$ pour obtenir une segmentation plus unifiée (c'est-à-dire avec moins d'une classe unique (singleton) parmi les observations des autres classes). Le nombre de couches cachées de *MLP* est fixé à 0 et à 1. MLP-0 correspond au perceptron linéaire (aucune couche cachée) et MLP-1 a donc une couche cachée dont le nombre de neurones est fixé à $(I + C)/2$. Random Forest correspond à un vote sur 500 arbres avec des variables $\left\lfloor \sqrt{I} \right\rfloor$ échantillonnés aléatoirement pour chaque division.

3.3.4.3 Analyse des performances

Afin de comparer les approches de classification, les mêmes indices que dans le tableau 3.6 sont calculés. Chaque indice est calculé sur les bases de données tests et reporté dans le tableau 3.7.

Premièrement, *RF* et *k-nn* sont capables, dans l'ensemble, d'isoler avec une plus grande précision les évènements. Ils les distinguent, pour tous les jeux de données avec un faible chevauchement (comme l'indique les ARI élevés, tableau 3.7). MLP-0 et MLP-1 sont capables d'identifier les 4 événements du cas simulé, mais ils ne sont pas efficaces pour les cas *in-situ* (n=6-8) en raison d'un manque d'observations et de classes inégales. Les techniques de regroupement des méthodes divisives comme *M-SC* ne souffrent pas des problèmes liés aux tailles de classes inégales et permettent de détecter plus facilement les événements de la série. De plus, *M-SC* a atteint le même objectif de nombre de motifs bien isolés que les techniques supervisées comme *RF* ou *k-nn*. Même si les scores des méthodes supervisées sont plus élevés que les scores de *M-SC*, ils sont à nuancer. En effet, les scores de connexité tels que l'ARI dépendent fortement du nombre de classes. Dans le cas supervisé, avec un nombre K fixe et calculé à partir de la base de données des tests uniquement, les scores ARI sont plus élevés que ceux des approches non supervisées.

Deuxièmement, les techniques d'apprentissage supervisé ne sur-segmentent pas les séries temporelles, contrairement aux méthodes de classification non-supervisées (Section 2.4.2). Encore une fois, cela est dû aux conditions initiales de la méthode qui imposent un nombre de classes fixe et connu. Mais il est important de rappeler que dans notre cas d'étude nous connaissons rarement le nombre de classes par avance.

Ainsi, la méthode *M-SC* est tout aussi performante qu'une méthode d'apprentissage pour labelliser une base de données, alors que c'est une méthode non supervisée. De plus, elle est plus facile à mettre en place car elle nécessite très peu d'*a priori* sur les données contrairement aux méthodes supervisées. Elle est donc tout à fait adaptée à nos problématiques d'études.

Tableau 3.7 – Indices de performance des différents algorithmes de classification supervisée pour chaque série temporelle avec un jeu d'entraînement à 20 ou 50 % du nombre de données totales. Chaque méthode est ordonnée en fonction du nombre de motifs bien isolés (#Iso). Puis les indices de performance : Indice Rand ajusté (ARI), indices Dunn et Silhouette (Sil.), précision totale (Tot.acc) et le nombre de clusters K sont décrits. La première ligne de chaque jeu (Labels Vrais) représente le score obtenu avec les vrais labels (soit le score maximum pour chaque indice). C : le nombre total de classes. (En gras : #Iso : nombre de classes bien isolées et $0,00$ résultat non nul mais $> 1.10^{-2}$). RF= Random Forest, MLP-l = Perceptron multi-couche avec l couche cachée, k-nn=k-plus proches voisins.

Jeux de données	20%-entraînement	K=C	ARI	Dunn	Sil.	Tot.acc	#Iso
5 Simulated	ground-truth	4	1,00	0,03	0,16	1,00	4
	RF	4	1,00	0,03	0,17	**1,00**	**4**
	1-nn	4	0,85	0,02	0,17	0,97	**4**
	MLP-0	4	0,90	0,02	0,17	0,97	**4**
	MLP-1	4	0,91	0,05	0,18	0,97	**4**
	7-nn	4	0,65	0,05	0,21	0,94	2
6 DYPHYMA leg1	ground-truth	3	1,00	0,00	-0,01	1,00	3
	RF	3	0,98	0,00	-0,01	0,99	**3**
	1-nn	3	0,77	0,00	0,002	0,91	**3**
	7-nn	3	0,56	0,001	0,02	0,82	**3**
	MLP-0	3	0,21	0,00	0,20	0,59	**3**
	MLP-1	3	-	-	-	0,50	1
7 DYPHYMA leg2	ground-truth	3	1,00	0,00	-0,04	1,00	3
	RF	3	0,98	0,00	-0,03	0,99	**3**
	1-nn	3	0,80	0,00	-0,02	0,91	**3**
	7-nn	3	0,75	0,00	-0,01	0,88	**3**
	MLP-0	3	0,58	0,002	0,24	0,74	2
	MLP-1	3	-	-	-	0,63	1
8 DYPHYMA leg3	ground-truth	3	1,00	0,00	-0,02	1,00	3
	RF	3	0,96	0,00	-0,01	0,98	**3**
	1-nn	3	0,70	0,00	-0,01	0,89	**3**
	7-nn	3	0,59	0,00	0,01	0,86	**3**
	MLP-0	3	0,28	0,001	-0,02	0,78	1
	MLP-1	3	-	-	-	0,70	1

Jeux de données	50%-entraînement	K	ARI	Dunn	Sil,	Tot,acc	#Iso
5 Simulated	ground-truth	4	1,00	0,03	-0,02	1,00	4
	RF	4	1,00	0,03	0,16	**1,00**	**4**
	1-nn	4	0,93	0,03	0,14	0,99	**4**
	7-nn	4	0,87	0,07	0,16	0,98	**4**
	MLP-0	4	0,99	0,03	0,15	0,99	**4**
	MLP-1	4	0,95	0,04	0,24	0,96	3
6 DYPHYMA leg1	ground-truth	3	1,00	0,00	-0,05	1,00	3
	RF	3	1,00	0,00	-0,02	**1,00**	**3**
	1-nn	3	0,83	0,00	-0,02	0,93	**3**
	7-nn	3	0,73	0,00	-0,02	0,90	**3**
	MLP-0	3	0,65	0,007	0,03	0,83	2
	MLP-1	3	-	-	-	0,51	1
7 DYPHYMA leg2	ground-truth	3	1,00	0,00	-0,015	1,00	3
	RF	3	0,97	0,00	-0,05	0,99	**3**
	1-nn	3	0,84	0,00	-0,05	0,92	**3**
	7-nn	3	0,80	0,00	-0,07	0,91	**3**
	MLP-0	3	0,75	0,001	0,14	0,79	2
	MLP-1	3	-	-	-	0,63	1
8 DYPHYMA leg3	ground-truth	3	1,00	0,00	0,16	1,00	3
	RF	3	0,98	0,00	-0,01	0,99	**3**
	1-nn	3	0,81	0,00	-0,01	0,93	**3**
	7-nn	3	0,70	0,00	0,003	0,90	**3**
	MLP-0	3	0,35	0,00	0,04	0,78	2
	MLP-1	3	-	-	-	0,69	1

3.4 Validation de la méthode sur un cas pratique : Données MAREL-Carnot

Dans la section précédente (section 3.3), la méthode *M-SC* est testée et validée sur plusieurs jeux de données référencées (artificielles et expérimentales). Nous allons donc maintenant mettre en application la méthode sur un premier cas pratique : le site de Manche orientale (MAREL-Carnot). Cette nouvelle phase a pour objectif d'évaluer l'efficacité de la méthode a définir des états environnementaux cohérents sur un jeu de données aux caractéristiques complexes de type MAREL (*e.g.* bases HF, volumineuses, non linéaires, avec un grand nombre de données manquantes (*NA*) ...). Ainsi, dans cette partie sont introduits dans un premier temps la base de données, puis les paramétrages de la méthode de classification. Ensuite les résultats de la classification sont présentés dans leur ensemble. Puis pour finir une labellisation est proposée.

3.4.1 Sélection de la base de données et paramétrage

3.4.1.1 Description du jeu de données : MAREL-Carnot

La station de mesures MAREL-Carnot est en service depuis 2004 et, est complètement opérationnelle depuis 2005. Les données sont donc disponibles pour les années 2005 à nos jours (Section 2.1.1). Pour cette application, nous travaillons sur la période de 2005 à 2009 (inclus). Ces travaux s'inscrivant dans la continuité de la thèse de Rousseeuw et al. 2015a, nous avons choisi une période identique à des fins de comparaison.

Avec une mesure toutes les 20 minutes, la base de données (2005-2009) comprend 131 472 instants d'acquisition (N) pour toutes les variables, excepté pour les concentrations en nutriments, où nous disposons de N =7 305 instants (fréquence d'échantillonnage biquotidienne). Sur cette période, 18 variables sont mesurées. Dans notre cas, nous avons sélectionné 9 variables non corrélé qui nous assure que le poids de chaque variable soit le plus équilibré et réparti possible. Ainsi, nous nous assurons que chaque variable aura son influence propre lors de l'analyse multivariée. Ces variables seront appelées les variables contributives du modèle (Tableau 3.8).

Dans leur étude, Rousseeuw et al. 2015a ont utilisé une ACP pour identifier les variables corrélées et ont sélectionné 10 variables non corrélées (Figure 3.8), de ces 10 variables nous avons, retiré la vitesse du vent en rafale et le niveau de la mer.

En effet, après plusieurs tests de classification, le niveau de la mer s'est avéré trop structurant. Son signal très périodique prend le pas sur la dynamique plus sporadique de certaines autres variables comme les sels nutritifs. Le clustering se retrouvait ainsi biaisé. La variable de vent par rafale correspond aux augmentations soudaines du vent, dépassant la vitesse symbolique de 10 nœuds, soit 18 km/h. Elle a donc été supprimée, car jugée peu représentative et donc non significative. Pour intégrer le vent de manière cohérente, nous aurions pu ajouter les données de vitesse et de direction du vent disponible à la station météorologique la plus proche. Mais cette intégration pose deux problèmes majeurs : (i) la station météorologique de Boulogne-sur-mer est assez éloignée de la station MAREL-Carnot, ce qui peut engendrer un biais important pour une étude aussi locale que la nôtre ; (ii) la méthode *M-SC* est basée sur des calculs de distance dans l'espace des variables et dans l'espace des vecteurs propres (Algorithme 9). Dans cette configuration les directions du vent 0° et 360° se retrouvent à une distance opposée alors qu'elles devraient être identiques. Une des solutions est de transformer les coordonnées polaires (distance, vitesse) en coordonnées cartésiennes (vecteurs vitesses). Mais cette solution engendrant une corrélation forte entre les deux variables, nous n'avons pas souhaité les intégrer dans la classification. Elles sont toutefois utilisées comme variables complémentaires.

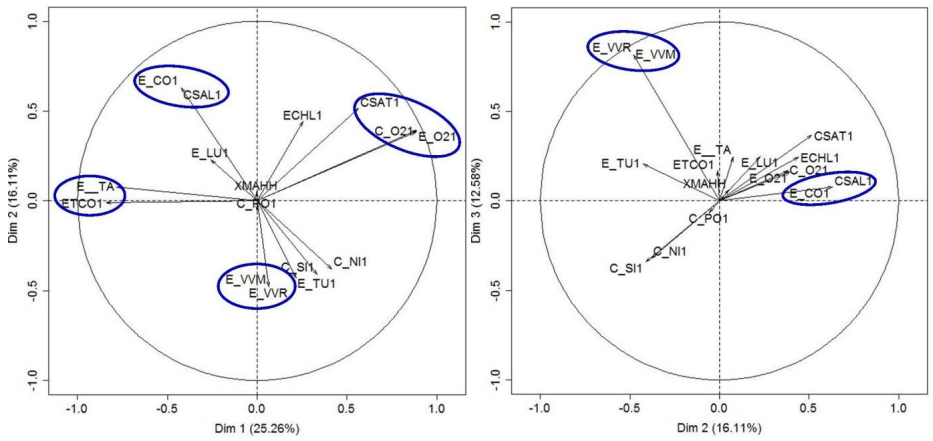
Ensuite, alors que Rousseeuw et al. 2015a utilisent le signal de fluorescence comme variable

Tableau 3.8 – Liste des variables mesurés par la station MAREL-Carnot avec leurs noms et acronymes associés. En vert les variables contributives sélectionnées pour le développement de la méthode *M-SC*. #Na % : pourcentage de données manquantes pour les variables contributives.

Noms					
Variables	**Coriolis**	**Acronyme**	**Unités**	Sélection	#Na%
Oxygène dissous corrigé	DOX1.LEVEL1	C_O21	$\mathrm{ml\,l^{-1}}$		2,32
Oxygène dissous non corrigé	MASS_DOXY.LEVEL1	E_O21	$\mathrm{ml\,l^{-1}}$		
Saturation en Oxygène	OSAT.LEVEL1	CSAT1	%		
Fluorescence	FLU3.LEVEL1	ECHL1	FFU		0,28
pH	PH.LEVEL1	E_PH1	//		
Salinité	PSAL.LEVEL1	CSAL1	PSU		2,67
Conductivité	CNDC.LEVEL1	E_CO1	Sm^{-1}		
Température de l'eau	TEMP.LEVEL1	E_TA	°C		0,00
Température de l'air	DRYT.LEVEL0	ETCO1	°C		
Niveau de la mer	SLEV.LEVEL1	XMAHH	m		
Vitesse du vent en moyenne	WSPD.LEVEL0	E_VVM	$\mathrm{m\,s^{-1}}$		
Vitesse du vent en rafale	WSPD.LEVEL0	E_VVR	$\mathrm{m\,s^{-1}}$		
Direction du vent	WDIR.LEVEL0	E_VDM	°		
PAR (surface)	LGH4.LEVEL1	E_LU1	$\mathrm{\mu mol\,m^{-2}\,s^{-1}}$		0,00
Turbidité	TUR4.LEVEL1	E_TU1	NTU		4,05
Nitrate	NTRZ.LEVEL1	C_NI1	$\mathrm{\mu mol\,l^{-1}}$		10,04
Phosphate	PHOS.LEVEL1	C_PO1	$\mathrm{\mu mol\,l^{-1}}$		14,07
Silicate	SLCA.LEVEL1	C_SI1	$\mathrm{\mu mol\,l^{-1}}$		15,17

de validation des schémas de variation observés, nous avons fait le choix de l'inclure dans la classification comme une information directe, indispensable afin de caractériser le processus de notre étude (la dynamique phytoplanctonique) et donc en prenant en compte à la fois les forçages, les différents facteurs de contrôle et les réponses. D'autres sources de données complémentaires comme les abondances taxonomiques, issues du REPHY, seront utilisées comme indicateurs.

Nous travaillons donc sur une base de données de dimension 131 472 ×9. Une fois alignée et complétée, la base de données compte au total ±48% de *NA*. Or la fonction *M-SC* supprime toutes les lignes qui présentent au moins un *NA*. Avec un pourcentage de données manquantes pour ces cinq années qui varie de 0 % à un peu plus de 15 % par variable et avec un nombre total de lignes contentant au moins un NA égale à 28 883, la base finale qui sera la référence pour les calculs ultérieurs est de dimension N égale à 102 859 après suppression des lignes concernées.



(a) ACP dimensions 1 et 2 (b) ACP dimensions 2 et 3

FIGURE 3.8 – Analyse en composantes principales des données MAREL-Carnot sur la période 2005-2008 issues de ROUSSEEUW et al. 2015a , a) sur les dimensions 1 et 2 et b) sur les dimensions 2 et 3, les cercles bleus rassemblent les paramètres corrélés entre eux.

3.4.1.2 Calibration de la méthode

Dans la section 3.2.3.2, plusieurs nouveaux paramètres (NivMax, sil.min, crit) sont intégrés au protocole de classification. Ces paramètres sont un point essentiel pour utiliser de façon optimale la fonction *M-SC*. Il est donc important de bien les définir. Pour cette analyse, la fonction *M-SC* est calibrée de la manière suivante :

- Le nombre maximal de niveaux de classification (NivMax) est choisi égal à 4. Pour rappel, plus le niveau est élevé, plus le niveau de détails augmente. L'analyse des résultats est stoppée au 3ème niveau car le nombre d'états multivariés est suffisamment important pour cette première analyse écologique. Ainsi seuls les trois premiers niveaux sont interprétés à ce stade afin de juger de la pertinence de la classification au regard des connaissances acquises. Une fois validé, le nombre de niveaux considéré pourra être plus important afin de permettre la mise en évidence des événements plus rares, intermittents, extrêmes.
- Le critère d'arrêt (sil.min) (Section 3.2.3.2) a été défini pour stopper la segmentation lorsqu'une classe est bien structurée. Ici, la classification s'arrêtera lorsque la silhouette d'une classe sera supérieure ou égale à 0,7. Cette valeur est modératrice ; assez faible pour limiter la sur-segmentation mais assez forte pour détecter des évènements de courte durée. Ainsi, elle est définie comme valeur par défaut dans le nouveau protocole.
- Le critère de coupe (crit) (Section 3.2.3.1) définit la méthode de détermination du nombre de K-classes. Dans ce cas, c'est une méthode d'analyse des amplitudes des valeurs propres qui est choisie : la technique des *PEV* ; C'est une méthode plus restrictive que la méthode initialement utilisée (Technique du "Gap"). Elle va limiter davantage le nombre de classes dans les couches profondes.

3.4.2 Phase de la classification : Des schémas généraux vers des schémas spécifiques

La fonction *M-SC*, de part son algorithme profond permet une identification des schémas de fonctionnement à plusieurs niveaux d'interprétations et d'échelles temporelles. Ainsi, sur le jeu de

données MAREL-Carnot, l'algorithme détecte deux classes au niveau 1 (Figure 3.9a, 3.9b, 3.9c), quatre classes au niveau 2 (Figure 3.9d, 3.9e, 3.9f), huit au niveau 3 (Figure 3.9g, 3.9h, 3.9i) et 15 classes au niveau 4 (Figure 3.9j, 3.9k, 3.9l). À chaque niveau le nombre de classes augmente, les échelles de temps diminuent et donc le niveau de détails augmente aussi. Ainsi cette approche offre la possibilité de choisir le niveau de détails adapté à l'étude.

Ici, nous faisons le choix de présenter les résultats par niveau avec une logique de succession temporelle afin d'être conforme à la dynamique environnementale du phytoplancton.

Le niveau 1 sépare les données en deux classes et met en évidence deux états avec des périodes bien distinctes. Ce sont des états qui couvrent des périodes relativement longues de l'ordre de plusieurs mois. D'après l'analyse de corrélation (méthode de Pearson, section 2.3.3.1) (Tableau 3.9), ce sont les variables de température et d'oxygène qui structurent en grande partie la classification (coefficients de corrélation de $0,60$ et de $-0,65$ pour la classe 1, et de $-0,38$ et de $0,59$ pour la classe 2, respectivement).Ce sont des variables marquées par une forte variabilité saisonnière. Cette variabilité se retrouve au niveau de la répartition des classes. En effet, **la classe 1** (cl1, rouge) concerne les mois d'avril à octobre tandis que la classe 2 (cl2, vert) concerne la période de novembre à mars avec un léger chevauchement entre les deux états (Figure 3.9b). La classe cl1 est caractérisée par des températures élevées, de faibles concentrations en oxygène et en nutriments. La classe cl2 se différencie de cl1 par des températures plus faibles, une augmentation des concentrations en oxygène et en nutriments. Ces augmentations sont caractéristiques des périodes de régénérations hivernales où le brassage et les échanges océan-atmosphère sont plus importants. À ce stade, c'est donc une dynamique saisonnière en termes de cycle biogéochimique qui est mis en évidence et non une dynamique claire liée à la production phytoplanctonique. Un deuxième niveau de classification est nécessaire pour identifier des structures caractéristiques.

Tableau 3.9 – Coefficients de corrélations entre les paramètres contributifs et les classes déterminées au 1er niveau de classification de M-SC sur la période 2005-2009 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 et aux variables structurantes des classes). L'astérisque (*) indique les corrélations non significatives (p-value $> 0,01$).

	cl1	cl2
Salinity (PSU)	0,29	-0,25
Temp (°C)	**0,60**	**-0,65**
OxyDissolved ($mg\,l^{-1}$)	-0,38	**0,59**
Turbidity (NTU)	-0,21	0,17
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	0,26	-0,18
N ($\mu mol\,l^{-1}$)	-0,38	0,32
P ($\mu mol\,l^{-1}$)	-0,18	0,05
Si($\mu mol\,l^{-1}$)	-0,41	0,40
Fluo ($\mu g\,l^{-1}$)	-0,09	0,24

Le niveau 2 subdivise chacune des deux classes du niveau 1 en sous-classes, réparties en sous-périodes. Ainsi, le passage à un niveau plus profond améliore l'identification des principaux états et de leurs périodes de transition. À ce stade, les périodes sont plus courtes. Les états sont d'une durée de l'ordre du mois à la quinzaine de jours (*e.g.* cl2 et cl4)) (Figure 3.9f). De plus, de nouvelles variables structurantes (en plus de la température et l'oxygène) vont déterminer la classification, notamment les concentrations en sels nutritifs. **La classe 3** (cl3, bleue) est détectée de novembre à février et semble assez continue dans le temps (Figure 3.9e). Elle est structurée par les faibles températures (coefficient $= -0,47$) et de forts concentrations de nutriments (coefficient de nitrate $= 0,49$ et de silicate $= 0,54$, Tableau 3.10). Cette classe correspond à la période

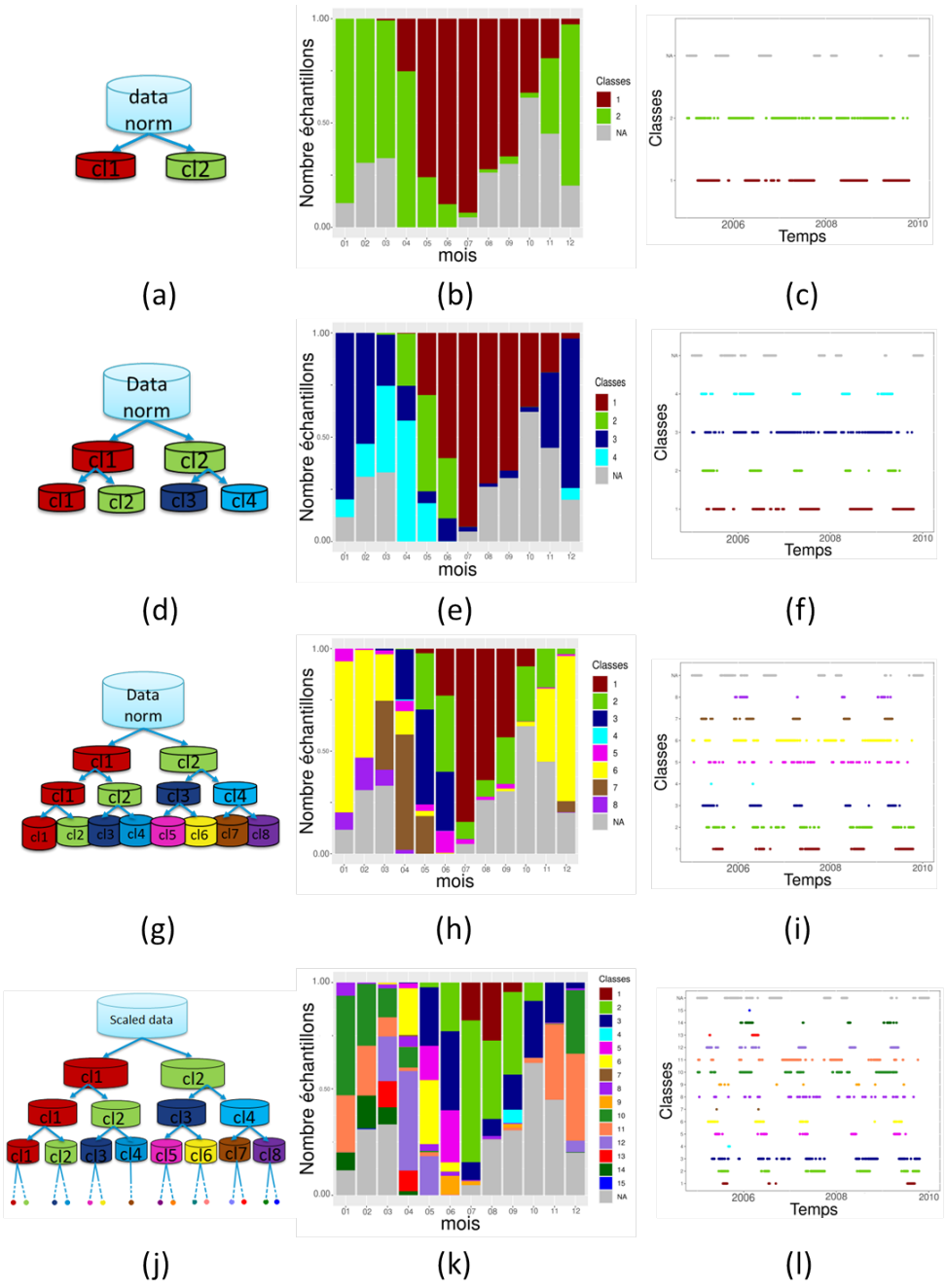


FIGURE 3.9 – Application de la méthode *M-SC* à 4 niveaux sur le jeu de données MAREL-Carnot. Schémas de la classification par niveau (figures a, d, g, j). Fréquence d'occurrence de chaque état par mois et par niveau (figures b, e, h, k) et dynamique de chaque état par niveau (figures c, f, i, l). La couleur grise représente les données non classées (au moins une donnée manquante (#NA)).

hivernale avec des concentrations en nutriments maximales (Tableau 3.10). **La classe 2** (cl2, vert) est détectée d'avril à juin et correspond à un état d'efflorescence printanière. Elle est caractérisée par une augmentation des concentrations en silice, mais il reste difficile de distinguer une variable réellement structurante pour cette classe (Tableau 3.10). **La classe 4** (cl4, cyan) correspond également à une période productive. Elle est détectée en période printanière (mars à mai) avec de légères occurrences en fin de période hivernale (Février) (Figure 3.9f). Elle est structurée par une forte concentration en biomasse phytoplanctonique (Tableau 3.10). Elle correspondrait à un déclenchement de bloom précoce en fin d'hiver qui perdurera au début du printemps. **La classe 1** (cl1, rouge) se produit principalement de juin à septembre (Figure 3.9f). Elle est caractérisée principalement par les fortes températures (coefficient de corrélation $= 0,64$, Tableau 3.10) et par des concentrations de nutriments faibles avec toutefois quelques maximums (*e.g.* coefficient de corrélation du nitrate $= -0,26$, Tableau 3.10). L'efflorescence printanière conduit à une diminution importante de la disponibilité en nutriments ce qui limite le développement en période estivale. Les concentrations en nutriments sont relativement faibles à cette période mais des apports par régénération et/ou par des apports fluviaux peuvent favoriser une efflorescence secondaire.

Tableau 3.10 – Coefficients de corrélations entre les paramètres contributifs et les classes déterminées au 2e niveau de classification de M-SC sur la période 2005-2009 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 et aux variables structurantes des classes). L'astérisque (*) indique les corrélations non significatives (p-value $> 0,01$).

	cl1	cl2	cl3	cl4
Salinity (PSU)	0,19	0,18	**-0,38**	0,14
Temp (°C)	**0,64**	-0,01	**-0,47**	**-0,33**
OxyDissolved ($mg\,l^{-1}$)	**-0,43**	0,06	**0,25**	**0,54**
Turbidity (NTU)	-0,12	-0,17	**0,28**	-0,12
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	0,17	0,18	-0,20	0,00*
N ($\mu mol\,l^{-1}$)	**-0,26**	-0,23	**0,49**	-0,19
P ($\mu mol\,l^{-1}$)	-0,12	-0,10	0,17	-0,15
Si($\mu mol\,l^{-1}$)	**-0,25**	**-0,29**	**0,54**	-0,14
Fluo ($\mu g\,l^{-1}$)	-0,16	0,11	-0,14	**0,53**

Ce deuxième niveau de classification met en évidence des schémas cohérents avec la dynamique phytoplanctonique. Il constitue la première étape de validation de l'algorithme *M-SC* capable d'identifier les périodes clefs de la dynamique du phytoplancton : les efflorescences de printemps (mars-juin ; cl2, cl4), les faibles abondances (octobre-mars ; cl3), les situations intermédiaires (juin à septembre : cl1). Toutefois, ce niveau ne permet pas l'identification d'évènements particuliers (aux caractères rares ou extrêmes). Pour aller plus loin dans l'identification d'états sous-jacents, la classification est donc réalisée à des niveaux encore plus profonds : les niveaux 3 et 4. Ces niveaux permettent d'aller au-delà de l'identification de modèles généraux fortement structurés par des facteurs environnementaux tels que la température et/ou la disponibilité des nutriments (Tableau 3.11). À partir de ces niveaux, des événements intermittents, rares ou extrêmes, dont la période peut être infra-hebdomadaire ou même horaire, peuvent être identifiés.

Au niveau 3, 8 classes sont identifiées. Certaines classes (*e.g.* cl2, vert) restent structurées par les mêmes variables que pour les niveaux précédents (température, l'oxygène) alors que pour d'autres classes (*e.g.* cl4, cyan, cl5, rose, cl8, violet), d'autres variables apparaissent comme majoritairement structurantes telles que la salinité, la fluorescence ou les sels nutritifs (Tableau 3.11). À ce stade des sous-états ponctuellement et sur de courtes durées sont identifiés (Figure 3.9i). Il faut aussi noter les classes cl3 (bleu) et cl7 (marron) qui sont identifiées exclusivement lors de la période productive. Toutefois, il y a toujours des classes dont les étendues des périodes

d'occurrences sont assez importantes telles que cl6 (jaune) (avec une période maximum égale à 58 jours).

TABLEAU 3.11 – Coefficients de corrélations entre les paramètres contributifs et les classes déterminées au 3e niveau de classification de M-SC sur la période 2005-2008 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 et aux variables structurantes des classes). L'astérisque (*) indique les corrélations non significatives (p-value $> 0,01$).

	cl1	cl2	cl3	cl4	cl5	cl6	cl7	cl8
Salinity (PSU)	0,10	**0,15**	**0,18**	0,02	0,01	**-0,40**	**0,15**	0,00*
Temp (°C)	**0,58**	**0,21**	-0,01	-0,01	0,02	**-0,49**	**-0,24**	**-0,25**
OxyDissolved ($mg\,l^{-1}$)	**-0,29**	**-0,26**	0,06	-0,00*	0,01	**0,25**	**0,42**	**0,32**
Turbidity (NTU)	-0,05	-0,10	**-0,1**	-0,02	0,02	**0,28**	-0,13	-0,01*
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	**0,15**	0,06	**0,18**	-0,00*	0,02	**-0,21**	0,02	-0,03
N ($\mu mol\,l^{-1}$)	**-0,35**	0,05	**-0,23**	-0,03	0,07	**0,48**	**-0,22**	0,02
P ($\mu mol\,l^{-1}$)	**-0,17**	0,03	-0,11	0,03	**0,40**	0,01	-0,12	-0,08
Si($\mu mol\,l^{-1}$)	**-0,21**	-0,09	**-0,29**	-0,02	**0,15**	**0,50**	**-0,15**	-0,01
Fluo ($\mu g\,l^{-1}$)	-0,10	-0,11	0,11	0,02	0,02	**-0,15**	**0,57**	0,04

Le niveau 4 identifie 15 classes encore plus courtes. La dynamique est très largement réduite dans ce cas de figure (Figure 3.9l) plus aucun état n'a de période supérieure à la semaine. Dans notre cas, le niveau 3 apparaît déjà comme suffisamment complexe pour identifier des schémas fonctionnels spécifiques et le niveau 4 ne fait que rajouter de la complexité sans apporter d'informations supplémentaires au vu de nos capacités d'interprétation actuelles, c'est pourquoi il n'est pas détaillé ici.

3.4.3 Phase de labellisation : Interprétations écologiques et taxonomiques

Comme présentée dans la section 3.4.2, l'approche multivariée et multi-niveau de *M-SC* permet une première caractérisation des classes. Néanmoins, il est important de relier ses classes à d'autres informations (Météorologie, biologique ...) afin de définir une labellisation cohérente des classes en états écologiques. Dans ce sens, une stratégie d'étude multivariée, multi-source et multi-échelle est mise en place. Un diagramme de Margalef étendu à 6 variables (Section 2.3.2.4) est utilisé pour faciliter l'inter-comparaison entre les classes et reproduire le caractère multivariée de *M-SC*. Une approche taxonomique est combinée à ce diagramme (Section 2.3.2.1 et 2.3.2.2) afin d'approfondir la phase de labellisation et de relier au niveau spécifique les classes à la dynamique du phytoplancton. Elle est réalisée à partir des données BF d'abondance phytoplanctonique issues d'une nouvelle source de données : les données du REPHY (section 2.1.2). Dans cette section, une première partie expose le pré-traitement des données BF et les choix d'analyse faits pour combiner les différentes échelles d'échantillonnage et les différentes variables. Une seconde partie présente les résultats issus de la classification, du diagramme de Margalef étendu et de l'approche taxonomique. À partir de cette analyse, un *label* (= état écologique) associé à chaque classe est déterminé.

3.4.3.1 Phase de pré-traitement

Les stratégies d'études multi-sources et multi-échelles nécessitent d'adapter les fichiers et les choix des périodes d'analyses pour combiner de manière cohérente les informations et les résultats. Plusieurs analyses préalables sont réalisées pour choisir les approches les plus adaptées afin de définir l'assemblage taxonomique relatif à chaque classe.

Pas de temps Premièrement, le pas de temps d'échantillonnage des campagnes REPHY (15 jours) est relativement faible par rapport au pas de temps des données MAREL-Carnot (20 minutes). Pour rester en accord avec les données HF et ne pas dégrader ce signal, nous avons choisi d'aligner les données BF au pas de temps HF et de dupliquer une partie du signal BF pour limiter le nombre de données manquantes (*NA*). Les valeurs d'abondance taxonomique sont donc dupliquées sur un pas de temps (dt) avant et après la mesure. En plus de limiter le nombre de *NA*, ce choix permet une meilleure identification des assemblages. Sans la duplication aucune correspondance ne serait trouvée entre les abondances et les évènements d'une classe dont la période est inférieure à la quinzaine de jours.

Afin de ne pas masquer des événements structurants lors de l'adaptation du pas de temps (dt) des données BF, celui-ci doit correspondre à une période où il y a très peu de changements de structure, de dynamique. L'analyse spectrale par décomposition modale empirique (*EMD*) (Méthodes section 2.3.1.1) est appliquée afin de définir cette période. La décomposition *EMD* produit plusieurs modes (*IMF*) et identifie plusieurs cycles caractéristiques du signal. Chaque mode est représenté pour une année et correspond à une moyenne des modes des 5 années (Figure 3.10). Les imf.6 et imf.7 sont les modes les plus proches de la tendance du signal MAREL-Carnot. Pour calculer leurs fréquences et donc leurs périodes, une analyse de Fourier est mise en œuvre. D'après le calcul des statistiques descriptives (Tableau 3.12) un pas de temps (dt) entre 2 et 4 jours nous assure de rester cohérents au regard de la dynamique phytoplanctonique. Le pas de temps dt est donc fixé à 3 jours. Par exemple, la mesure du 10 janvier sera donc aussi utilisée pour couvrir les données acquises entre le 7 et le 13 janvier.

Tableau 3.12 – Statistiques descriptives des périodes des imf.6 et imf.7. Le minimum (Min), la moyenne (Moy), la médiane (Med), le maximum (Max) et le premier (Q1) et troisième (Q3) quantiles des périodes sont présentés en minutes, en heures et en jours.

imf.6	**minutes**	**heures**	**jours**	imf.7	**minutes**	**heures**	**jours**
Min	1429	24	0,9	**Min**	3030	51	2
Q1	1843	131	1	**Q1**	4162	69	2
Med	2389	40	2	**Med**	5838	97	4
Moy	2673	44	2	**Moy**	6569	109	4
Q3	3354	56	3	**Q3**	8340	139	6
Max	5243	88	4	**Max**	13846	230	10

Période significative Deuxièmement, toutes les classes ont une dynamique propre et certaines d'entre elles ont des périodes d'apparitions moins fréquentes. De plus, chaque classe est composée de plusieurs évènements dont les périodes peuvent différer. Cette variabilité temporelle rend difficile la concordance entre les classes et la distribution phytoplanctonique. Nous avons donc fait le choix de définir les assemblages taxonomiques à partir des périodes les plus significatives possibles. Ces périodes sont définies en fonction du nombre de points appartenant à une classe (Nombres d'occurrences) par mois. Nous avons choisi d'identifier des périodes par mois pour être concordants avec les mesures BF (= 2 données par mois). Une période sera considérée comme significative en fonction d'un seuil arbitraire du nombre total d'occurrences (Tot) (Tableau 3.13) : (i) Pour Tot > 10 000 sont retenus les mois dont le nombre d'occurrences est supérieur à 2 000 ; (ii) sinon, sont sélectionnés les mois supérieurs à 1 000. Dans le tableau 3.13, ces périodes significatives sont représentées en gras. Par exemple, pour la classe 7, les mois de 03-Mars, 04-Avril, 05-Mai sont retenus. Pour la classe 8 la période a été étendue aux mois 01-Janvier et 03-Mars compte tenu des valeurs proches de 1 000.

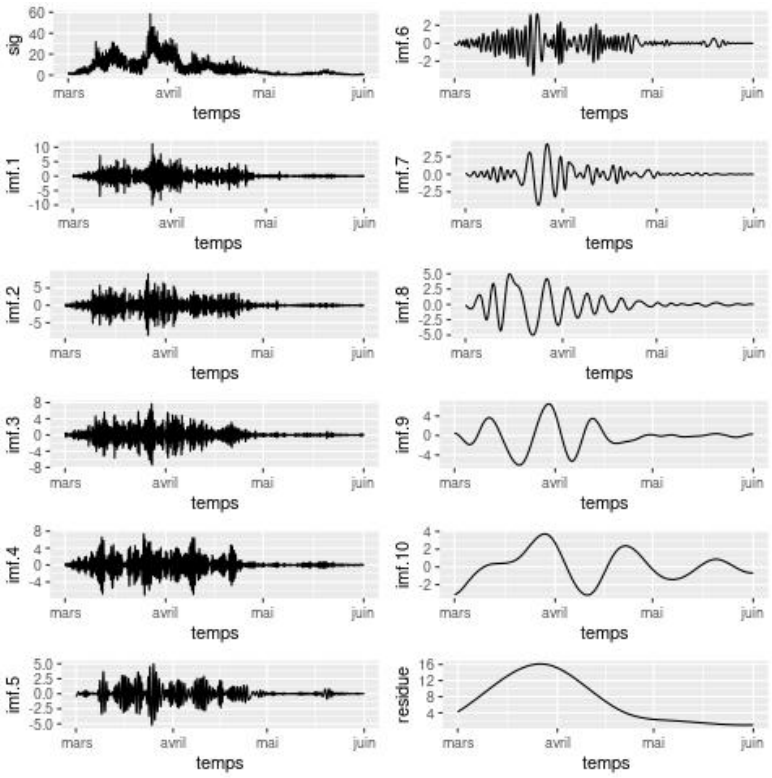


FIGURE 3.10 – Décomposition modale empirique (EMD) du signal de fluorescence MAREL-Carnot de 2005 à 2009, moyennée sur 1 an. Présentation des imf 1 à 10 ainsi que du signal initial (sig) et du résidu (residue).

TABLEAU 3.13 – Tableau de contingence : Nombres de points (occurrences) appartenant à une classe en fonction du mois. En gras les périodes significatives, soit pour Tot >10 000 le nombre d'occurrences > 2 000 et Tot <10 000 le nombre d'occurrences > 1 000.

Classe	1	2	3	4	5	6	7	8	non classé
mois									
01	0	0	0	0	691	**8217**	8	**937**	1307
02	0	0	0	0	67	**5329**	0	**1607**	3149
03	0	0	98	0	210	**2528**	**3753**	**870**	3701
04	0	38	**2642**	67	547	1229	**6074**	201	2
05	252	**3062**	**5164**	0	351	282	**2049**	0	0
06	**2475**	**4007**	**3107**	10	**1154**	38	9	0	0
07	**9427**	934	10	0	258	0	0	0	531
08	**7156**	900	0	0	176	0	0	0	2928
09	**4669**	**2458**	0	0	242	141	0	0	3290
10	965	**2981**	12	0	11	252	0	0	6939
11	0	**2041**	7	0	52	**3851**	4	0	4845
12	0	306	0	0	97	**7897**	600	26	2234
Tot	24944	16727	11040	77	3856	29764	12497	3641	28926

3.4.3.2 Définition des états environnementaux par classes

Basé sur le principe du diagramme de Margalef (Section 2.3.2.4), une nouvelle version étendue à 6 variables est présentée pour caractériser les propriétés biogéochimiques de chaque classe (Figure 3.13).

Ce diagramme rend compte de l'équilibre entre : la disponibilité en sels nutritifs, des conditions d'éclairement et les degrés de mélange, sur la base de 6 variables explicatives : Turbidité, *PAR*, Nitrate, Phosphate, Silicate et le rapport N/P. À partir du gradient de ces trois caractéristiques, il est possible d'identifier rapidement les facteurs d'influence de chaque classe. Il est ainsi possible de définir les conditions écologiques et de les relier à trois stratégies de vie correspondant à : des stratégies colonialistes-invasives (C) favorisées pas des milieux stables et riches en sels nutritifs, des stratégies rudérales (R) favorisées pas des milieux riches en sels nutritifs et des conditions de mélange fortes et d'éclairement faible, des stratégies tolérantes au stress (S) qui offrent des avantages de développement en conditions instables (turbulent, pauvre en nutriments). Ainsi, pour chaque variable et pour chaque classe, le barycentre des données sur la période 2005-2009 est calculé. Le barycentre est le centre de gravité de chaque nuage de points correspondants à une classe. Il rend compte de la valeur moyenne des variables pour chacune des classes. Les unités de mesures de chaque variable sont différentes. Pour faciliter l'inter-comparaison entre chacune, tous les barycentres sont normalisés (centré-réduit). Puis les variables sont représentées deux à deux dans un graphique d'ensemble où chaque classe est désignée par son numéro et une couleur (Figure 3.13). Cette méthode permet de synthétiser les informations multivariées mais tous les résultats peuvent être calculés par variables

Des analyses supplémentaires (répartition des classes sur le signal (Exemple Figure 3.11), boîte de dispersion par classe et par variable (Exemple Figure 3.12), tableau de corrélation sont aussi mis en liens avec les conclusions.

Pour aller plus loin, une étude taxonomique est associée afin de déterminer les assemblages caractéristiques pour chaque état (Figure 3.15). Ces assemblages sont déterminés à partir des données du réseau de surveillance REPHY (section 2.3.2.2). Ce fichier contient les abondances de 97 unités taxonomiques différentes (Annexe A). Un assemblage correspond à un regroupement de taxons dominants, c'est-à-dire dont l'abondance cumulée est >95 % de l'abondance totale.

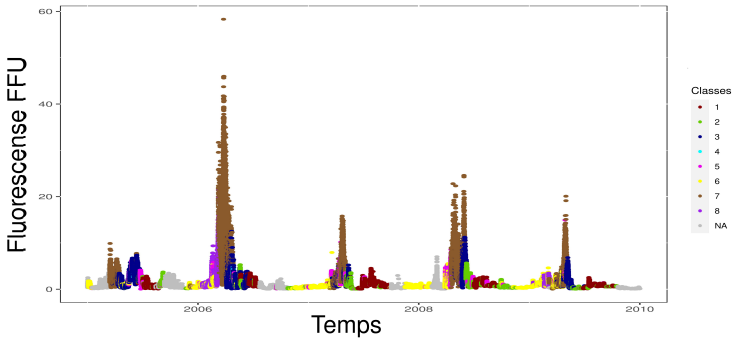
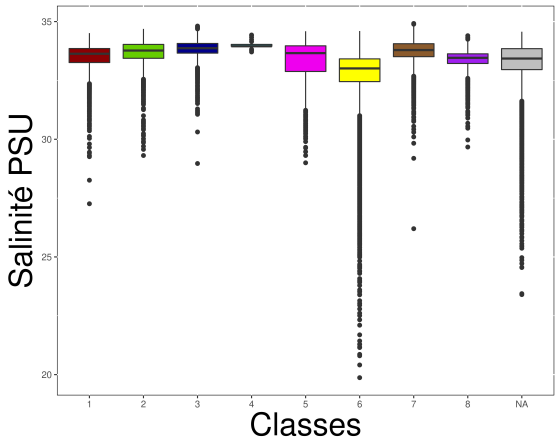
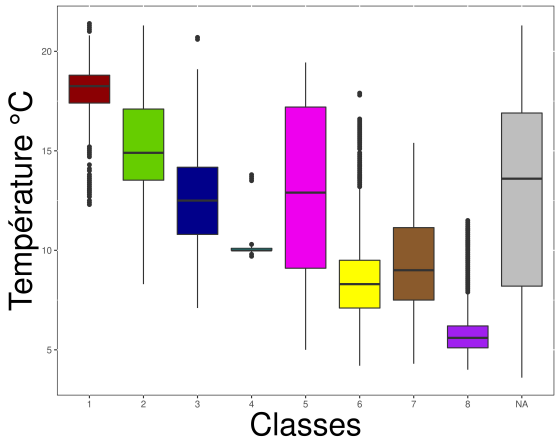


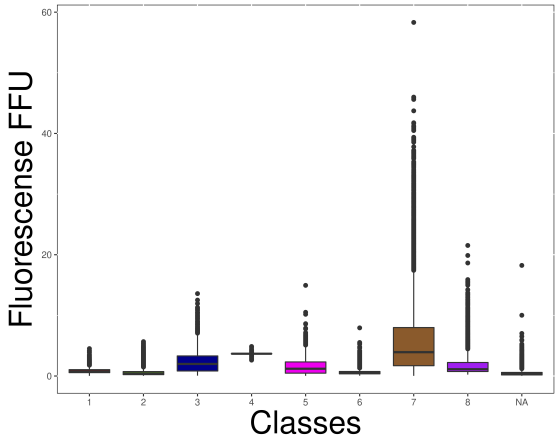
FIGURE 3.11 – Distribution des classes M-SC au niveau 3 reportées sur le signal de fluorescence de MAREL-Carnot 2005-2009.



(a)



(b)



(c)

FIGURE 3.12 – Classification au niveau M-SC 3 des données MAREL-Carnot sur la période 2005-2009. Boîte de dispersion par classe (a) de la salinité (PSU), (b) de la température (°C) et (c) de la fluorescence (FFU).

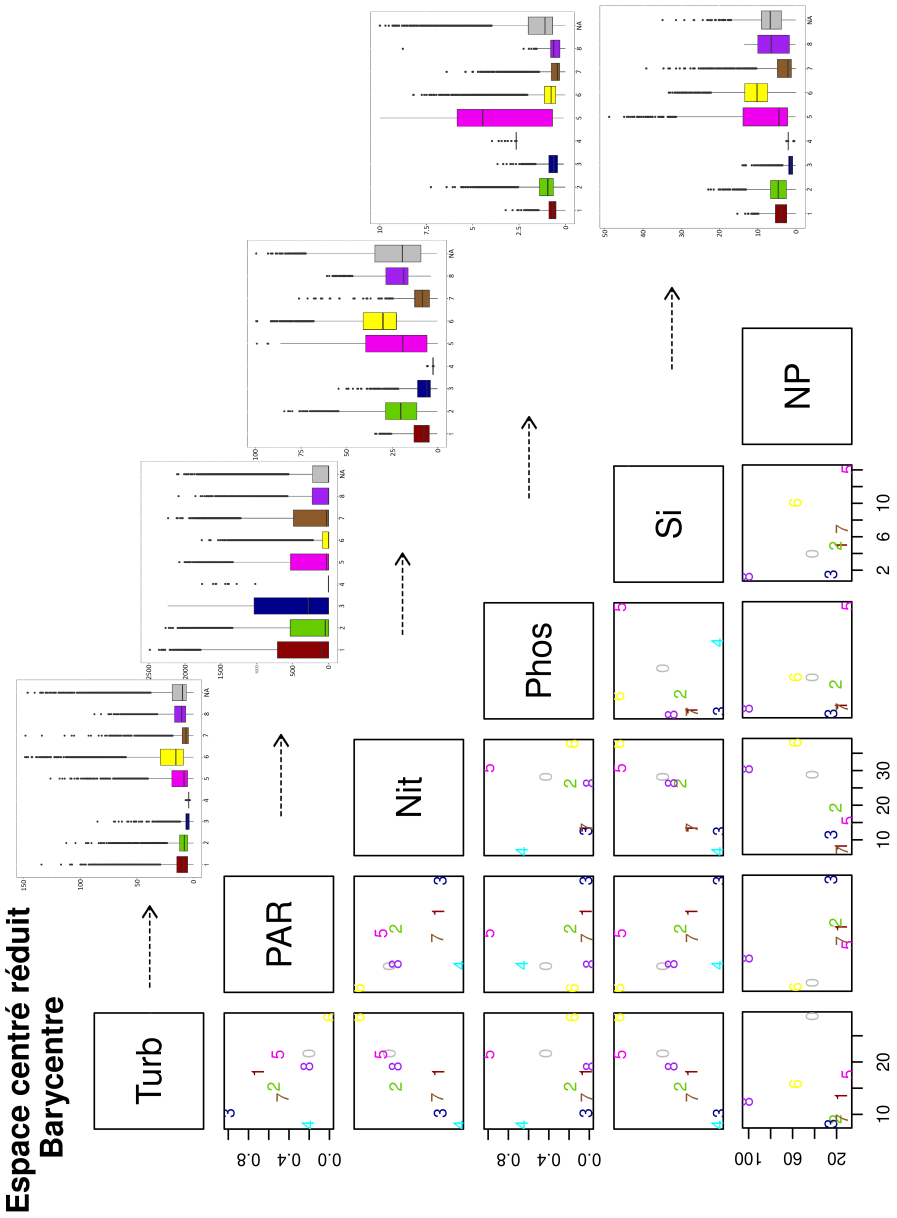


FIGURE 3.13 – Représentation, inspirée de l'approche de Margalef, avec 6 facteurs de contrôles principaux de la dynamique phytoplanctonique (Turb : Turbidité, PAR : *Photosynthetically active radiation*, Nit : Nitrate, Phos : Phosphate, Si : Silicate, NP : rapport de Redfield Nitrate sur Phosphate). Chaque classe est représentée par le barycentre normalisé des données sur la période 2005-2009. Elle est identifiée par son numéro et une couleur associée (cl1 rouge, ...). Pour chaque variable est ajoutée, sur la ligne correspondante au paramètre, la boîte de dispersion par classe.

La méthode de classification *M-SC* permet d'identifier 8 classes réparties sur un ou plusieurs mois de l'année et généralement restreintes à une saison (Figure 3.9i, Tableau 3.13). À l'aide des différentes approches présentées dans les paragraphes précédents, nous allons labelliser ces classes (cl_i) en états environnementaux et leur attribuer un label (l_i).

En Période hivernale, la classe 6 (cl6, en jaune) est identifiée comme classe caractéristique de **la période hivernale non productive (l6)** (Figure 3.9h). Elle est structurée par un faible niveau de lumière et de fortes concentrations en sels nutritifs (Figure 3.13, colonne PAR, Ligne Nit et Si). Lors de cette saison, il est fréquent de rencontrer des épisodes de tempêtes qui génèrent un brassage intensif et des apports d'eau douce importants. Ainsi, cl6 est aussi structurée par les plus fortes baisses de salinité (Figure 3.12a). À cette classe est bien associée un assemblage taxonomique hivernal (Figure 3.15) composé de taxons comme *Thalassiosira*, *Thalassionema*, *Chaetoceros* et d'espèces euryhalines telles que *Skeletonema costatum* ou encore *Asterionellopsis sp.* ce qui est cohérent avec les dessalures. Bien que restreinte aux mois les plus significatifs (En gras Tableau 3.13), cette classe s'étend sur une période de 5 mois de novembre à mars. Ainsi, des espèces plus automnales et printanières telles que les *Crypthophytes* ou *Gymnodiniales* sont donc aussi retrouvées.

Comme cl6, la classe 8 (cl8, en violet) est identifiée en fin de période hivernale (Figure 3.9h), elle présente néanmoins des conditions d'éclairement plus favorables au développement phytoplanctonique que cl6. cl8 se différencie aussi de cl6 par des concentrations en sels nutritifs plus faibles notamment pour les concentrations en Silicate (Figure 3.13, ligne Si). Globalement, cl8 présente des concentrations en sels nutritifs plus élevées que la majorité des autres classes. De plus, cl8 est caractérisée par un rapport N/P bien plus élevé (Figure 3.13, ligne NP). Ce rapport suggère que cette classe est **un état de transition entre la période productive et non productive (une amorce ou le début du bloom) (l8)**. Dans l'assemblage taxonomique de cette classe restreinte (mois ≥ 1000 occurrences), sont effectivement retrouvées des taxons hivernaux comme *Thalassiosira*, *Chaetoceros*, *Asterionellopsis sp.* mais aussi, en abondance plus importante que pour cl6, des espèces printanières du genre *Pseudo-nitzschia*. cl8 peut être définie comme un état pré-printanier. Des efflorescences en fin d'hiver sont de plus en plus fréquemment observées [Lefebvre, Guiselin et al. 2011] et des blooms d'algues vertes peuvent se produire avant l'efflorescence printanière de diatomées [Breton, Brunet et al. 2000].

En période printanière se succèdent les classes 7 (cl7, en marron) et 3 (cl3, en bleu)(Figure 3.9h). Elles sont caractérisées par des concentrations de nutriments relativement faibles et un éclairement moyen pour cl4 et fort pour cl3 (Figures 3.13, ligne PAR et clonne Nit, ligne Phos). Un gradient est observé au niveau des nutriments avec des concentrations plus importantes pour cl8 que pour cl7 (où les nutriments chutent considérablement) et cl3 (où ils sont presque épuisés) (Figure 3.13, boîte de dispersion Nit, Phos et Si). Ces deux classes se distinguent aussi des autres classes par une fluorescence élevée (Figure 3.12c). cl7 est particulièrement structurée par cette variable (Tableau 3.11). Elles correspondent donc toutes **les deux à un état d'efflorescence où la biomasse et la consommation des nutriments sont importantes (l3, l7)**. cl3 se différencie globalement de cl7 par des niveaux d'éclairement beaucoup plus forts et une turbidité plus faible (Figures 3.13, colonne Turb, ligne PAR). cl3 a aussi des concentrations en silicate plus faibles que cl7 où il est encore présent dans le milieu (Figure 3.13, boîte de dispersion Si). Ces deux classes ont donc des facteurs d'influence bien distincts et cette différence se retrouve au niveau des assemblages phytoplanctoniques. cl7 est exclusivement dominée par les ***Phaeocystis* (l7)**. Pour cl3, aux *Phaeocystis* s'ajoutent les genres ***Pseudo-nitzschia***, *Cerataulina*, *Rhizosolenia* **(l3)**. Il est donc possible, à partir de cette classification, de dissocier les efflorescences nuisibles de *Phaeocystis*, de celles avec la combinaison *Phaeocystis* et *Pseudo-nitzschia* dont les conséquences

pourraient être plus importantes puisque cumulant le développement d'un taxon nuisible et d'un taxon potentiellement toxique.

Dans le cas présent, *M-SC* identifie une succession printanière de trois états environnementaux l8, l7 et l3, représentés par cl8, cl7 et cl3 (Figure 3.11). Ces trois évènements peuvent s'apparenter à un état d'efflorescence où la biomasse phytoplanctonique et donc la consommation des nutriments sont importantes. Toutefois, chaque classe à des facteurs d'influences bien distincts. Durant ces trois états, se produit une chute des sels nutritifs et un gradient au niveau des nutriments de cl8 à cl3 est observé. De même, cl8 a un niveau d'éclairement plus faible que cl3 (Figure 3.13).

Une partie de la période printanière mais aussi la période automnale est associée à la classe 2 (cl2, en vert) (Figure 3.9h). Tout comme cl8, elle se distingue des autres classes par des concentrations en nutriments élevées. Mais le rapport N/P de cl2 est plus faible que celui de cl8 ou encore cl6, qui sont des états où les stocks de tous les nutriments sont au maximum (Figures 3.13, colonne Nit). Ce plus petit rapport indique que l'état écologique qui correspond à cl2 aura tendance à se produire lors d'une ré-augmentation des concentrations en nutriments, mais après une période d'épuisement des stocks de nutriments. De plus, l'assemblage est représenté par deux taxons dominants estivaux, *Cryptophytes* et *Chaetoceros*, mais aussi des taxons hivernaux comme *Thalassiosira*, *Gymnodiniales*, *Asterionellopsis sp.*, même si les conditions de regroupement sont identiques, il est plus cohérent, d'un point de vue temporel et taxonomique d'étudier ces deux périodes indépendamment l'une de l'autre.

Ainsi, il est possible d'observer la classe cl2 sur deux périodes biologiques bien distinctes : cl2.1 **en fin de bloom printanier (l2.1)**, après les blooms de *Phaeocystis* (cl7) et de *Pseudo nitzchia* (cl3), et cl2.2 **en automne (l2.2)**. Cette distinction se fait aussi au niveau des assemblages taxonomiques. Au niveau de cl2.1, c'est un assemblage printanier commun à cl3, mais plus diversifié qui est retrouvé. La combinaison *Phaeocystis* et *Pseudo-nitzschia* est dominante, mais l'on retrouve aussi en plus faible quantité des taxons présents toute l'année tels que les taxons *Chaetoceros*, *Cryptophycées*, *Asterionellopsis sp.* . L'assemblage de cl2.2 est aussi un assemblage mixte, on y retrouve un assemblage hivernal commun avec cl6 et mélangé avec un assemblage plus estival comme cl1. Ainsi, deux sous-classes cl2.1 et cl2.2 sont donc définies : cl2.1 pouvant être apparentées à la fin du bloom printanier et cl2.2 a un potentiel bloom automnal.

En période estivale, la classe 1 (cl1, en rouge) est identifiée (Figure 3.9h). Elle est structurée par de faibles concentrations en nutriments et un fort éclairement (Figure 3.13, ligne PAR et ligne Si). C'est **un état non productif avec une limitation de la croissance phytoplanctonique essentiellement par les nutriments et aussi potentiellement par les trop fortes lumières en surface (l1)**. La transition entre les états cl3 (bloom printanier) et cl1 (période estivale) est liée à des températures plus importantes (Figure 3.12b) et des maximums d'éclairements même si en moyenne les taux d'éclairement sont plus importants pour cl3 (Figure 3.13, ligne PAR). À l'inverse, le passage vers des états plus automnaux (cl2) ou hivernaux (cl6) se traduit par une différence au niveau des concentrations en nutriments (beaucoup plus faibles pour cl1) (Figure 3.13, colonne Nit, Si, Phos). C'est un état de transition défini par une combinaison de taxons aux stratégies différentes. La classe cl1 est une des classes les plus diversifiées avec au total 62 taxons identifiés dont 13 dominants (<95 % de l'abondance totale). De ces 13 taxons dominants, le taxon *Chaetoceros* ressort largement. Il est courant de retrouver ce taxon, composé d'espèces de petite taille, sur la zone d'avril à octobre. Des taxons pré-estivaux tels que les *Cryptophytes* ou encore des diatomées estivales comme *Guinardia* ou *Rhizosolenia* sont aussi retrouvés dans cl1.

Un évènement particulier est mis en évidence par la classe 5 (cl5, en rose). Cette classe a une dynamique de plus courte durée (Figure 3.9i), plus proche des évènements qualifiés de rares,

extrêmes. cl5 est fortement structurée par les fortes concentrations en nutriments (Figure 3.13, colonne Nit, Phos et Si). Cette classe courte, identifiée à des périodes différentes, correspond à des apports de nutriments brefs et importants. Il correspond aussi à des périodes de diminutions de la salinité (Figure 3.12a). De plus, elle est caractérisée par un assemblage constitué majoritairement de *Chaetoceros* et de *Phaeocystis*.

Ainsi, cet état peut être qualifié d'**évènement hivernal et printanier de quelques jours où les concentrations en nutriment sont particulièrement élevées (l5)**. Cet évènement est caractéristique d'une réponse du phytoplancton à un apport de nutriment dans le milieu. Ces apports peuvent être liés à de multiples facteurs comme une tempête, des périodes de crues importantes ou encore à des activités anthropiques telles que les dragages du port. À ce moment, les sels nutritifs deviennent disponibles dans la colonne d'eau (alors qu'ils sont en théorie à de faibles concentrations) et ils pourront donc être consommés par le phytoplancton. Ces conditions environnementales particulières pourraient expliquer certaines proliférations secondaires se produisant parallèlement à la dynamique générale par exemple les études des proliférations d'algues nuisibles. Toutefois, nous n'avons pas pu corréler ces apports à un facteur précis. En effet, la réponse est conditionnée par un effet cumulé. Ainsi, pour un même facteur, elle ne sera pas forcément analogue suivant les autres conditions. En ce qui concerne cl5, l'apport doit précéder une période d'appauvrissement en sel nutritif, dans le cas contraire, un apport lié à une période de crue ne sera pas forcement identifié dans la classe. Afin d'approfondir l'identification des facteurs d'influence, une classification avec l'ajout des périodes de crue ou des périodes de dragages comme variables contributives aurait été intéressante. Nous avons toutefois décidé de les utiliser seulement comme variables complémentaires compte tenu de la forme et de la faible quantité de données que nous avions à disposition. La détection de ces évènements peut aider aussi bien à étudier les facteurs environnementaux qui déclenchent ces proliférations qu'à la calibration des modèles.

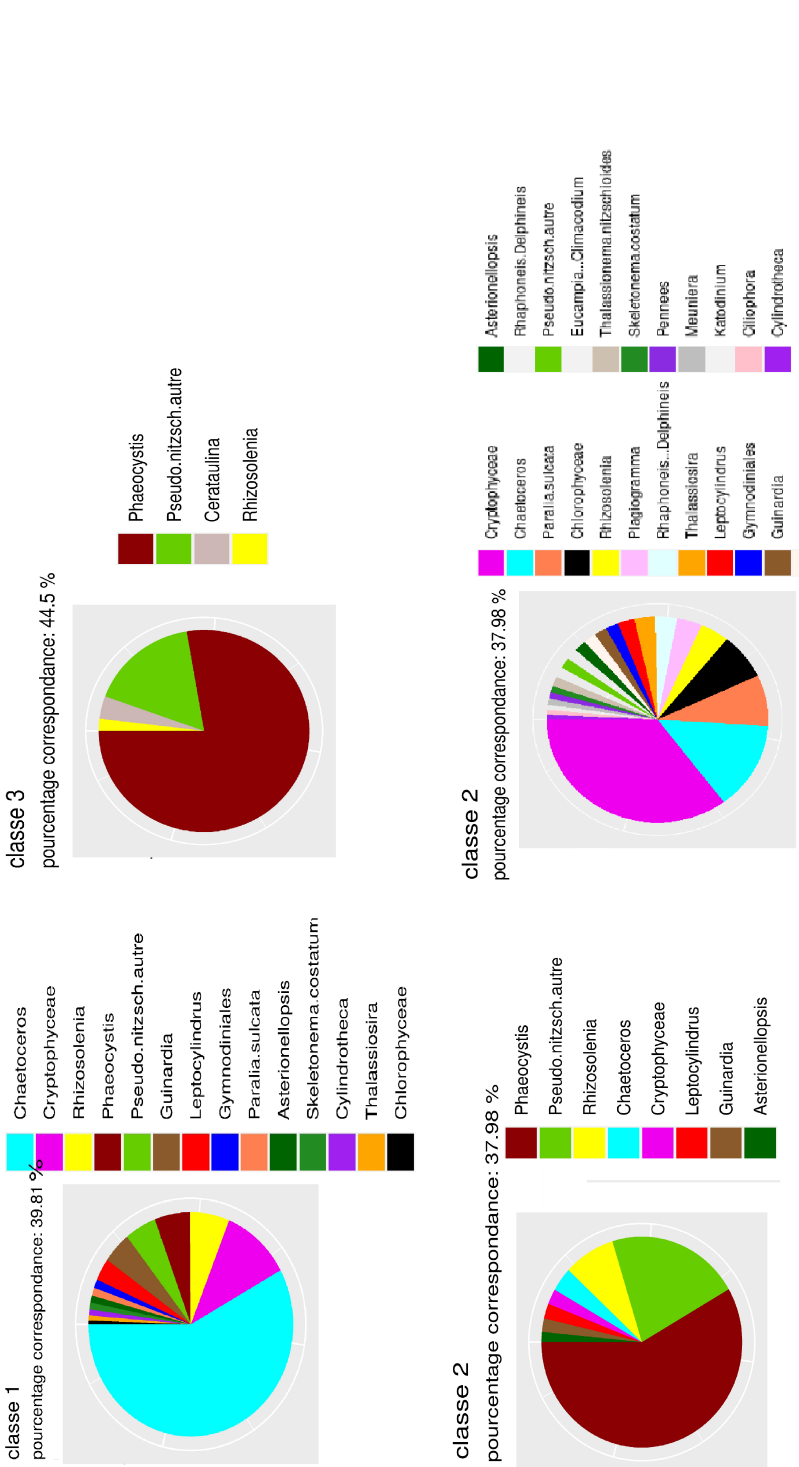


FIGURE 3.14 – Diagramme circulaire des assemblages des taxons dominants par classe. L'abondance cumulée de ces assemblages est $> 95\%$ de l'abondance totale. Le pourcentage de correspondance représente le pourcentage d'évènements pour lesquels un prélèvement REPHY a été trouvé.

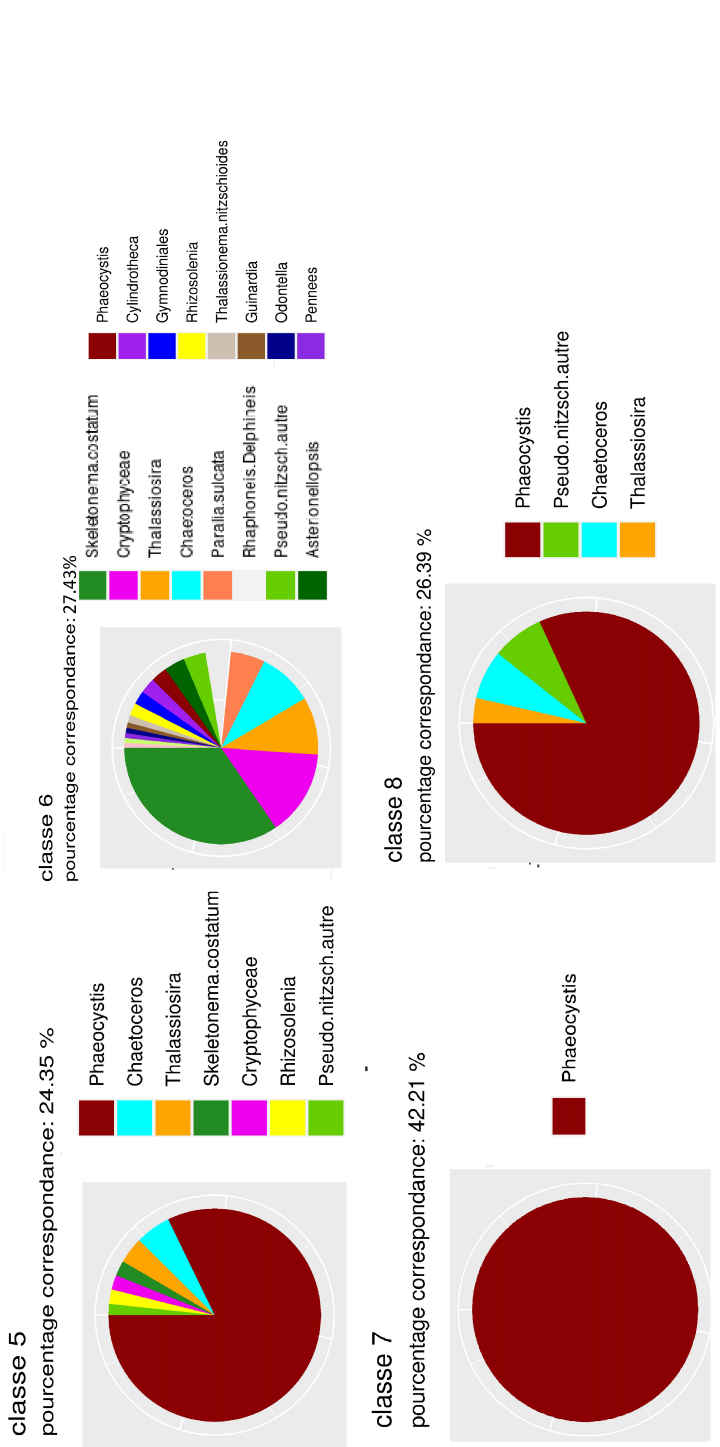


FIGURE 3.15 – Diagramme circulaire des assemblages des taxons dominants par classe. L'abondance cumulée de ces assemblages est $> 95\%$ de l'abondance totale. Le pourcentage de correspondance représente le pourcentage d'évènements pour lesquels un prélèvement REPHY a été trouvé.

3.5 Conclusion

Dans ce chapitre, une méthode de classification non supervisée *M-SC* est proposée afin d'identifier des états environnementaux dans des séries temporelles multivariées. Pour répondre au mieux à cette problématique plusieurs développements méthodologiques et optimisations ont été proposés tels que l'approche multi-niveau (NivMax) et le critère d'arrêt (sil.min) ou encore le protocole de pré-traitement. Ainsi *M-SC* combine actuellement 3 aspects de la classification : l'aspect spectral, l'aspect hiérarchique et l'aspect de cohésion via l'approche par densité (Section3.2.1).

Cette méthode a été validée, par une analyse comparative entre différentes méthodes de classification supervisées et non supervisées sur différents jeux de données (Section 3.3.3). Les résultats ont démontré que la méthode *M-SC* est bien adaptée à notre application. En effet, que ce soit pour une approche spatiale ou temporelle, elle compte parmi les méthodes les plus performantes pour définir des classes dans le cadre de signaux non linéairement séparables avec des formes complexes.

Enfin, l'approche a été appliquée sur le jeu de données MAREL-Carnot. Cette étape a permis de détecter et d'identifier avec plusieurs niveaux d'interprétations des états environnementaux et leurs caractéristiques biogéochimiques. Puis, l'analyse complémentaire combinant les résultats de la classification et les données taxonomiques ont permis de définir des schémas saisonniers, des successions d'évènements et des états extrêmes. Ainsi, pour chaque classe, nous sommes capables (i) de définir les paramètres biogéochimiques et de diversité, et (ii) de fournir une labellisation du jeu de données MAREL-Carnot. Cette labellisation va pouvoir être utilisée afin de détecter par apprentissage ces évènements sur d'autres sites d'études.

Chapitre 4

Vers un système d'aide à la décision sur d'autres sites d'études

4.1 Introduction

Au chapitre précédent, nous avons développé une méthodologie complète permettant d'aboutir à un système d'aide à la labellisation d'événements détectés dans des séries temporelles acquises à HF. Cette méthodologie a permis la labellisation pour la période 2005-2009 du site de référence de la station MAREL-Carnot. Dans le cas de la dynamique phytoplanctonique, la base labellisée peut permettre, par le biais de méthodes de *Machine Learning* supervisées, de développer des modèles d'apprentissage et de prédiction (Agent de classification), ce qui fournit un outil de reconnaissance automatique des événements, tels que les efflorescences d'algues potentiellement nuisibles (Figure 4.1). Ainsi, les classes (cl_i) issues de la classification *M-SC*, puis labellisées par expertise en labels (l_i) seront prédites sur de nouveaux jeux de données par le biais d'une méthode de classification supervisée (Figure 4.1).

Dans la première section de ce chapitre, nous allons analyser le potentiel de cette approche pour construire un système d'interprétation des événements phytoplanctoniques sur une période non apprise de la station MAREL-Carnot. Ensuite, les réponses du modèle de prédiction construit et les performances de reconnaissance seront évaluées sur d'autres jeux de données. Cette étape doit permettre de réduire le temps de labellisation des nouveaux jeux de données et améliorer la compréhension et la comparaison des réponses de chaque écosystème. Ici, ce sont les bases de données des stations MAREL-Iroise (Rade de Brest, mer d'Iroise) et MesuRho (Golfe du Lion, Méditerranée).

Nous rappelons qu'au chapitre 1, nous avons détaillé ces 3 stations avec pour objectif de retrouver des dynamiques saisonnières générales communes et de pouvoir analyser leurs différences. Il a été émis comme hypothèse qu'au regard de leurs implantations géographiques et de leurs conditions environnementales respectives, il était possible de mettre en évidence à la fois un aspect global au travers de schémas communs par rapport à une dynamique des systèmes tempérés, mais aussi un aspect plus spécifique au niveau de la variabilité de ces schémas, qui font la particularité de chaque site. Nous espérons donc déterminer et cibler des événements communs liés aux forçages globaux à grande échelle, et différencier des événements particuliers propres à chaque site d'études liés à des forçages plus locaux.

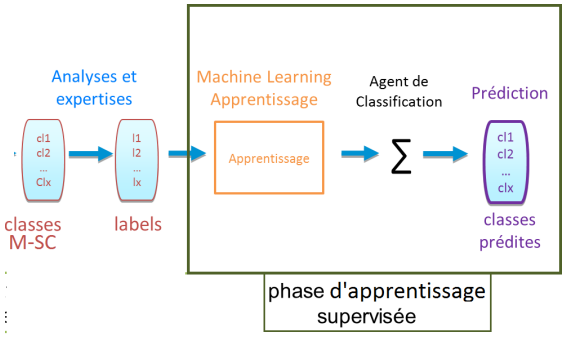


FIGURE 4.1 – Schématisation de la phase d'apprentissage de l'approche semi-supervisée.

4.2 Protocole d'étude

Nous reprenons ici le protocole associé à l'exploitation des connaissances acquises sur le site MAREL-Carnot et les différentes analyses faites sur deux autres sites géographiques.

4.2.1 Protocole d'apprentissage et jeux de données

Ainsi, trois bases de données sont exploitées :

— La première base est constituée du jeu de données MAREL-Carnot ainsi que les labels (l_i) associés à chaques observations au chapitre 3 (Section 3.4.3).

— Les deuxième et troisième bases sont des jeux de données non labellisés issus des stations de mesures MAREL-Iroise et MesuRho (Section 2.1.1.2). Ces deux stations mises en place respectivement en 2000 et en 2009, mesurent entre 5 et 10 variables avec un pas de temps de 20 minutes pour MAREL-Iroise et de 30 minutes pour MesuRho.

Pour chacune de ces bases, le même protocole de pré-traitement est appliqué : alignement, correction des gammes, suppression des points avec le code qualité 4, et complétion. Ce protocole défini pour la base MAREL-Carnot est détaillé Section 3.2.2.

Afin de labelliser les jeux de données MAREL-Iroise et MesuRho, il est nécessaire d'avoir une base de connaissance (observations, labels) pour appliquer une méthode de classification supervisée. N'ayant pas à ce jour de bases labellisées sur ces sites, l'algorithme *M-SC* sera appliqué pour détecter les événements particuliers de chaque site (cl_i). La même calibration que pour MAREL-Carnot est utilisée : crit = PEV, sil.min= 0,7 et niveau = 3 (Section 3.4.1.2). Ces clusters d'événements issus de *M-SC* seront utilisés comme labels vrais. Néanmoins, il est important de noter que ces regroupements sont basés uniquement sur la géométrie des données et font office d'expertise / de vérité, puisqu'aucun expert scientifique n'a labellisé les données. Le degré de labellisation est donc plus faible que celui présenté pour MAREL-Carnot dans le chapitre 3. Toutefois, il permet d'avoir un ordre d'idée de la répartition des données aux stations MAREL-Iroise et MesuRho et d'identifier les points communs et les différences de dynamique avec les classes labellisées de MAREL-Carnot (l_i).

Pour réaliser la classification et la labellisation automatique, le nombre et le type d'entrées présentés à un classifieur doivent être identiques. Par conséquent, seules les EOVs communes aux trois bases sont utilisées dans ce cas d'étude, à savoir : la salinité (PSU), la température (°C), l'oxygène dissous ($\mathrm{mg\,l^{-1}}$), la turbidité (NTU), le *PAR* ($\mathrm{\mu mol\,m^{-2}\,s^{-1}}$). Le signal de fluorescence

est ajouté à ces 5 entrées uniquement lorsque celle-ci est disponible en unité (FFU). Il est mesuré en FFU ou en $\mu g\,l^{-1}$ en fonction des stations et des périodes (Section 2.1) avec aucune conversion fiable disponible entre ces deux unités. Le signal de fluorescence en FFU est disponible uniquement pour MAREL-Iroise à partir de 2012, il sera utilisé seulement dans ce cas de figure en entrée de notre modèle sinon à titre de variable d'interprétation. Le modèle de prédiction est construit à partir d'une base d'apprentissage (observations, labels). Une fois validé, il permettra de labelliser une nouvelle base, appelée base de test. La base de test permet de vérifier les capacités du modèle à prédire l'état environnemental (label l_i) d'une nouvelle observation. Différentes configurations ont été testées.

Dans le premier cas (Section 4.2.2), nous cherchons à évaluer la capacité d'apprentissage, c'est-à-dire la capacité de la méthode à construire un modèle de reconnaissance performant, sur le jeu de données MAREL-Carnot (Figure 4.2 phase 1). Cela revient à vérifier que l'algorithme d'apprentissage choisi est capable d'apprendre correctement les labels de la base MAREL-Carnot issus de la classification *M-SC*. Pour valider la robustesse du système appris, nous avons fait plusieurs tests avec des jeux d'apprentissage et de test différents. Pour chaque configuration, le jeu test correspond à une année particulière A_i et le jeu d'apprentissage aux autres années disponibles $A\backslash\{A_i\}$ dans l'ensemble $A = \{2005 - 2009\}$. Le tableau 4.1 reprend les 5 premières configurations testées fonction de l'année test isolée. Par exemple, si 2005 appartient au jeu test, le jeu d'apprentissage correspond aux autres années disponibles dans l'ensemble 2005-2009 soit 2006, 2007, 2008, 2009.

Tableau 4.1 – Récapitulatif des différents jeux d'entraînement et de test l'évaluation de la capacité d'apprentissage.

Configurations	**Jeux d'entraînement**	**Périodes**	**Jeux test**	**Périodes**
1	MAREL-Carnot	2006,2007,2008,2009	MAREL-Carnot	2005
2	MAREL-Carnot	2005,2007,2008,2009	MAREL-Carnot	2006
3	MAREL-Carnot	2005,2006,2008,2009	MAREL-Carnot	2007
4	MAREL-Carnot	2005,2006,2008,2009	MAREL-Carnot	2008
5	MAREL-Carnot	2005,2006,2007,2008	MAREL-Carnot	2009

Pour le second cas (Section 4.3.2), c'est la capacité de généralisation du modèle à reconnaître un type d'événement sur un autre site que nous cherchons à estimer (Figure 4.2 phase 2). Ainsi, nous sélectionnons la totalité de la base labellisée de MAREL-Carnot comme base d'entraînement et de validation. Les bases MAREL-Iroise ou MesuRho, avec les groupes issus de *M-SC* non labellisés par un expert scientifique, seront utilisées comme bases de test (Tableau 4.2).

Tableau 4.2 – Récapitulatif des différents jeux d'entraînement et de test pour l'évaluation de la capacité de généralisation.

Configurations	**Jeux d'entraînement**	**Périodes**	**Jeux test**	**Périodes**
6	MAREL-Carnot	2005 à 2009	MAREL-Iroise	2006
7	MAREL-Carnot	2005 à 2009	MAREL-Iroise	2014 à 2016
8	MAREL-Carnot	2005 à 2009	MesuRho	2009
9	MAREL-Carnot	2005 à 2009	MesuRho	2010 à 2014

Cette capacité de généralisation sur les autres sites sera évaluée en fonction de deux points. Premièrement, nous allons évaluer les similitudes et les différences de structures dans les jeux de

données (Figure 4.2 phase 2.1). Pour cela nous comparons les classes (cl_i) issues d'une classification par *M-SC* aux labels prédits (l_i) par le modèle de reconnaissance construit à partir de la base de données labellisée MAREL-Carnot. Cela permet de comparer les partitionnements entre les cl_i *M-SC* de MAREL-Iroise et MesuRho et les labels prédits l_i sur ces mêmes jeux de données. Il est ainsi possible d'identifier si la dynamique qui régit le modèle construit à partir de MAREL-Carnot est similaire à la dynamique des nouveaux jeux de données MAREL-Iroise et MesuRho et donc de déduire si le modèle est compétent pour effectuer une prédiction significative. Deuxièmement, c'est la répartition temporelle des labels prédits sur les nouveaux jeux de données MAREL-Iroise et MesuRho qui est comparé à la répartition temporelle des classes issues de *M-SC* à MAREL-Carnot (Figure 4.2 phase 2.2) ; ces mêmes classes labellisées définissent les labels l_i appris, c'est-à-dire utilisés pour construire le modèle de prédiction. Cela permet d'évaluer si la prédiction sur les sites MAREL-Iroise et MesuRho, qui ont des caractéristiques environnementales différentes, permet de retrouver une cohérence temporelle vis-à-vis des labels prédits et de leurs dynamiques environnementales.

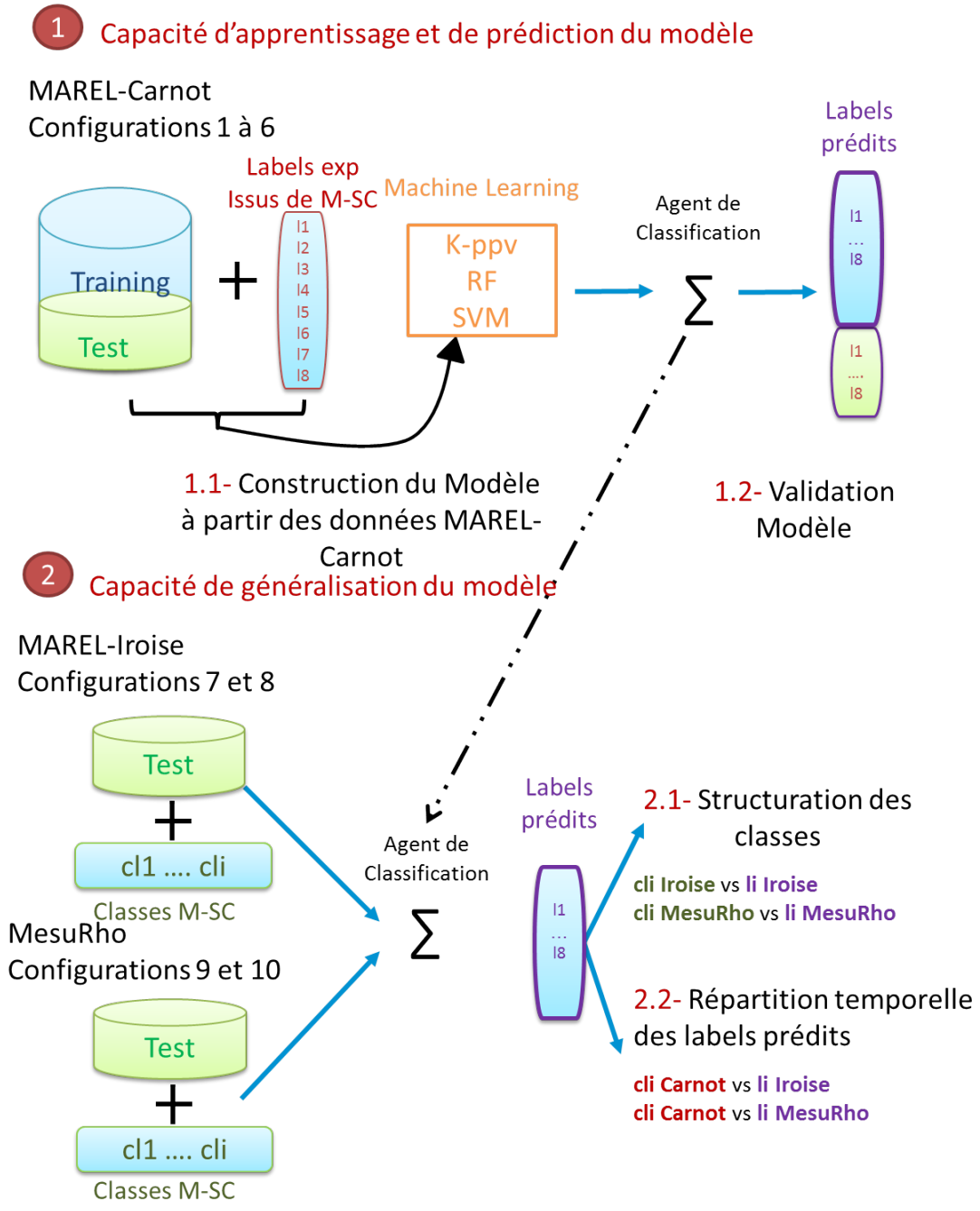
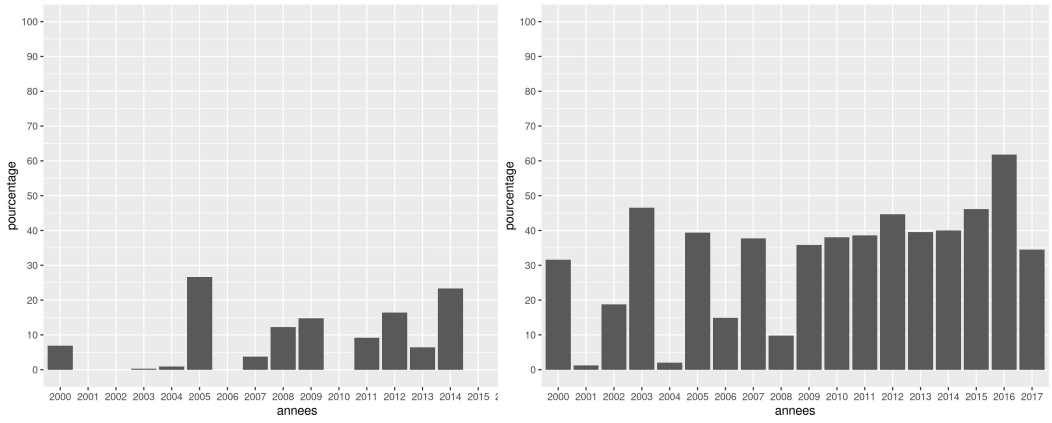


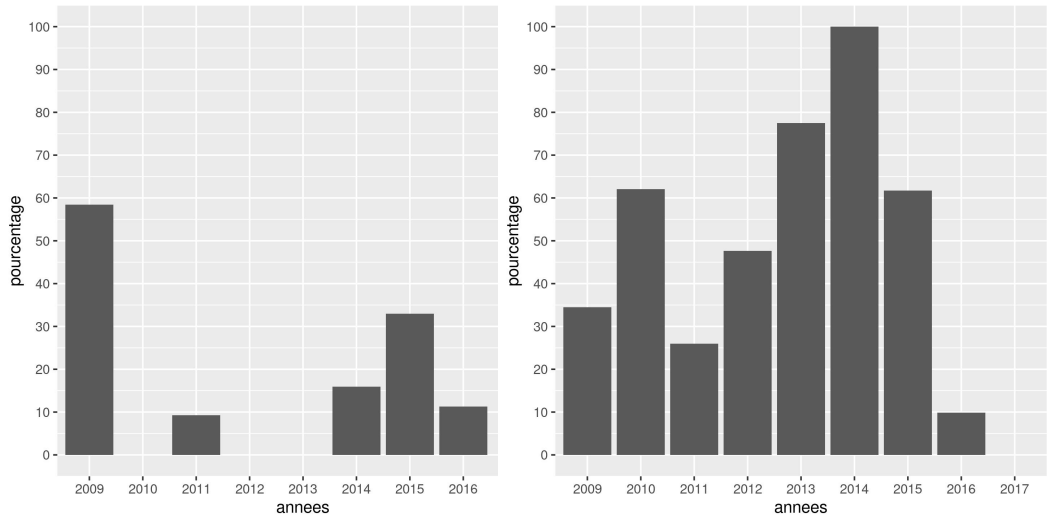
FIGURE 4.2 – Schématisation du protocole d'évaluation de (1) la capacité d'apprentissage et de prédiction et (2) la capacité de généralisation du modèle.

Les bases de données MAREL-Iroise et MesuRho, après imputation des données manquantes par *DTWBI* avec une taille de fenêtre maximum (acceptHole) de 1 mois, sont largement incomplètes. Les figures 4.3 et 4.4 illustrent ce nombre important de données manquantes.



(a) Température (b) Salinité

FIGURE 4.3 – Pourcentage de données manquantes par année dans la base de données MAREL-Iroise de 2000 à 2017 pour (a) la température et (b) la salinité.



(a) Température (b) Salinité

FIGURE 4.4 – Pourcentage de données manquantes par année dans la base de données MesuRho de 2000 à 2017 pour (a) la température et (b) la salinité.

Comme pour *M-SC*, les méthodes de classification non supervisées ne tolèrent pas les valeurs manquantes (*NA*), ainsi un jeu avec trop de données *NA* peut considérablement modifier la distribution du signal et donc, les résultats. C'est pourquoi, nous avons sélectionné différentes périodes, les plus complètes possibles, pour constituer les jeux tests et non le fichier complet

(Tableau 4.2). De plus, nous souhaitions avoir une idée de la capacité de généralisation et de prédiction du modèle sur des années communes aux périodes du jeu d'apprentissage MAREL-Carnot (2005-2009), mais aussi la capacité de prédiction sur une période antécédente et plus longue à celle-ci. Un jeu test sur de nouvelles années permettra d'évaluer le potentiel d'identification de nouveaux évènements et un jeu test plus long permet de multiplier les informations et donc potentiellement une meilleure classification. Par conséquent, nous avons sélectionné pour la base MAREL-Iroise, comme années communes, un premier jeu de données composé de l'année 2006 et comme période différente, un second, allant de janvier 2014 à décembre 2016. Pour la base MesuRho, les jeux tests sont composés pour le premier de l'année commune 2009, puis pour le second des années 2010 à 2014.

4.2.2 Choix des méthodes d'apprentissage

Dans le but de prédire des états environnementaux particuliers, nous avons sélectionnés des méthodes classiques de classification supervisées (Détaillées Section 2.4.3). Ces méthodes créent des modèles de prédiction en se fondant sur des données labellisées. Chaque modèle est une fonction inférée, déterminée à partir d'un objet d'entrée et d'une valeur de sortie souhaitée. Ce modèle de prédiction peut ensuite être utilisé sur de nouveaux exemples non labellisés et ainsi, prédire de nouveaux résultats. Il existe de nombreux modèles et algorithmes d'apprentissage supervisés. Pour notre application, nous avons sélectionné trois algorithmes, et pour chacun, deux configurations différentes ont été choisies :

- L'algorithme des k-plus proches voisins (*k-ppv*), en anglais *k-Nearest Neighbors* (*k-nn*), effectue un calcul de distance pour attribuer la valeur inconnue à son plus proche voisin. Il est possible de définir la méthode de mesure des distances, ici choisie euclidienne, ainsi que le nombre de k voisins, pris égal à $k = 1$ et $k = 7$ dans notre cas. *k-ppv* ne nécessite pas de phase d'apprentissage des données contrairement aux méthodes ci-après qui optimisent leurs paramètres internes en fonction de la base d'apprentissage.
- La méthode de classification Randum Forest (*RF*) [Breiman 1999], permet de construire une « forêt d'arbres de décision » de manière aléatoire. C'est une méthode d'ensemble. Le principe est de construire une multitude de modèles «faibles» puis de les combiner pour créer un modèle robuste. Chaque arbre de la forêt est construit sur une fraction ("*in bag*") des données d'entraînement originales. Elles sont échantillonnées de manière aléatoire, ce qui génère de petits sous-ensembles de données constitués de la nouvelle fraction et du restant ("*out of bag*"). Ces sous-ensembles sont également connus sous le nom d'échantillons bootstrap. Chaque arbre est entraîné sur ces échantillons bootstrap. Tous les arbres de décision constituent l'ensemble de *RF*. Le résultat final du modèle d'ensemble est déterminé par un vote majoritaire de tous les arbres de décision. On peut déterminer le nombre d'arbres (ntree) et la profondeur de ces arbres (mtry), respectivement fixés à ntree=100 et mtry=10 pour la première configuration et à ntree=500 et mtry=20 pour la seconde.
- Les Séparateurs à Vaste Marge (*SVM*) [Vapnik 1998], en anglais, support vector machine, consiste à chercher une règle de décision basée sur une séparation par hyperplan de marge optimale. Cela revient à chercher un hyperplan dont la plus petite distance aux observations est maximale. Tout comme les méthodes spectrales, les *SVM* permettent de transformer l'espace d'entrée en un nouvel espace (Espace Hilbertien). Cette projection dans un espace de dimension plus grand a pour but d'augmenter la probabilité que les données soient davantage linéairement séparables. Elle requiert le calcul d'un produit scalaire dans l'espace initial ou l'espace projeté via une fonction noyau. Ici, nous avons choisi deux méthodes usuels : un noyau linéaire (SVM-linéaire) et un noyau gaussien radial (SVM-radial).

4.2.3 Indices de validation des approches

Pour évaluer à la fois les capacités d'apprentissage et de généralisation, trois indices sont retenus : l'indice de précision, le rappel (Recall) et le score F1 (section 2.4.4.2) pour chaque classe. Chacun de ces indices est défini à partir des valeurs issues du tableau de confusion entre les classes prédites et les labels vrais. Plus ces indices sont proches de 1, plus l'algorithme est performant. Ensuite quatre indices sont calculés pour rendre compte de la qualité globale et non par classe. Les deux premiers indices globaux sont issus de l'analyse de partition : (1) le Rand Index RI et (2) sa version corrigée ARI (section 2.4.4.1). Ces indices permettront d'étudier si la structuration des événements identifiés est proche des événements appris ou si certains événements détectés sont nouveaux. On y ajoute deux indicateurs définis à partir du tableau de confusion entre les classes prédites et les labels vrais déterminés après vote majoritaire : (3) le taux de reconnaissance global (accuracy - acc.) et (4) le nombre de classes bien isolées (#Iso). Une classe sera considérée comme bien isolée si elle est représentée par plus de la moitié des vraies observations positives (*i.e.*, acc.>0,5).

4.3 Évaluation de la capacité d'apprentissage et de généralisation

Cette section présente les résultats de l'apprentissage, puis de la prédiction, réalisés sur la base labellisée MAREL-Carnot (Tests 1 à 5, Tableau 4.2). Ces tests ont pour but d'évaluer la capacité des algorithmes à apprendre et à prédire les données utilisées pour le construire.

4.3.1 Résultats de l'apprentissage

Les tableaux de 4.3 à 4.8, récapitulent pour chaque méthode d'apprentissage, les indices de performance (Precision, Recall, F1, #Iso et acc.) des résultats obtenus sur les jeux d'entraînement des 5 configurations d'apprentissage sur la base de données MAREL-Carnot avec les 8 labels issus de la classification *M-SC* au niveau 3. Ici, c'est donc la capacité d'apprentissage qui est évaluée, la base de test étant identique à la base d'apprentissage.

De manière générale, les 3 systèmes de reconnaissance - *k-ppv*, *SVM*, *RF* - ont une bonne capacité d'apprentissage. Le nombre de classes isolées (colonne #Iso) le plus faible est égal à 5 sur les 8 apprises et le pourcentage de reconnaissance (colonne acc.) est supérieur à 64 % (Tableau 4.7). De plus, tous les scores sont proches de 1 et même égaux à 1 pour la méthode *k-ppv* avec 1 voisin et la méthode *RF* avec 100 et 500 arbres (Tableau 4.9, 4.11, 4.12). Ces trois méthodes sont donc efficaces pour apprendre les données, mais elles peuvent conduire au phénomène de sur-apprentissage ; une adaptation trop restrictive à la géométrie des données conduira à des frontières trop strictes des classes et une difficulté à assigner une nouvelle observation en dehors de ces frontières. Les labels *M-SC* au niveau 3 sont bien séparés et sont déterminés en fonction de la géométrie des données. De plus, elles ont au moins une silhouette supérieure à 0,7 d'après le critère sil.min. Ces caractéristiques semblent simplifier l'apprentissage, ce qui peut expliquer la facilité des méthodes à apprendre. Le label l4 est identifié uniquement en 2005 et 2006 avec un nombre d'observations faible ce qui conduit automatiquement selon la base d'apprentissage choisie à un rappel (Recall) nul et donc une précision et un F1 non calculable (NA dans les tableaux).

Les plus faibles scores des méthodes *SVM* et *k-ppv* à 7 voisins sont liés à l'apprentissage des labels l4 et l5. Par exemple, la méthode *k-ppv* à 7 voisins a des scores assez faibles de rappel (Recall) de l'ordre de 0,3 (Tableau 4.4) et a donc des difficultés à apprendre cette classe. De même, il est difficile pour les méthodes *SVM* linéaires et radiales d'apprendre l5 (Recall

$\pm 0,5$). Elles n'apprennent pas non plus l4 avec des scores de 0 (Tableau 4.4, 4.7 et 4.8). Cela se justifie pleinement par la dynamique temporelle de ces classes. Ces deux labels sont qualifiés d'évènements extrêmes et ont des périodes d'occurrences plus courtes et un nombre d'observations plus faibles, ils sont donc plus complexes à apprendre. Elles sont aussi dominantes sur des années particulières, il sera alors parfois difficile d'obtenir 7 données proches appartenant au label dans la base d'apprentissage. Une idée non testée ici serait d'utiliser le principe d'apprentissage par renforcement : élargir le nombre d'observations de labels sous-représentées. Cependant, au vu de leur caractérisation et du fait que nos entrées sont très limitantes, il semble difficile de construire des données artificielles. Le label l5 étant caractérisé par une chute des nutriments, une perturbation des 5 variables d'entrées n'est pas suffisante et conduirait plutôt à une mauvaise assignation.

Tableau 4.3 – Capacité d'apprentissage de l'algorithme **kppv à 1 voisin** pour les jeux de données d'entraînement **de la base MAREL-Carnot** (2005 à 2009). L'année test n'est pas insérée dans le jeu d'entraînement. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux d'entraînement	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
privé de 2005	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2006	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2007	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2008	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2009	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		

TABLEAU 4.4 – Capacité d'apprentissage de l'algorithme **kppv à 7 voisins** pour les jeux de données d'entraînement **de la base MAREL-Carnot** (2005 à 2009). L'année test n'est pas insérée dans le jeu d'entraînement. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux d'entraînement	indices	l1	cl2	l3	l4	l5	l6	l7	l8	#Iso	acc.
privé de 2005	Precision	0,84	0,80	0,80	0,91	0,73	0,89	0,90	0,84	7	0,85
	Recall	0,95	0,73	0,73	0,76	0,31	0,95	0,87	0,61		
	F1	0,89	0,76	0,76	0,83	0,44	0,92	0,89	0,71		
privé de 2006	Precision	0,82	0,83	0,80	NA	0,77	0,90	0,88	0,84	5	0,85
	Recall	0,96	0,71	0,81	0,00	0,34	0,96	0,83	0,43		
	F1	0,89	0,76	0,80	NA	0,47	0,93	0,85	0,57		
privé de 2007	Precision	0,81	0,80	0,81	0,91	0,76	0,87	0,90	0,82	7	0,83
	Recall	0,94	0,71	0,86	0,66	0,34	0,93	0,84	0,60		
	F1	0,87	0,75	0,83	0,77	0,47	0,90	0,87	0,70		
privé de 2008	Precision	0,84	0,81	0,81	0,91	0,79	0,87	0,88	0,82	7	0,84
	Recall	0,96	0,67	0,84	0,66	0,35	0,93	0,84	0,63		
	F1	0,90	0,73	0,83	0,77	0,49	0,90	0,86	0,71		
privé de 2009	Precision	0,81	0,84	0,80	0,91	0,77	0,89	0,91	0,87	7	0,85
	Recall	0,96	0,70	0,85	0,66	0,36	0,95	0,86	0,64		
	F1	0,88	0,76	0,82	0,77	0,49	0,92	0,88	0,74		

TABLEAU 4.5 – Capacité d'apprentissage de l'algorithme **RF à 100 arbres et 10 niveaux de profondeur** pour les jeux de données d'entraînement **de la base MAREL-Carnot** (2005 à 2009). L'année test n'est pas insérée dans le jeu d'entraînement. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux d'entraînement	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
privé de 2005	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2006	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2007	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2008	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2009	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		

Tableau 4.6 – Capacité d'apprentissage de l'algorithme **RF à 500 arbres et 20 niveaux** de profondeur pour les jeux de données d'entraînement **de la base MAREL-Carnot** (2005 à 2009). L'année test n'est pas insérée dans le jeu d'entraînement. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux d'entraînement	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
privé de 2005	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2006	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2007	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2008	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		
privé de 2009	Precision	1	1	1	1	1	1	1	1	8	1
	Recall	1	1	1	1	1	1	1	1		
	F1	1	1	1	1	1	1	1	1		

Tableau 4.7 – Capacité d'apprentissage de l'algorithme **SVM linéaire** pour les jeux de données d'entraînement **de la base MAREL-Carnot** (2005 à 2009). L'année test n'est pas insérée dans le jeu d'entraînement. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux d'entraînement	indices	l1	l2	l3	l4	l5	l6	cl7	l8	#Iso	acc.
privé de 2005	Precision	0,82	0,73	0,78	NA	0,10	0,83	0,82	0,78	5	0,79
	Recall	0,90	0,67	0,65	0,00	0,04	0,95	0,78	0,41		
	F1	0,86	0,70	0,71	NA	0,06	0,89	0,80	0,54		
privé de 2006	Precision	0,83	0,65	0,63	NA	0,13	0,82	0,77	0,77	5	0,75
	Recall	0,84	0,65	0,75	0,00	0,04	0,96	0,53	0,43		
	F1	0,83	0,65	0,68	NA	0,07	0,89	0,63	0,55		
privé de 2007	Precision	0,73	0,62	0,76	NA	0,31	0,83	0,81	0,47	5	0,72
	Recall	0,84	0,58	0,68	0,00	0,05	0,80	0,77	0,88		
	F1	0,78	0,60	0,72	NA	0,08	0,82	0,79	0,62		
privé de 2008	Precision	0,76	0,36	0,75	NA	0,12	0,78	0,87	0,79	5	0,64
	Recall	0,38	0,66	0,57	0,00	0,18	0,95	0,73	0,56		
	F1	0,51	0,46	0,65	NA	0,14	0,86	0,80	0,65		
privé de 2009	Precision	0,81	0,71	0,75	NA	0,14	0,83	0,74	0,38	5	0,77
	Recall	0,89	0,68	0,66	0,00	0,04	0,95	0,80	0,10		
	F1	0,85	0,70	0,70	NA	0,07	0,89	0,77	0,16		

TABLEAU 4.8 – Capacité d'apprentissage de l'algorithme **SVM non linéaire** pour les jeux de données d'entraînement **de la base MAREL-Carnot** (2005 à 2009). L'année test n'est pas insérée dans le jeu d'entraînement. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux d'entraînement	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
privé de 2005	Precision	0,85	0,87	0,91	1,00	0,95	0,90	0,95	0,89	7	0,88
	Recall	0,97	0,75	0,89	0,64	0,31	0,96	0,93	0,69		
	F1	0,91	0,80	0,90	0,78	0,46	0,93	0,94	0,78		
privé de 2006	Precision	0,84	0,89	0,89	NA	0,93	0,92	0,92	0,93	6	0,88
	Recall	0,98	0,74	0,92	0,00	0,36	0,97	0,90	0,61		
	F1	0,90	0,81	0,90	NA	0,52	0,94	0,91	0,74		
privé de 2007	Precision	0,83	0,87	0,89	1,00	0,94	0,90	0,94	0,89	7	0,87
	Recall	0,97	0,75	0,94	0,55	0,36	0,94	0,93	0,69		
	F1	0,89	0,81	0,91	0,71	0,52	0,92	0,93	0,77		
2008	Precision	0,86	0,90	0,88	1,00	0,92	0,89	0,94	0,88	7	0,88
	Recall	0,98	0,70	0,94	0,55	0,29	0,95	0,91	0,69		
	F1	0,92	0,79	0,91	0,71	0,44	0,92	0,92	0,77		
privé de 2009	Precision	0,84	0,90	0,88	1,00	0,92	0,91	0,95	0,91	8	0,89
	Recall	0,98	0,76	0,93	0,55	0,36	0,96	0,92	0,79		
	F1	0,91	0,82	0,91	0,71	0,51	0,94	0,93	0,85		

Les résultats des 3 systèmes de reconnaissance utilisés permettent de démontrer qu'il est possible d'apprendre l'ensemble des états labellisés à partir d'un jeu de variables réduit. Il est toutefois plus difficile mais possible de reconnaître les événements extrêmes dont les paramètres nutriments ne sont pas insérés dans l'agent de prédiction.

4.3.2 Résultats de la généralisation sur la base MAREL-Carnot

Les tableaux de 4.9 à 4.11, récapitulent pour chaque technique utilisée les indices de performance en généralisation pour les configurations 1 à 5 de la base MAREL-Carnot avec les 8 labels issus de la classification *M-SC* au niveau 3.

Les résultats obtenus en test sont moins bons par rapport aux résultats d'apprentissage, mais la prédiction est tout de même satisfaisante. En effet, le taux de reconnaissance (acc.) est supérieur à 0,55 quelle que soit la méthode. Comme pour les résultats sur le jeu d'entraînement, les labels l4 et l5 sont les plus difficiles à prédire. Ce sont des labels présents seulement en 2005 et 2006, les scores sont donc légèrement plus forts pour les années tests 2007-2008-2009, le jeu d'entraînement comprenant plus d'exemples de l4 et l5. Par exemple, la méthode *RF* à 100 arbres a un rappel de 0,84 pour 2008 contre 0,01 pour 2006 (Tableau 4.11). Toutefois même si les indices sont forts pour les dernières années, le nombre de classes bien isolées est plus faible. Seulement 1 label est isolé par *k-ppv* à 1 voisin pour les années 2008 et 2009 (Tableau 4.9) et 2 labels par *RF* à 100 arbres (Tableau 4.11). Pour ces années, la confusion est donc plus importante. Les années 2005 et 2006 sont des années particulières avec un taux de fluorescence faible enregistré pour 2005 et fort pour 2006 en comparaison aux années 2007, 2008 et 2009. Les années 2005 et 2006 sont donc moins représentatives et peuvent rendre la prédiction plus complexe. Le label l8, au contraire, est mal prédit sur 2005, 2006, et bien prédit sur le reste des années. Ces résultats semblent cohérents, puisque l8 a très peu d'occurrences en 2005 et 2006 (Figure 4.5). De plus, les labels printaniers l3 et l7 et le label hivernal l6 sont bien prédits, quelle que soit la méthode , *i.e.* Recall $> 0,5$ pour l3 pour *k-ppv* à 7 voisins (Tableau 4.10).

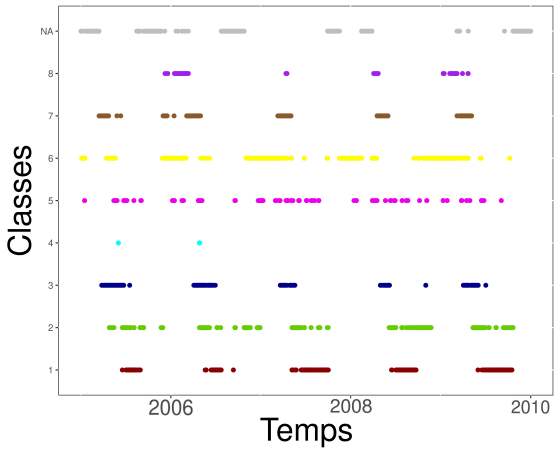


FIGURE 4.5 – Dynamique temporelle de chaque label au niveau trois de classification M-SC pour la station MAREL-CARNOT sur la période 2005-2009

Cette étude sur la base MAREL-Carnot a démontré la faisabilité de proposer un système de reconnaissance des événements phytoplanctoniques pour des observations non apprises (1) sur un même site et (2) sur un ensemble restreint d'EOVs.

Il est important de souligner que nous avons fait le choix d'explorer des solutions simples, rapides en temps de mise en œuvre et calculatoire, et ne nécessitant pas un grand nombre d'observations pour aboutir à des résultats satisfaisants. Les approches neuronales peu profondes approchent les mêmes performances et la quantité actuelle d'observations ne permettent pas de paramétrer simplement les techniques d'apprentissage profond (deep learning).

Pour la suite de ce chapitre, nous avons fait le choix de ne présenter que les résultats sur le classifieur *k-ppv* à 1 voisin. Un $k > 1$ (*i.e.* 7) nécessite un *a priori* fort sur le nombre d'observations appartenant à un même événement, i.e. événement extrême, isolé ou rare (nombre d'observations $< k/2$).

TABLEAU 4.9 – Capacité de généralisation de l'algorithme **kppv à 1 voisin** pour les jeux de données test **de la base MAREL-Carnot** (2005 à 2009). L'année test correspond à l'année utilisée pour le jeu test. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de rappel, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux de test	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
2005	Precision	0,55	0,22	0,75	0,00	0,02	0,62	0,53	0,01	4	0,55
	Recall	0,72	0,19	0,66	0,00	0,01	0,83	0,36	0,12		
	F1	0,62	0,20	0,70	NA	0,01	0,71	0,43	0,02		
2006	Precision	0,70	0,38	0,72	NA	0,02	0,65	0,70	0,55	5	0,59
	Recall	0,70	0,53	0,53	0,00	0,03	0,72	0,91	0,16		
	F1	0,70	0,44	0,61	NA	0,02	0,68	0,79	0,25		
2007	Precision	0,87	0,38	0,32	0,00	0,06	0,87	0,73	0,02	3	0,68
	Recall	0,69	0,66	0,58	NA	0,05	0,79	0,63	0,02		
	F1	0,77	0,48	0,41	NA	0,06	0,83	0,67	0,02		
2008	Precision	0,62	0,72	0,27	0,00	0,07	0,89	0,66	0,04	1	0,65
	Recall	0,87	0,39	0,79	NA	0,04	0,80	0,67	0,06		
	F1	0,72	0,51	0,40	NA	0,05	0,84	0,67	0,05		
2009	Precision	0,74	0,44	0,66	0,00	0,02	0,69	0,72	0,20	1	0,58
	Recall	0,73	0,43	0,47	NA	0,07	0,58	0,60	0,43		
	F1	0,73	0,44	0,55	NA	0,03	0,63	0,65	0,28		

TABLEAU 4.10 – Capacité de généralisation de l'algorithme **kppv à 7 voisin** pour les jeux de données test **de la base MAREL-Carnot** (2005 à 2009). L'année test correspond à l'année utilisée pour le jeu test. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux de test	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
2005	Precision	0,57	0,28	0,76	NA	0,02	0,62	0,59	0,01	3	0,59
	Recall	0,90	0,16	0,68	0,00	0,00	0,87	0,38	0,08		
	F1	0,70	0,21	0,72	NA	0,01	0,72	0,46	0,01		
2006	Precision	0,69	0,40	0,73	NA	0,02	0,62	0,71	0,49	5	0,62
	Recall	0,83	0,53	0,52	0,00	0,01	0,74	0,92	0,08		
	F1	0,75	0,45	0,61	NA	0,01	0,68	0,80	0,13		
2007	Precision	0,89	0,46	0,35	0,05	0,87	0,79	0,03	0,89	4	0,75
	Recall	0,80	0,66	0,66	0,01	0,87	0,64	0,02	0,80		
	F1	0,84	0,54	0,46	0,02	0,87	0,71	0,02	0,84		
2008	Precision	0,61	0,82	0,28	0,00	0,04	0,87	0,70	0,05	1	0,67
	Recall	0,95	0,37	0,81	NA	0,01	0,86	0,60	0,04		
	F1	0,74	0,51	0,42	NA	0,01	0,86	0,65	0,04		
2009	Precision	0,75	0,53	0,60	0,02	0,71	0,73	0,23	0,75	2	0,64
	Recall	0,86	0,42	0,50	0,03	0,65	0,59	0,41	0,86		
	F1	0,80	0,47	0,55	0,03	0,68	0,66	0,30	0,80		

TABLEAU 4.11 – Capacité de généralisation de l'algorithme **RF à 100 arbres et 10 niveaux de profondeur** pour les jeux de données test **de la base MAREL-Carnot** (2005 à 2009). L'année test correspond à l'année utilisée pour le jeu test. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits. Les cellules grisées sont une aide à la lecture et représentent les chiffres cités comme exemple dans les résultats.

Jeux de test	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
2005	Precision	0,55	0,28	0,83	NA	0,17	0,65	0,62	0,00	4	0,59
	Recall	0,70	0,27	0,70	0,00	0,10	0,77	0,53	0,00		
	F1	0,62	0,28	0,76	NA	0,12	0,71	0,57	NA		
2006	Precision	0,78	0,46	0,74	NA	0,01	0,69	0,80	0,65	5	0,67
	Recall	0,64	0,66	0,76	0,00	0,01	0,82	0,81	0,22		
	F1	0,71	0,54	0,75	NA	0,01	0,75	0,81	0,33		
2007	Precision	0,91	0,46	0,36	0,07	0,90	0,71	0,00	0,91	3	0,73
	Recall	0,75	0,77	0,74	0,04	0,79	0,73	0,00	0,75		
	F1	0,82	0,58	0,49	0,05	0,84	0,72	NA	0,82		
2008	Precision	0,67	0,83	0,23	0,31	0,94	0,63	0,07	0,67	2	0,69
	Recall	0,97	0,43	0,79	0,14	0,84	0,62	0,01	0,97		
	F1	0,79	0,57	0,36	0,19	0,89	0,63	0,01	0,79		
2009	Precision	0,73	0,45	0,70	0,01	0,70	0,73	0,22	0,73	2	0,61
	Recall	0,74	0,43	0,65	0,01	0,58	0,67	0,48	0,74		
	F1	0,74	0,44	0,67	0,01	0,63	0,70	0,31	0,74		

Tableau 4.12 – Capacité de généralisation de l'algorithme **RF à 500 arbres et 20 niveaux de profondeur** pour les jeux de données test **de la base MAREL-Carnot** (2005 à 2009). L'année test correspond à l'année utilisée pour le jeu test. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits.

Jeux de test	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
2005	Precision	0,55	0,27	0,83	NA	0,17	0,66	0,62	0,00	4	0,59
	Recall	0,70	0,26	0,69	0,00	0,10	0,76	0,55	0,00,		
	F1	0,62	0,27	0,75	NA	0,13	0,71	0,58	NA		
2006	Precision	0,80	0,45	0,73	NA	0,01	0,69	0,80	0,63	5	0,66
	Recall	0,65	0,67	0,76	0,00	0,01	0,81	0,78	0,25		
	F1	0,71	0,54	0,74	NA	0,01	0,74	0,79	0,35		
2007	Precision	0,91	0,46	0,37	0,00	0,08	0,90	0,71	0,91	3	0,74
	Recall	0,74	0,77	0,74	NA	0,04	0,80	0,73	0,74		
	F1	0,82	0,58	0,49	NA	0,06	0,84	0,72	0,82		
2008	Precision	0,67	0,83	0,24	0,33	0,93	0,64	0,10	0,67	2	0,70
	Recall	0,97	0,43	0,80	0,13	0,86	0,64	0,01	0,97		
	F1	0,79	0,57	0,37	0,19	0,89	0,64	0,02	0,79		
2009	Precision	0,73	0,44	0,69	0,01	0,70	0,74	0,22	0,73	2	0,60
	Recall	0,74	0,42	0,66	0,01	0,58	0,67	0,46	0,74		
	F1	0,73	0,43	0,68	0,01	0,63	0,70	0,29	0,73		

Tableau 4.13 – Capacité de généralisation de l'algorithme **SVM linéraire** pour les jeux de données test **de la base MAREL-Carnot** (2005 à 2009). L'année test correspond à l'année utilisée pour le jeu test. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits.

Jeux de test	indices	l1	l2	l3	l4	l5	l6	l7	l8	#Iso	acc.
2005	Precision	0,64	0,68	0,79	NA	0,00	0,67	0,77	0,00	4	0,68
	Recall	0,95	0,19	0,78	0,00	0,00	0,96	0,56	0,00		
	F1	0,76	0,30	0,78	NA	NA	0,79	0,65	NA		
2006	Precision	0,77	0,47	0,67	NA	0,00	0,68	0,85	0,78	4	0,64
	Recall	0,57	0,40	0,72	0,00	0,01	0,96	0,68	0,22		
	F1	0,65	0,43	0,70	NA	0,00	0,79	0,75	0,34		
2007	Precision	0,92	0,61	0,41	0,26	0,88	0,80	NA	0,92	4	0,81
	Recall	0,80	0,83	0,56	0,03	0,97	0,69	0,00	0,80		
	F1	0,85	0,70	0,48	0,05	0,92	0,74	NA	0,85		
2008	Precision	0,62	0,56	0,42	0,05	0,91	0,76	0,00	0,62	3	0,68
	Recall	0,50	0,64	0,42	0,05	0,94	0,81	0,00	0,50		
	F1	0,55	0,60	0,42	0,05	0,92	0,79	NA	0,55		
2009	Precision	0,83	0,46	0,87	0,00	0,70	0,78	0,46	0,83	3	0,66
	Recall	0,61	0,75	0,48	0,00	0,94	0,54	0,20	0,61		
	F1	0,70	0,57	0,62	NA	0,80	0,64	0,28	0,70		

TABLEAU 4.14 – Capacité de généralisation de l'algorithme **SVM non linéaire** pour les jeux de données test **de la base MAREL-Carnot** (2005 à 2009). L'année test correspond à l'année utilisée pour le jeu test. Pour chaque jeu d'entraînement, les indices de performance : la précision, le score de Recall, le score F1, le nombre de classes bien isolées (#Iso) et l'accuracy (acc.) sont décrits.

Jeux de test	indices	l1	l2	l3	l4	l5	l6	l7	l8	Iso	acc.
2005	Precision	0,59	0,56	0,82	NA	0,00	0,65	0,67	0,00	4	0,65
	Recall	0,99	0,17	0,74	0,00	0,00	0,93	0,43	0,08		
	F1	0,74	0,27	0,77	NA	NA	0,76	0,53	0,01		
2006	Precision	0,41	0,42	0,77	NA	0,03	0,64	0,52	0,74	4	0,55
	Recall	0,85	0,53	0,63	0,00	0,01	0,78	0,27	0,07		
	F1	0,55	0,47	0,69	NA	0,01	0,70	0,36	0,12		
2007	Precision	0,90	0,56	0,44	0,32	0,89	0,65	0,00	0,90	4	0,78
	Recall	0,83	0,71	0,66	0,01	0,89	0,73	0,00	0,83		
	F1	0,87	0,63	0,53	0,02	0,89	0,69	NA	0,87		
2008	Precision	0,58	0,87	0,35	0,33	0,90	0,69	0,00	0,58	1	0,70
	Recall	0,98	0,43	0,80	0,01	0,87	0,65	0,00	0,98		
	F1	0,73	0,58	0,48	0,01	0,88	0,67	NA	0,73		
2009	Precision	0,76	0,61	0,74	0,01	0,67	0,71	0,23	0,76	3	0,75
	Recall	0,87	0,45	0,55	0,00	0,65	0,63	0,39	0,87		
	F1	0,81	0,52	0,63	0,00	0,66	0,67	0,29	0,81		

4.4 Évaluation de la capacité de généralisation sur d'autres zones d'études

Dans cette section, la capacité de généralisation du modèle construit à partir de la base labellisée de MAREL-Carnot est testée sur d'autres jeux de données, ici MAREL-Iroise et MesuRho (configurations 6 à 9, Tableau 4.2). Ces résultats vont permettre, dans un premier temps, d'évaluer la capacité à labelliser des données acquises sur un autre site, dans un second temps, de mettre en évidence les points communs et les différences entre les labels de la base MAREL-Carnot et les classes de MAREL-Iroise et MesuRho.

Dans le chapitre 1 (Section 1.4), nous avons émis l'hypothèse que les trois stations, bien que leurs répartitions géographiques soient différentes et qu'elles effectuent des mesures dans des écosystèmes différents, des schémas communs pourront être identifiés par rapport à une dynamique des systèmes tempérées. En effet, les stratégies d'implantations sont telles que les conditions environnementales sont en moyenne assez proches, ce qui doit permettre une inter-comparaison. À des échelles plus fines, en revanche une certaine variabilité de ces schémas, qui fait la particularité de chaque site, pourrait être observée.

La caractérisation des états environnementaux de MAREL-Iroise et MesuRho à partir des états définis selon MAREL-Carnot a donc un sens naturel pour les schémas généraux (niveau *M-SC* 1 et 2) que nous devrions retrouver. Les labels définis pour MAREL-Carnot rendent compte d'une dynamique saisonnière globale caractéristique des zones tempérées (Section 1.2.3) pour les trois stations. Au niveau supérieur vers les événements extrêmes, la démarche est exploratoire et nous ne sommes pas certains d'observer les mêmes types événements.

Pour cette analyse, les données de MAREL-Iroise et MesuRho ont été classées par la méthode *M-SC*. Ces classes feront office de vérité (*true label*), mais il est important de rappeler que ce sont des classes qui rendent compte de la géométrie des données et non des états environnementaux comme pour MAREL-Carnot. La capacité de reconnaissance est évaluée à partir de 2 critères (1) un critère temporel, en observant la répartition des classes par mois et sur le signal de fluorescence et (2) des critères de comparaison de la partition non supervisée (cl_i) et celle supervisée (l_i) : le Rand index (RI) et l'Adjusted Rand Index (ARI). La répartition des classes, les tableaux de

contingences entre les classes des stations tests (MAREL-Iroise ou MesuRho) (cl_i) et les labels prédits (l_i) seront également commentés.

4.4.1 Prédictions sur MAREL-Iroise pour 2006

Structuration des classes

La structure des événements de la base test MAREL-Iroise peut différer de celle de la base d'apprentissage MAREL-Carnot en raison, d'une part, de la base de données elle-même (variables EOVs différentes), et d'autre part, du fait des différences de conditions environnementales entre les sites. Les tableaux 4.15, 4.16 et 4.17, permettent de quantifier le niveau de similitude de structure entre les observations de la base MAREL-Iroise, définies par les classes non labellisées par un expert scientifique (Classe non exp. cl_i) et les labels prédits (l_i) par le modèle de prédiction construit à partir des observations et des labels de MAREL-Carnot. Le score RI, qui représente le taux de similarité entre les classes, est supérieur à 60 % quel que soit le niveau d'interprétation (niveau *M-SC*), et même s'il est faible, le taux ARI est positif ($> 0,2$). Ces deux indices démontrent qu'il est pertinent de proposer une première labellisation de cette base à partir de MAREL-Carnot et que le partitionnement obtenu n'est pas aléatoire. Nous étudions ensuite la capacité de notre système à proposer les schémas de fonctionnements. Au niveau *M-SC* 1, plus de 80 %, soit 10 726/(10726+2 313)(Tableau 4.15), des occurrences de la classe cl2 ont été reconnues comme appartenant à la saison froide (l2 label prédit) et la moitié des observations de cl1 à la saison chaude (label l1). La température est un paramètre structurant du niveau 1 et en effet, on constate une différence de variation de température entre les deux sites, ce qui permet d'expliquer ce chevauchement. La température mesurée à MAREL-Carnot varie entre 0 et 20 °C et celle de MAREL-Iroise entre 7 et 17 °C pour 2006. 60 % des appariements d'observations issus de *M-SC* sont en concordance avec les labels assignés.

Tableau 4.15 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 1** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MAREL-Iroise 2006 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	RI	ARI
cl1	5477	**5164**	0,60	0,20
cl2	**10726**	2313		

Pour le niveau *M-SC* 2, une certaine dynamique commune semble ressortir. Par exemple, cl2 non exp. pourrait être qualifiée d'état correspondant à une efflorescence printanière. En effet, 81 % (3561 et 3492, Tableau 4.16) des occurrences de cl2 sont assignées majoritairement dans l2 et l4, classes toutes deux identifiées sur la période de mars à juin et labellisées comme efflorescence printanière. Ensuite, les observations liées à cl4 sont assignées principalement au label l1 qui correspond à une situation intermédiaire de post-bloom entre juin et septembre. Avec deux tiers de ces occurrences (202 contre 14, 95 et 0, Tableau 4.16) affectés à l2, cl4 semble être un épisode de diminution des sels nutritifs dans le milieu, l2 étant structurée par des chutes de nutriment et en particulier de silicate (Section 3.4.2 chapitre 3 , Tableau 3.10). Mais cette conclusion reste à confirmer puisque les nutriments ne font pas partie des données d'entrée et ne peuvent pas ici être vérifiés. Ainsi, une pré-labellisation de classes est possible par le biais du modèle de prédiction, même si le nombre de variables d'entrée est moins importante.

TABLEAU 4.16 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 2** (Classes non exp. cl_i) et les labels prédits (l_i) prédits obtenus par Kppv à 1 voisin sur le jeu de données MAREL-Iroise 2006 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Labels prédits	l1	l2	l3	l4	RI	ARI
cl1	443	**1294**	91	118	0,62	0,21
cl2	191	**3561**	1454	**3492**		
cl3	**6638**	3870	2192	21		
cl4	14	**202**	95	0		

Enfin pour le niveau 3 *M-SC*, la confusion est plus importante (TableauN4.17). Seule une distinction entre les principales phases de bloom est possible. En effet, les labels des évènements extrêmes l4 et l5 ne correspondent à aucune des classes identifiées par *M-SC* sur MAREL-Iroise. Un quart des occurrences de cl5 est affecté à l5 et le reste est confondu avec l1, l2, l3. Ceci pourrait s'expliquer par une classe cl5 encore non homogène où un niveau plus profond *M-SC* pourrait être utile à la segmentation. Le niveau d'arrêt *M-SC* a été figé par absence d'information fine. En effet, nous ne disposons pas des labels expertisés pour les niveaux plus profonds de MAREL-Carnot. L'expertise a été faite au niveau 3 pour rester en accord avec le niveau de précision que pouvaient nous fournir les données complémentaires telles que le dénombrement taxonomique. En ce qui concerne les similitudes à ce niveau, la phase de faible productivité (l1) et la phase de transition entre la période estivale et printanière (l8) ne semblent pas isolées dans une classe non expertisée cl_i. l8 et même quasi absent, seulement 9 occurrences sont en correspondance avec l8 (Tableau 4.17). Il ressort majoritairement des classes proches de l7 et l3 les deux labels printaniers et de l2 plutôt qualifiées de bloom automnal.

TABLEAU 4.17 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 3** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MAREL-Iroise 2006 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

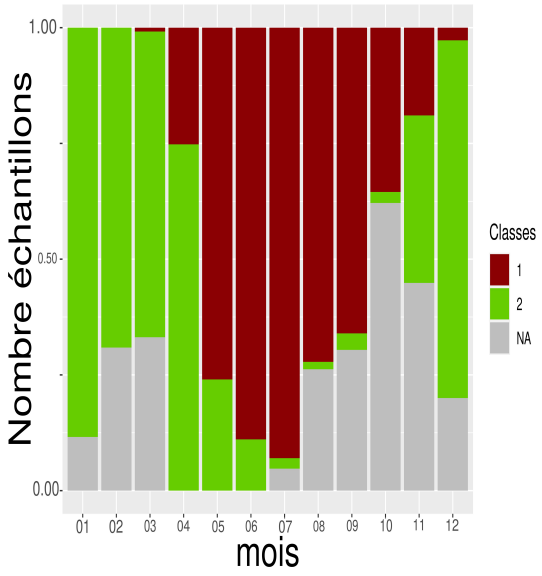
Classes non exp. vs Label prédits	l1	l2	l3	l4	l5	l6	l7	l8	RI	ARI
cl1	0	39	**787**	0	50	15	99	3	0,63	0,13
cl2	234	169	**488**	0	21	3	22	0		
cl3	9	167	1303	0	68	420	**4284**	0		
cl4	0	20	2267	5	14	954	**3060**	5		
cl5	3107	**3493**	**3862**	2	2178	10	21	1		
cl6	9	**24**	10	0	5	0	0	0		
cl7	0	14	**203**	0	0	95	0	0		

Répartition temporelle des labels prédits

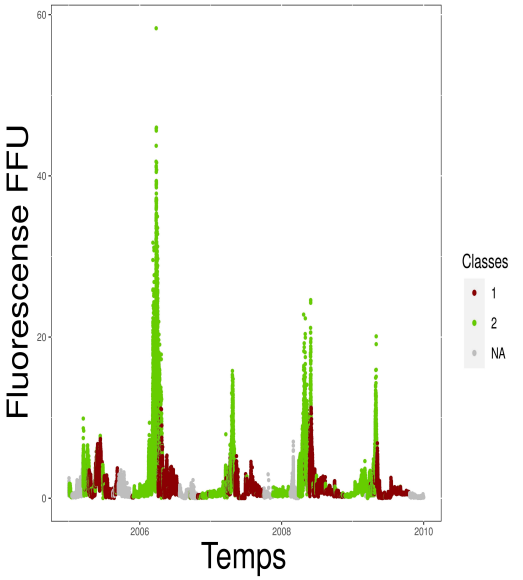
Les résultats de classification obtenus pour MAREL-Carnot entre 2006 et 2009 sont comparés aux labels prédits sur MAREL-Iroise pour l'année 2006. Les figures 4.6, 4.7, 4.8 illustrent pour les 3 niveaux de classifications, en a) la distribution par mois des classes labellisées par l'algorithme de *clustering* *M-SC* pour la base de données MAREL-Carnot, en b) leurs distributions sur le signal de fluorescence à MAREL-Carnot, en c) la distribution par mois des labels prédits pour la base de données test, ici MAREL-Iroise en 2006 et en d) la distribution des labels prédits sur le signal de fluorescences à MAREL-Iroise pour l'année 2006.

Pour cette station, nous retrouvons une dynamique proche de celle des classes de MAREL-Carnot (décrite au chapitre 3, section 3.4.3). Ainsi, au niveau 1 (figure 4.6) une dynamique

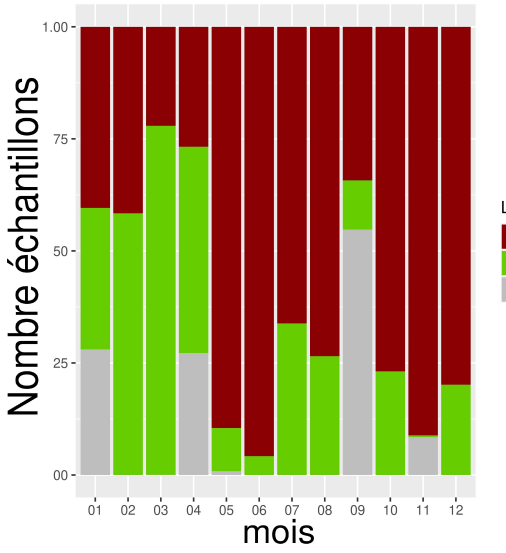
saisonnière avec une séparation entre une période printanière/estivale et une période automnale/-hivernale est bien observée. Elle est toutefois moins marquée que pour MAREL-Carnot. Une des raisons pouvant expliquer cette différence est que le jeu de données MAREL-Iroise contient un nombre important de données manquantes (*NA*) sur la période printanière, la période est donc moins bien prédite et moins bien représentée. De plus, ce sont les labels prédits qui sont présentées figure 4.6c et 4.7d et non les classes issues de *M-SC* comme pour la figure 4.6a et 4.6b, il y a donc plus de confusions. Il en est de même pour les niveaux 2 (Figure 4.7) et 3 (Figure 4.8), où les labels prédits ont sensiblement la même dynamique que les classes MAREL-Carnot. Il est toutefois possible d'identifier quelques différences plus importantes, notamment au niveau 3. Par exemple, l6 est bien un état hivernal comme pour MAREL-Carnot mais semble être défini principalement sur les périodes de janvier à mars et non sur la totalité de la période (novembre à mars) (Figure 4.8a et 4.8c). Les évènements extrêmes l4 et l5 sont peu représentés et ont des périodes différentes, mais ces labels font partie des plus difficiles à apprendre de part leur caractère extrême. De plus leur dynamique plus courte et ponctuelle n'est pas contradictoire avec une période d'occurrences différente. Il est en effet imaginable que ces évènements qualifiés d'évènements extrêmes à la station MAREL-Carnot soientt plus récurrents ou se produisent à des périodes différentes à la station MAREL-Iroise. Par exemple l5 est relatif à une augmentation des nutriments dans le milieu. Un événement de ce type peut être lié à différents apports anthropiques et peut donc avoir une périodicité différente en fonction de la zone d'étude. Les deux labels printaniers l3 et l7 sont bien présents, mais il semble que la l7 soit plus précoce et débute en janvier au lieux de mars et que l3 soit plus étendu sur l'année. Néanmoins, l7, est fortement structuré par le signal de fluorescence, or, lors de l'apprentissage, le signal de fluorescence n'est pas en entrée du modèle de prédiction, ce qui peut expliquer cette différence.



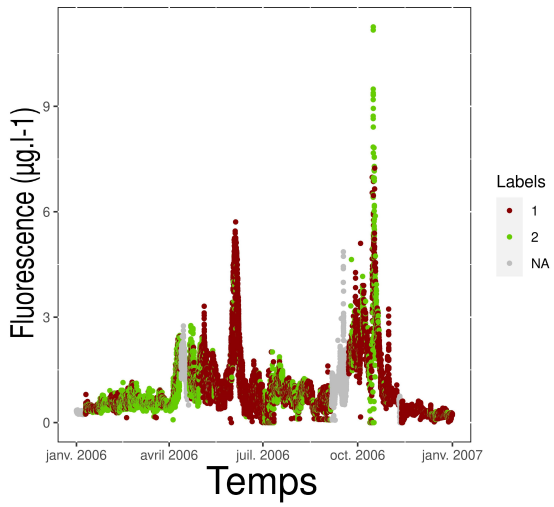
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot

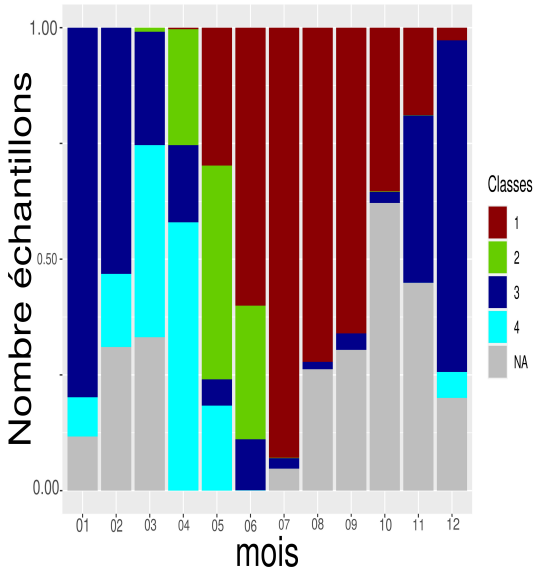


(c) Distribution par mois des labels prédits pour MAREL-Iroise

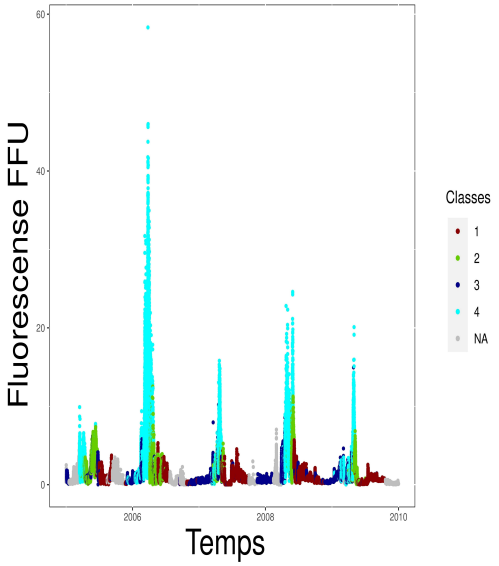


(d) Distribution des labels prédits sur le signal de fluorescence ($\mu g\, l^{-1}$) MAREL-Iroise

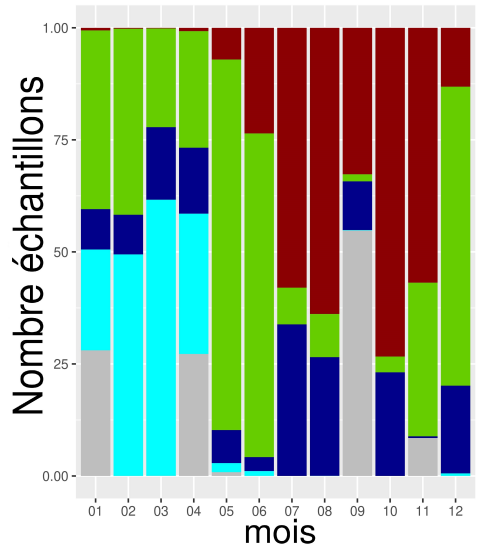
FIGURE 4.6 – Comparaison **au niveau 1** de la dynamique temporelle entre les classes M-SC de **MAREL-Carnot de 2005 à 2009** et les labels prédits pour **MAREL-Iroise sur la période 2006**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise. La répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot sur 2005-2009, d) labels prédits pour MAREL-Iroise sur 2006.



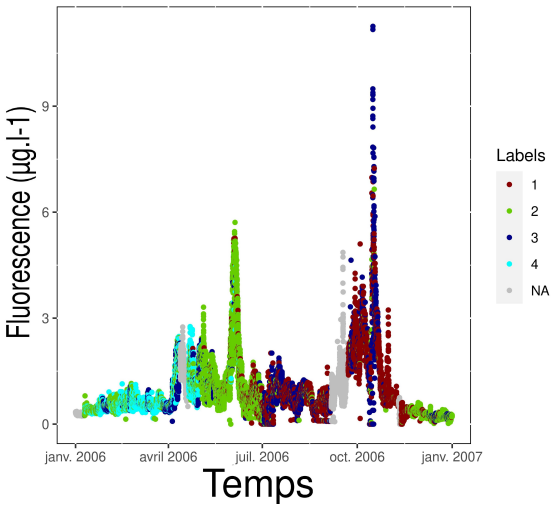
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot

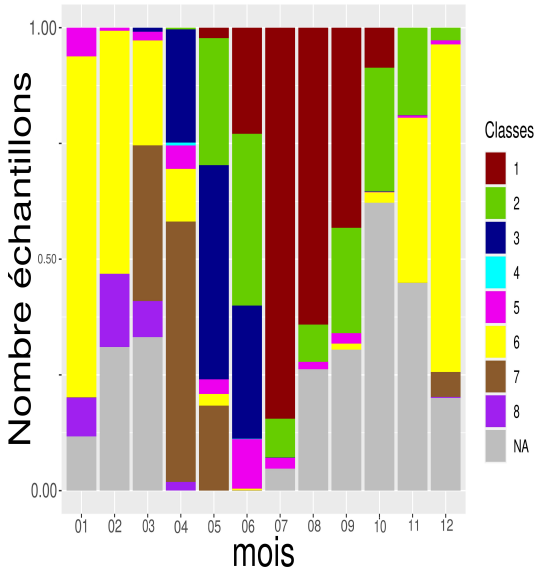


(c) Distribution par mois des labels prédits pour MAREL-Iroise

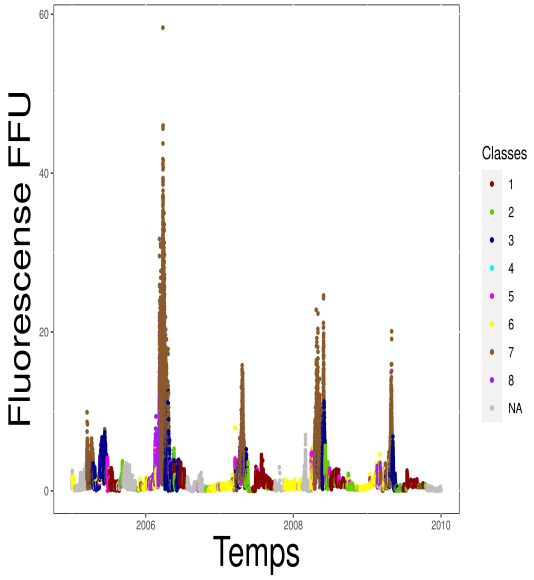


(d) Distribution des labels prédits sur le signal de fluorescence ($\mu g\,l^{-1}$) MAREL-Iroise

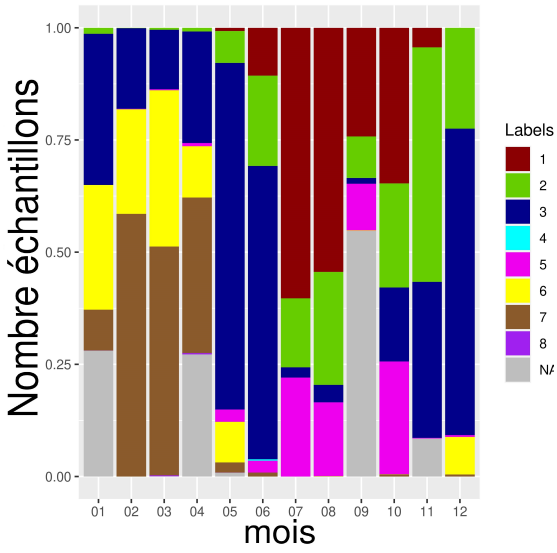
FIGURE 4.7 – Comparaison **au niveau 2** de la dynamique temporelle entre les classes M-SC de **MAREL-Carnot de 2005 à 2009** et les labels prédits pour **MAREL-Iroise sur la période 2006**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise. La répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot sur 2005-2009, d) labels prédits pour MAREL-Iroise sur 2006.



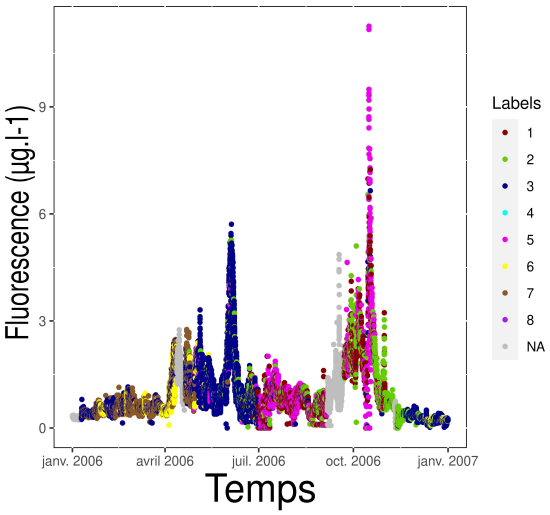
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot



(c) Distribution par mois des labels prédits pour MAREL-Iroise



(d) Distribution des labels prédits sur le signal de fluorescence ($\mu g\,l^{-1}$) MAREL-Iroise

FIGURE 4.8 – Comparaison **au niveau 3** de la dynamique temporelle entre les classes M-SC de **MAREL-Carnot de 2005 à 2009** et les labels prédits pour **MAREL-Iroise sur la période 2006**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise. La répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot sur 2005-2009, d) labels prédits pour MAREL-Iroise sur 2006.

4.4.2 Prédictions sur MAREL-Iroise de 2014 à 2016

Pour la période 2014-2016, le signal de fluorescence en unité FFU est disponible, il est donc ajouté en entrée du modèle de prédiction. Dans ce cas, le modèle appris n'est plus le même et les résultats de reconnaissance sont donc différents. Toutefois, le modèle étant stable, les différences se font pour les niveaux les plus profonds. Ainsi, pour éviter les redondances, seuls les résultats du niveau 3 sont présentés. Les tableaux et résultats des niveaux précédents sont mis en Annexe E.

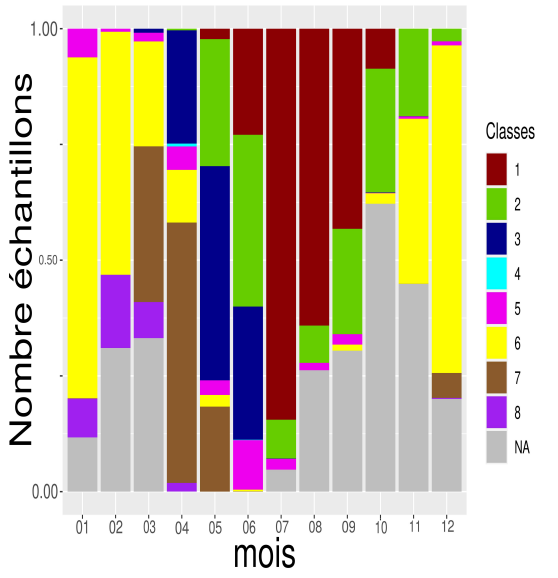
Tout comme pour le jeu test MAREL-Iroise 2006, les évènements extrêmes (l4 et l5) sont peu représentés et ne correspondent pas à des périodes identiques (Tableau 4.18). Toutefois, il est intéressant de voir que l5 - augmentation forte de la concentration en nutriments - est prédit pour la majorité des mois. Dans le cas de MAREL-Iroise, il semble que cet événement extrême soit plus à considérer comme un évènement de courte durée, mais plus périodique que pour le site d'étude MAREL-Carnot. La zone de MAREL-Iroise est sous influence forte des apports fluviaux de l'Aulne au sud et de l'Elorn au Nord-Est et du régime de marée (section 1.3.2.1), ce qui génère des apports brefs et intenses de nutriments, potentiellement plus réguliers que pour MAREL-Carnot. Les périodes d'occurrence du label l6 sont beaucoup plus courtes et le label l3 à une période plus longue que pour MAREL-Carnot (Figure 4.9a et 4.9c). l6 est caractérisé par de faibles températures et de faibles concentrations en nutriments et il correspond à une période hivernale avec de faibles productivités, et l3 est une période productive. Cependant, le site MAREL-Iroise est une zone eutrophe avec des apports en nutriments importants et sa période productive est plus longue que celle de MAREL-Carnot et s'étend de mars à octobre. Ces deux caractéristiques peuvent expliquer la prédiction plus importante de l3 (période productive) et la diminution de la période caractéristique de l6. Les labels l1 et l2 sont aussi peu prédits, mais le jeu de données contient un nombre important d'observations à données manquantes *NA* sur les périodes où elles sont dominantes (Figure 4.9d). La multiplication du nombre d'années dans le jeu test n'a pas permis de pallier ce manque d'informations lié au nombre de *NA* conséquent. En revanche, l'ajout de la fluorescence dans le modèle de prédiction permet une meilleure prédiction des labels l7 et l3 - efflorescence printanière. Ainsi, la période de forte productivité est bien identifiée par l7 et la période avec une productivité légèrement plus faible par l3.

Tableau 4.18 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) au niveau 3 (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MAREL-Iroise 2014-2016 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

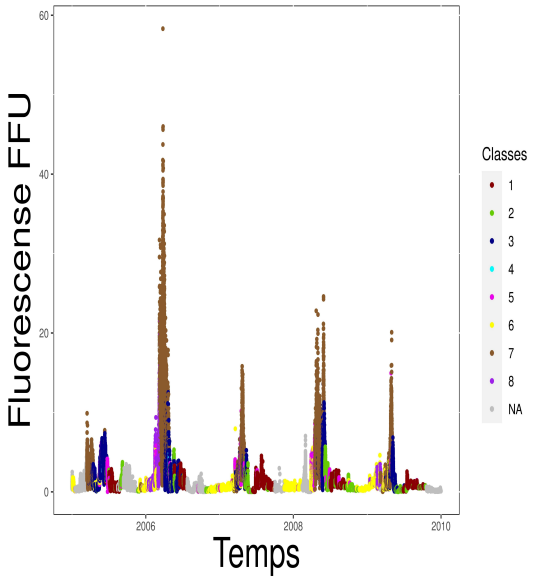
Classes non exp. vs Label prédits	l1	l2	l3	l4	l5	l6	l7	l8	RI	ARI
cl1	2	10	435	1	94	39	**2909**	14		
cl2	233	269	1150	1	75	45	**1305**	11		
cl3	0	74	857	0	277	**987**	817	1	0,67	0.17
cl4	69	174	**1445**	8	211	22	75	0		
cl5	24	76	**904**	4	79	0	73	0		

Dans le cas de MAREL-Iroise, il est donc possible, à ce niveau de classification, de caractériser la distribution et le séquencement des principaux états et d'identifier les grandes variations saisonnières sur un nouveau jeu de données à partir d'un modèle appris sur une autre base de données. Ainsi, le modèle de prédiction peut être une aide à la labellisation de nouvelles bases de données. Toutefois, les processus biogéochimiques, et donc les états, peuvent varier d'une région à une autre et le modèle peut ne pas être adapté à la nouvelle variabilité du jeu de données. Afin d'évaluer les limites de généralisation du modèle, la même étude a été reproduite sur le jeu de données MesuRho, dont les caractéristiques de l'écosystème sont plus marquées qu'avec celles de

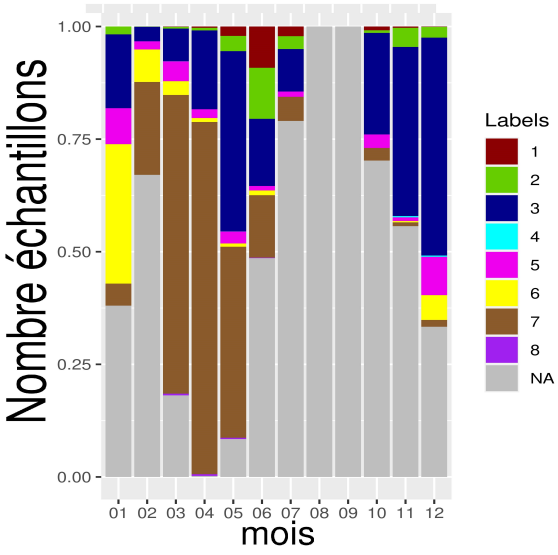
MAREL-Carnot.



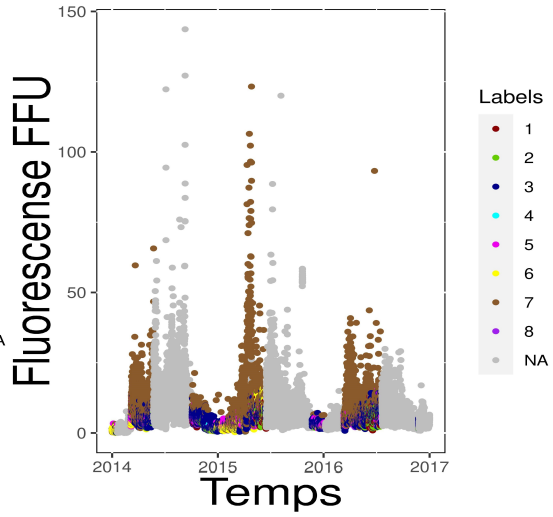
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot



(c) Distribution par mois des labels prédits pour MAREL-Iroise



(d) Distribution des labels prédits sur le signal de fluorescence (FFU) MAREL-Iroise

FIGURE 4.9 – Comparaison **au niveau 3** de la dynamique temporelle entre les classes M-SC à **MAREL-Carnot de 2005 à 2009** et les labels prédits pour **MAREL-Iroise sur la période 2014-2016**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise. La répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot, d) labels prédits pour MAREL-Iroise sur 2010-2014.

4.4.3 Prédictions sur MesuRho de 2010 à 2014

Les caractéristiques physico-chimiques et biologiques du site d'étude MesuRho étant sensiblement différentes de celles de MAREL-Carnot, un jeu de données test de plusieurs années est plus approprié. Le nombre d'observations plus important permet, en effet, de multiplier les informations et permet donc une meilleure classification de celles-ci. C'est pourquoi les résultats de prédiction sur le jeu de données MesuRho sont présentés pour la période 2010 à 2014. Les résultats pour le jeux test de l'année 2009 sont mis en Annexe E.

Structuration des classes

Dès le niveau *M-SC* 1 (Tableau 4.19), il est possible de constater une différence de structuration des variables entre les jeux de données MesuRho et MAREL-Carnot. En effet, une répartition inégale des classes est observable, avec une majorité des observations assignées au label l1, mais cette prédiction coïncide avec les caractéristiques du site. En effet, l1 est principalement structuré par de fortes températures et les températures du site d'étude MesuRho sont dans l'ensemble plus chaudes, avec des maximums enregistrés aux alentours de 34 °C pour MesuRho contre 23 °C et des températures médianes respectives de 16,37 °C et 13,00 °C. Ainsi la classe cl1 correspond à un état global de fortes températures et la classe cl2 à des périodes de baisse de température. Les scores RI et ARI, indiquant respectivement 0,75 et 0,14, font aussi ressortir une structuration des données en phase avec les labels de MAREL-Carnot.

Tableau 4.19 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 1** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MesuRho 2010-2014 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	RI	ARI
cl1	**5475**	494	0,75	0,14
cl2	432	**160**		

Pour le niveau *M-SC* 2, la structure des données s'exprime en 3 classes non exp. contre 4 labels pour la base d'apprentissage MAREL-Carnot. Les classes cl1 et cl2 ont, respectivement, des caractéristiques communes avec une période non productive (l1) et productive (l2) (Tableau 4.20). Les labels l2 et l4 sont définis comme une période productive, avec une concentration en chlorophylle a plus faible pour l2 que l4 ; la mise en évidence d'une classe non exp. proche des caractéristiques de l2 concorde encore une fois avec les caractéristiques environnementales de la zone d'étude. La classe cl3 a par contre une structuration qui est différente des classes du jeu d'apprentissage.

Tableau 4.20 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 2** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MesuRho 2010-2014 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	l3	l4	RI	ARI
cl1	**3035**	167	137	14		
cl2	1166	**1109**	267	72	0,59	0,19
cl3	**422**	11	160	1		

Au niveau *M-SC* 3, cela devient plus difficile de distinguer des structures communes bien marquées (Tableau 4.21). Comme pour le niveau *M-SC* 2, plusieurs classes, *i.e.* cl1 et cl2 , cl4 et

cl5, se démarquent comme ayant plus une affinité de faible production et des classes, *i.e.* cl3 et cl4, ayant une affinité plus productive.

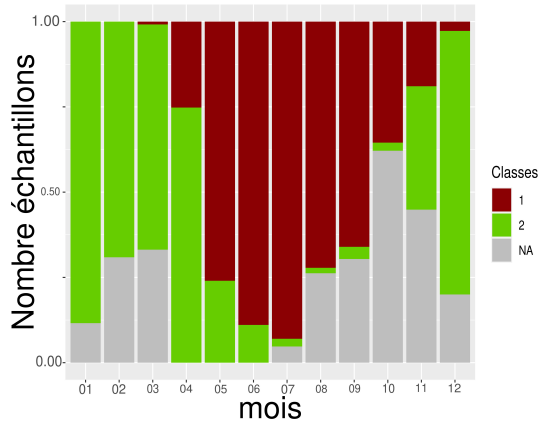
TABLEAU 4.21 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 3** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MesuRho 2010-2014 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	l3	l4	l5	l6	l7	l8	RI	ARI
cl1	**2069**	956	182	0	90	32	13	0	0,64	0,18
cl2	**24**	13	2	0	2	14	1	0		
cl3	316	298	**490**	0	70	123	50	2		
cl4	79	449	**602**	0	20	50	20	0		
cl5	**286**	131	11	0	20	56	1	0		
cl6	**80**	8	0	0	1	0	0	0		

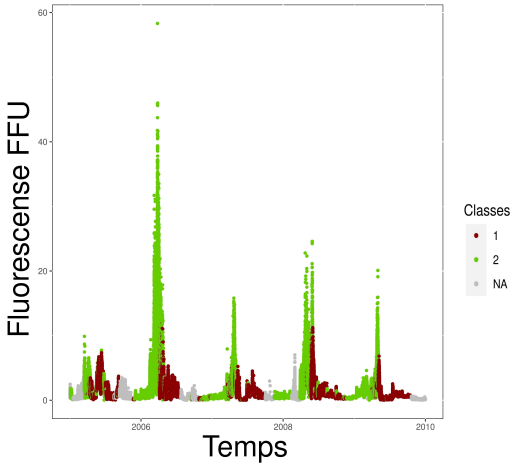
Répartition temporelle des labels prédits

En ce qui concerne la dynamique temporelle, il est difficile de retrouver un schéma équivalent à celui de MAREL-Carnot. Au premier niveau (Figure 4.10), le label l1 est prédit pour la majorité des observations du fait de sa structuration par les fortes températures, dominantes tout au long de l'année pour le site d'étude, relativement à MAREL-Carnot. Comme dans la partie précédente, une répartition inégale des labels est constatée. Ces différences de structure se retrouvent aussi pour les niveaux plus profonds (Figure 4.11 et 4.12). Par exemple, les périodes de bloom ne sont pas du tout identifiées dans ce cas de figure. Les différences de conditions environnementales entre les sites d'étude MAREL-Carnot et MesuRho sont plus importantes que celles avec MAREL-Iroise. C'est en grand partie pour cette raison que le modèle de prédiction a des difficultés à identifier des périodes clefs qui sont par définition propres à MAREL-Carnot. De plus, le nombre de variables EOVs communes utilisé pour l'apprentissage et la prédiction étant plus restreint que pour la classification, le modèle de prédiction peut se retrouver biaisé par le manque d'informations, surtout dans un cas de figure où les variables d'entrée ont une distribution différente forte. L'ajout de variables structurantes de la classification spectrale, comme les concentrations en Nitrate, Phosphate et Silicate ou encore le signal de fluorescences en FFU, aurait permis de diminuer ce biais et donc d'améliorer la qualité de la prédiction. Cependant, ces données ne sont pas toujours disponibles à ce jour pour cette station de mesure. Ce biais est aussi alimenté par le nombre important de *NA* dans les séries temporelles de la base de données.

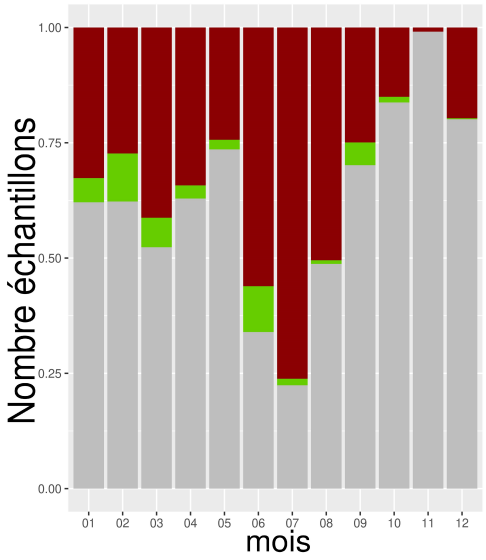
Ainsi, les résultats démontrent que le modèle est capable de prédire des états en cohérence avec la zone d'étude (exemple de label 1 au premier niveau), mais laisse percevoir que ce modèle, en l'état, n'est pas adapté à cette zone d'étude (manque d'informations essentielles pour définir les schémas de fonctionnement phytoplanctonique). Ici, le modèle de prédiction issu de MAREL-Carnot est proposé afin de prédire les potentiels états environnementaux à la station MesuRho. Bien entendu, dans le cas où une base de données labellisée est disponible sur le site d'étude. Il semble plus cohérent d'utiliser un modèle de prédiction local construit à partir du jeu de données MesuRho pour identifier des états caractéristiques de la zone.



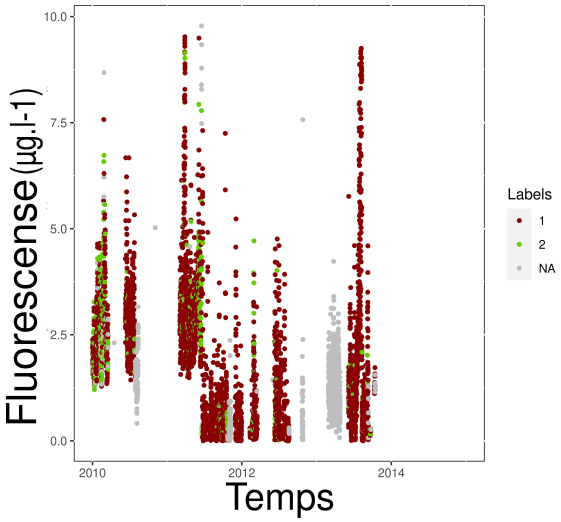
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot

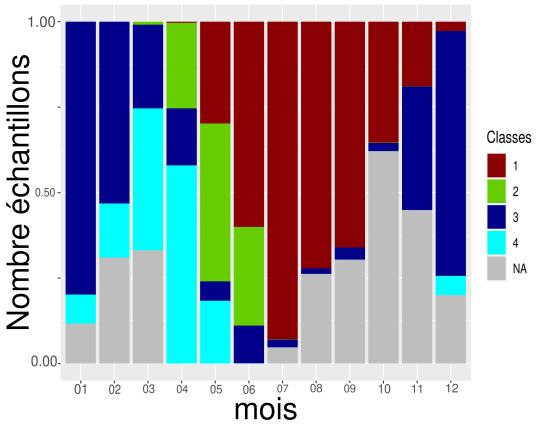


(c) Distribution par mois des labels prédits pour MesuRho

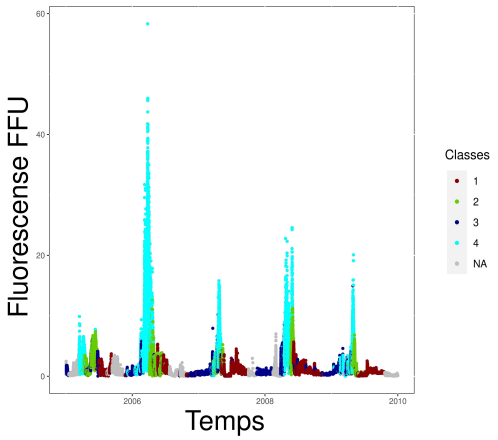


(d) Distribution des labels prédits sur le signal de fluorescence ($\mu g\, l^{-1}$) MesuRho

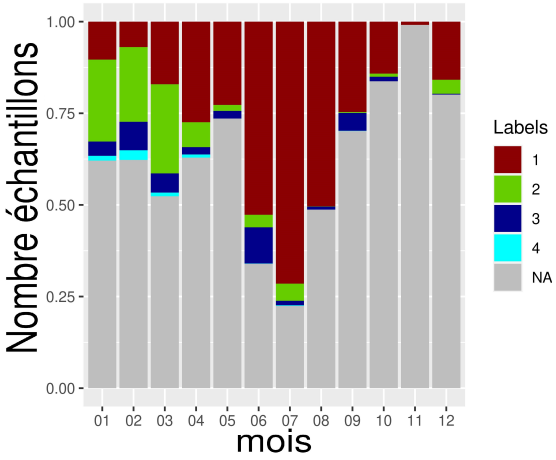
FIGURE 4.10 – Comparaison **au niveau 1** de la dynamique temporelle entre les classes M-SC à **MAREL-Carnot de 2005 à 2009** et les labels prédits **pour MesuRho sur la période 2010-2014**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MesuRho. La répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot, d) labels prédits pour MesuRho sur 2010-2014.



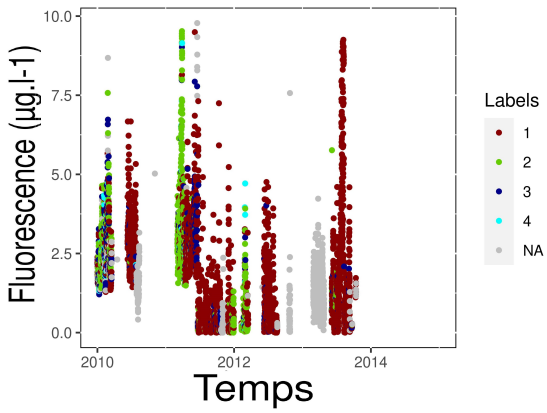
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot

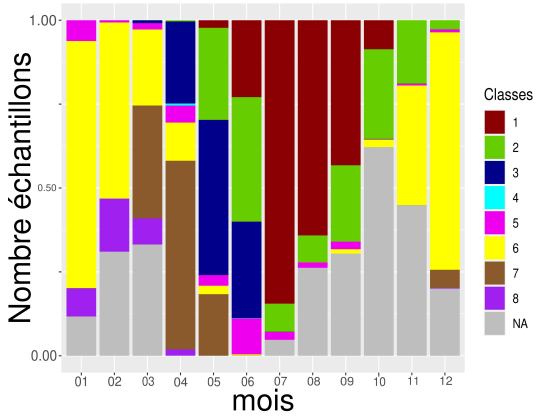


(c) Distribution par mois des labels prédits MesuRho

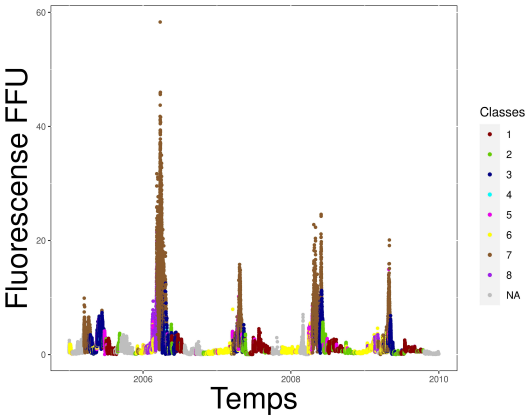


(d) Distribution des labels prédits sur le signal de fluorescence ($\mu g\, l^{-1}$) MesuRho

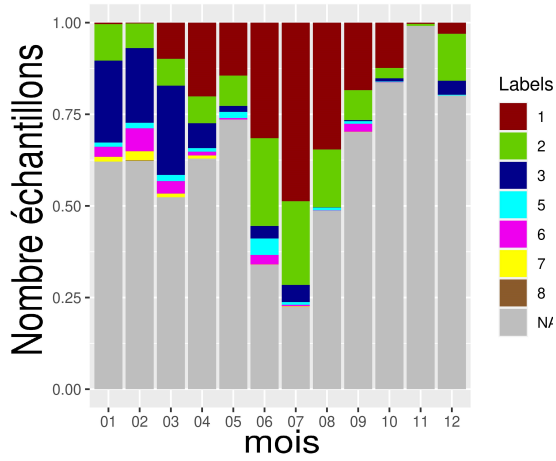
FIGURE 4.11 – Comparaison **au niveau 2** de la dynamique temporelle entre les classes M-SC à **MAREL-Carnot de 2005 à 2009** et les labels prédits pour **MesuRho sur la période 2010-2014**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MesuRho. La répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot, d) labels prédits pour MesuRho sur 2010-2014.



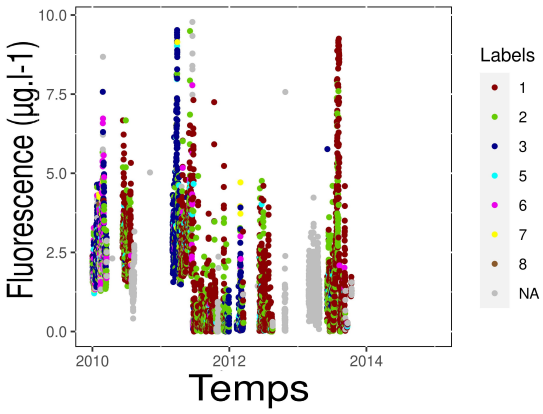
(a) Distribution par mois des labels appris via MAREL-Carnot



(b) Distribution des labels appris sur le signal de fluorescence (FFU) MAREL-Carnot



(c) Distribution par mois des labels prédits MesuRho



(d) Distribution des labels prédits sur le signal de fluorescence ($\mu g\, l^{-1}$) MesuRho

FIGURE 4.12 – Comparaison **au niveau 3** de la dynamique temporelle entre les classes M-SC à **MAREL-Carnot de 2005 à 2009** et les labels prédits pour **MesuRho sur la période 2010-2014**. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MesuRho. La répartition sur le signal de fluorescence b) des classes de M-SC pour MAREL-Carnot, d) labels prédits pour MesuRho sur 2010-2014.

4.5 Conclusions et perspectives

L'identification d'états environnementaux multi-critères (combinaison de paramètres physico-chimiques et biologiques) par une approche conjointe multivariée est une tâche complexe. Nous avons donc proposé de développer des modèles de prédiction adaptés afin d'identifier et de prédire ces états. La particularité de notre approche est de construire le modèle de prédiction supervisé à partir d'une base de données labellisée de façon non supervisée et de proposer, ainsi, une approche automatique et semi-supervisée.

La première phase de ce travail a été d'évaluer la capacité du modèle supervisé à prédire les états environnementaux définis par la méthode non supervisée *M-SC* et labellisés après expertise dans le chapitre 3. Les résultats aquis sur la base de données MAREL-Carnot démontrent une bonne capacité d'apprentissage et de généralisation des 3 systèmes de reconnaissances : *k-ppv*, *SVM*, *RF*. Ainsi, ces classifieurs supervisés sont capable de prédire de façon pertinente les labels issus de la classification spectrale non supervisée. Le modèle de prédiction construit permet de proposer un outil de labellisation des bases de données qui identifie les caractéristiques communes entre un jeu d'apprentissage et un jeu de données non labellisé avec un ensemble de variables identiques ou restreintes. Les méthodes, telles que *k-ppv* qui proposent une solution simple et rapide en temps de mise en œuvre et calculatoire, permettent une optimisation du temps de labellisation et fournissent une première estimation des états caractéristiques de la zone.

Les premiers tests effectués sur les autres sites d'études montrent qu'il est possible d'identifier les caractéristiques environnementales d'une nouvelle base de données à partir d'un modèle issu de l'apprentissage d'une autre base. Néanmoins, les résultats de prédiction plus performants sur la base de données MAREL-Iroise que sur la base de données MesuRho démontrent les limites du modèle de prédiction et suggèrent une utilisation régionalisée du protocole d'apprentissage semi-supervisé. Ainsi, nous pouvons considérer que cette approche peut être un outil d'aide à l'interprétation de nouvelles sources de données fiables dans un cadre régionalisé. Dans la zone Manche - Mer du Nord, le modèle d'apprentissage construit à partir de la base de données labellisée MAREL-Carnot pourrait être appliqué sur des jeux de données issus de la bouée du réseau COAST-HF de l'IR-ILICO : SMILE installée en Baie de Seine ou encore la bouée ASTAN au large de Roscoff. Un autre cas de figure envisageable serait l'utilisation d'un modèle appris sur la base de données issue d'une bouée fixe et la caractérisation des états par le modèle lors d'un déploiement ponctuel lors d'une campagne spécifique proche de la zone.

Assurément, ces conclusions et recommandations sont basées sur des ensembles réduits d'EOVs et d'années exploitables, liées aux maintenances et modifications de plateformes importantes des stations de mesure MAREL. Dans le chapitre suivant, nous démontrons le potentiel d'identification et de prédiction des schémas de fonctionnement de l'outil *M-SC* combiné à des approches supervisées afin de définir, étudier et prédir la dynamique spatio-temporelle au regard de l'existence de perturbations multi-échelles environnementales.

Chapitre 5

Autres applications

5.1 Éco-Régions marines - données CGFS

Au chapitre précédent, la méthode de classification *M-SC* est appliquée dans un contexte de segmentation temporelle avec une illustration sur la caractérisation d'événements à partir des données du système de mesures MAREL. Dans le chapitre 3 section 2.4.4, elle a été validée sur des données temporelles ou spatialisées. Ici, nous abordons le cadre spatial pour une application opérationnelle avec un jeu de données issu de la campagne en mer CGFS (Channel Ground Fish Survey), dans le but d'identifier des éco-régions marines.

5.1.1 Contexte Général

Dans le cadre du programme de surveillance des descripteurs 1-Habitat Pélagique et 5-Eutrophisation de la DCSMM [DCSMM 2008/56/CE], un suivi saisonnier des paramètres physico-chimiques et biologiques a été mis en place à l'échelle des Sous-Régions Marines (SRM) Manche Mer du Nord et Mer Celtique en 2018. Ce suivi est effectué lors de plusieurs campagnes, dont la campagne halieutique CGFS (section 2.2.2). Pour compléter les évaluations DCSMM via des métriques imposées, les outils de mesures HF automatisées (FerryBox et Pocket FerryBox) ont été proposés comme complément aux méthodes conventionnelles Basses Fréquences. Pour la campagne CGFS, un FerryBox est déployé depuis 2017 sur le Navire Océanographique la Thalassa.

L'utilisation d'un dispositif de mesures automatisées à HF embarqué à bord de navire océanographique, tels que les FerryBox, offre l'opportunité de pouvoir mesurer en continu différents paramètres physico-chimiques et biologiques (température, salinité, oxygène, turbidité, concentration en chlorophylle a de groupe phytoplanctonique, ...) (Section 2.2.3). Les échelles d'observation rendues ainsi accessibles par ces dispositifs (une mesure par minute) permettent d'étudier des processus impossibles à observer avec des mesures à fréquences dites « conventionnelles » (effectuées à l'aide d'une bouteille Niskin ou d'un filet WP2). Néanmoins, l'utilisation de ces données HF n'est pas sans contrainte : les séries de données obtenues peuvent contenir de nombreuses données manquantes ou aberrantes (pannes, maintenances, dérive de capteurs, ...). La quantité de données générées par ces systèmes nécessite l'emploi de méthodes numériques spécifiques afin d'optimiser les phases de pré-traitement et de traitement des données pour en extraire le maximum d'informations et comprendre leurs structures ainsi que leurs dynamiques. Le déploiement de systèmes HF implique également d'être capable de mettre à disposition les données ainsi que les

résultats en temps quasi-réel. Aussi, dans le contexte de l'évolution des systèmes d'observation et de surveillance du milieu marin, le développement des outils d'aide aux traitements de telles données et d'aide à la prise de décisions (échantillonnages adaptatifs, prévisions d'événements) est essentiel.

Parmi les outils déjà développés, l'outil FBdataM [Lefebvre et Devreker 2019b], permet de fusionner les différents fichiers de données produites quotidiennement par la FerryBox du Navire Océanographique la Thalassa (ainsi que celles du Pocket FerryBox du Navire Océanographique l'Europe) avec les autres données du bord stockées dans TECHSAS/CASINO et de visualiser rapidement leur contenu à travers une interface. Un outil tel que TTAinterface (R-CRAN : : TTAinterfaceTrendAnalysis) [Devreker et Lefebvre 2020] permet d'effectuer des analyses statistiques de base et des diagnostics (boxplots, anomalies, autocorrelations, ...) sur des séries temporelles univariées et d'en extraire la tendance temporelle. Mais ce n'est pas un outil d'analyses de données HF et il est nécessaire de convertir les données pour les rendre compatibles avec l'interface.

Pour aller plus loin, nous avons donc testé notre approche HF et multidimensionnelle (*M-SC*) sur les données FerryBox (issues de la campagne CGFS 2018). L'idée est d'identifier, à partir des données HF et multivariées, des éco-régions caractérisées par la structuration des données dans un espace spectral et d'analyser leur répartition spatiale. D'autres travaux basés sur les données de campagne CGFS ont déjà défini des éco-régions, par exemple, par le SHOM sous forme de paysages marins dans le cadre du descripteur 7- Condition hydrologique de la DCSMM [Tew-Kai et al. 2020] ou pour la répartition spatiale de communautés de poissons [Vaz et al. 2007]. Nos résultats sont donc comparés à ces études.

5.1.2 Présentation de la base de données

Pour CGFS 2018 [Coppin 1988], le jeu de données FerryBox est issu d'une campagne en Manche sur la période du 12 septembre au 11 octobre 2018 (Figure 5.1).

Les données sont référencées via la plate-forme SISMER mais ne sont pas encore en accès libre. Prochainement, toutes les données seront disponibles sur la plateforme Coriolis.

Le FerryBox collecte 31 variables en plus de la date, de l'heure, de la latitude et de la longitude. Avec un échantillonnage toutes les minutes, le fichier aligné se compose de 42 218 observations par variable avec environ 10 % *NA*. Les *NA* peuvent être dues à des phases d'amorçage et de rinçage du FerryBox, ou aux entrées et sorties de port. Seules 8 variables, faisant partie des EOVs, sont utilisées comme variables contributives pour la classification (résumées dans le tableau 5.1).

Les quatre premières variables sont des paramètres physico-chimiques : Température (°C), Salinité (PSU), Turbidité (NTU), Oxygène dissous ($\mu mol\,l^{-1}$). Les quatre autres variables sont des paramètres issus de l'Algae Online Analyser (AOA)(Section 2.2.3) : AOA Algues Vertes (éq. $\mu g\,l^{-1}$), AOA Algues Bleues (éq. $\mu g\,l^{-1}$), AOA Algues Brunes (éq. $\mu g\,l^{-1}$), AOA Cryptophytes (éq. $\mu g\,l^{-1}$).

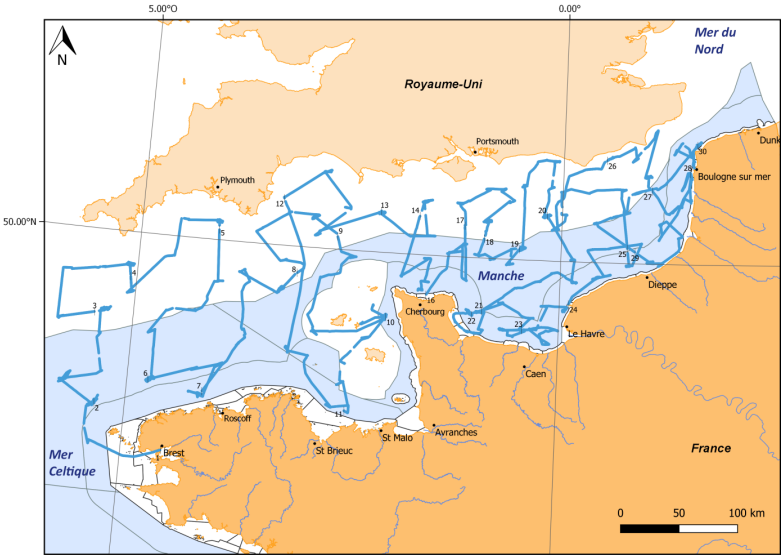


FIGURE 5.1 – Trajet de la campagne CGFS 2018 représenté par les points de mesure du FerryBox du N/O Thalassa. Les numéros le long du parcours représentent les jours successifs de campagne. (source : rapport ODE [LEFEBVRE et DEVREKER 2009])

TABLEAU 5.1 – Caractéristiques des variables FerryBox utilisée pour la classification : minimum (min), maximum (max), moyenne, médiane et quartiles (Q1-Q3), % de données manquantes NA.

Paramètre (unités)	Moyenne-medianne	min-max	Q1-Q3	# % NA
Température (°C)	17,30 - 17,56	13,67 - 18,84	16,97 - 17,86	9,90
Salinité (PSU)	34,67 - 35,01	26,26 - 35,38	34,67 - 35,17	9,90
Turbidité (NTU)	5,67 - 5,20	0,41 - 515,03	1,80 - 8,94	9,90
Oxygène dissous ()	259,3 - 257,5	237,7 - 302,0	252,8 - 263,8	9,90
AOA Algues Vertes ($\mu g\,l^{-1}$)	0,29 - 0,33	0,00 - 1,78	0,00 – 0,57	9,90
AOA Algues Bleues ($\mu g\,l^{-1}$)	0,002 - 0,00	0,00 - 5,88	0,00 - 0,00	9,90
AOA Algues Brunes ($\mu g\,l^{-1}$)	0,96 - 0,63	0,00 - 9,81	0,09- 1,30	9,90
AOA Cryptophytes ($\mu g\,l^{-1}$)	0,83 - 0,82	0,00 - 20,18	0,57 - 0,99	9,90

5.1.3 Résultats de la classification M-SC

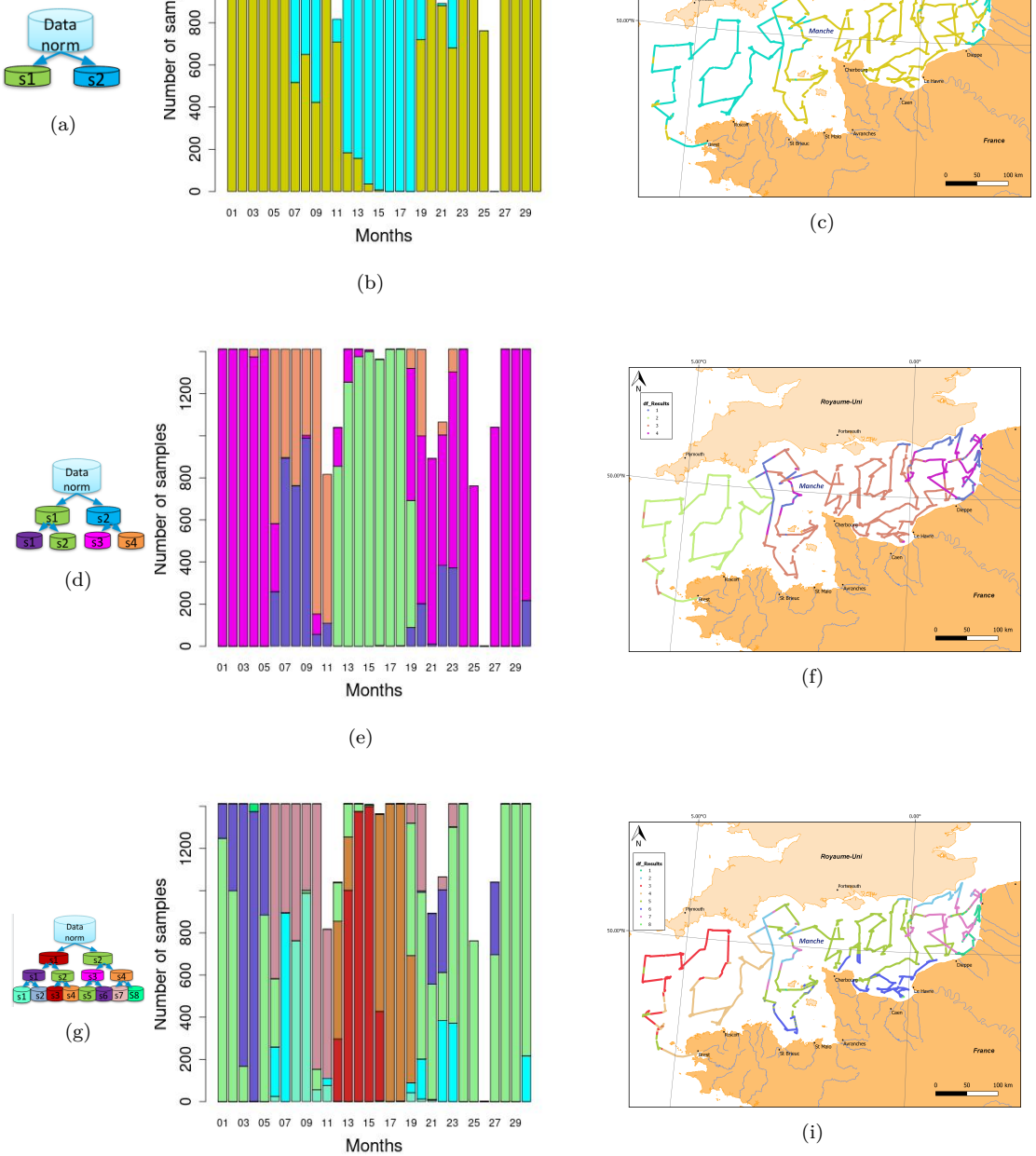
Nous présentons les résultats issus de la classification non supervisée *M-SC* par niveau.

Le niveau 1 divise les données en deux classes : cl1 en vert et cl2 en cyan. Ce niveau de classification fait une distinction entre les deux bassins de la Manche (Figure 5.2c). Il différencie la Manche occidentale dont les eaux sont plus froides, moins turbides et moins chargées en algues vertes que les eaux de la Manche orientale.

Le niveau2 subdivise cl1 et cl2 en quatre sous classes : cl1 en violet, cl2 en vert, cl3 en rose et cl4 en corail. Ce niveau identifie clairement 3 masses d'eau (Figure 5.2f) : une masse d'eau du large sous influence de l'océan atlantique (cl2) plus froide, une masse d'eau au centre de la manche orientale (cl4) avec des caractéristiques plus côtières (moins salée et avec une présence d'algues vertes et marrons fortes) et une masse d'eau plus à l'Ouest sous influence des eaux de la mer du Nord (cl1) plus chaude. cl3 par contre est moins localisée.

Le niveau 3 met en évidence 8 zones avec une différence environnementale du point de vue de la signature phytoplanctonique (Figure 5.2i). Une distinction, notamment au niveau des zones côtières françaises, apparaît. La classification isole par exemple la baie de Somme (cl8 en vert flache), la baie de Seine et le Golf Normano-breton (cl6 en violet), ainsi que des zones de front tel que la zone de front de Ouessant (cl3 en rouge). Ces dernières sont toutes les trois chargées en algues brunes et ont des salinités relativement faibles. La différentiation entre ces trois classes se fait au niveau de la température. cl3 est la classe la plus froide et cl8 la plus chaude. D'autres classes, comme cl4 en corail ou cl5 en vert affichent une présence d'algues bleu-vert et des cryptophytes. Elles sont étendues et correspondent à des masses d'eaux localisées au large.

À chaque niveau, la classification met en avant des éco-régions cohérentes avec la géographie et les masses d'eaux caractéristiques de la zone d'étude. L'arbre multi-niveaux *M-SC* obtenu permet de définir des éco-régions de plus en plus spécifiques. Ces éco-régions peuvent être mises en lien avec d'autres classifications réalisées à partir de protocoles différents.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIGURE 5.2 – Application de la méthode M-SC à 3 niveaux sur le jeu de données CGFS 2018. Schémas de la classification par niveau (Figures a, d, g). Fréquence d'occurrence de chaque état par jour et par niveau (Figures b, e, h) et carte de la répartition des classes sur le trajet de la campagne (Figures c, f, i).

5.1.3.1 Validation des résultats : Exemple des paysages marins

Afin de tester la capacité de la méthode à identifier différentes éco-régions de manière non supervisée, une comparaison est effectuée entre les éco-régions définies par *M-SC* et les paysages marins calculés à partir de 11 paramètres hydrodynamiques par le SHOM [Tew-Kai et al. 2020] (Figure 5.3). Les paysages marins utilisés sont les zones définies par l'équipe du descripteur 7 (Changements Hydrographiques) du SHOM pour l'évaluation DCSMM 2018 du descripteur 1 (Habitat Pélagique) [DCSMM 2008/56/CE]. Il y a 12 paysages marins, énumérés de 1 à 12, identifiés pour toute la façade Atlantique-Manche-Mer du Nord (Tableau 5.2). Nous nous focalisons uniquement sur la zone Manche-Mer du Nord, soit les paysages marins numérotés de 1 à 8 et le numéro 10, le numéro 12 étant les plaines abyssales et ne concernant que la façade Atlantique.

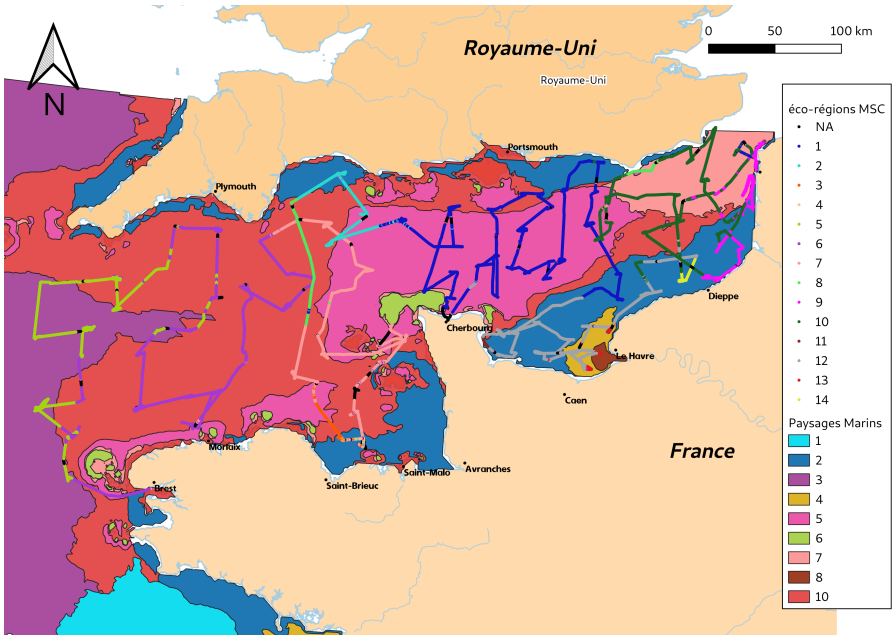
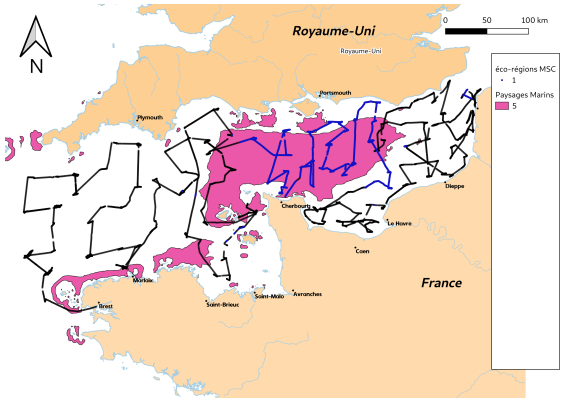


Figure 5.3 – Les 14 éco-régions définies par Classification Spectrale Multi-niveaux (M-SC) et transposées sur le trajet de la campagne GCFS 2018 et superposées aux 10 paysages marins définis par le SHOM pour les descripteurs 1 (Habitat pélagique), 5 (Eutrophisation) et 7 (Changements Hydrographiques).

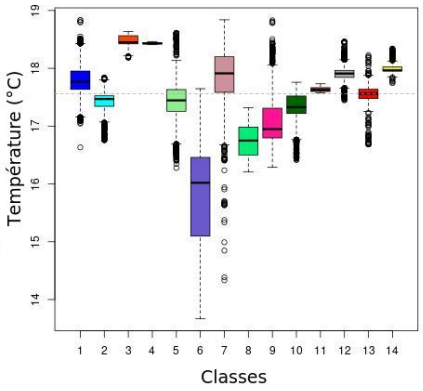
Tableau 5.2 – Liste des 10 passages marins et de leurs dénominations.

Paysages Marins	Dénominations
PM1	plateau large GdG
PM2	zones cotières
PM3	large MC
PM4	panaches
PM5	zones Manche soumises à la marée
PM6	zones fortement énergétiques
PM7	zones énergétiques soumises à la marée
PM8	estuaires
PM10	zones sud Gascogne
PM12	plaines abyssales

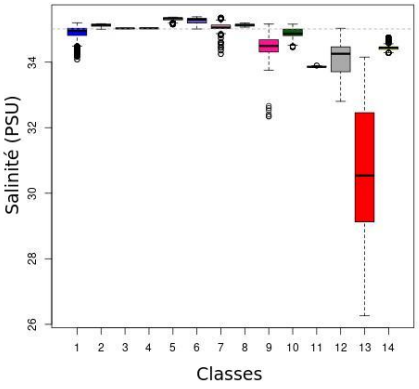
Au niveau 3 de la classification *M-SC*, 8 classes sont détectées, ce qui demeure inférieur aux nombres de paysages marins du SHOM. Pour être plus proche de ce découpage par paysages marins, un 4^{e} niveau de classification est calculé. À ce niveau, 14 classes sont identifiées. Il est important de noter que la concordance ne peut pas être parfaite, car les paysages marins et les éco-régions ne sont pas calculés à partir des mêmes variables. Toutefois, la superposition montre certaines similitudes. Ainsi, plusieurs éco-régions définies par *M-SC* sont analogues aux paysages marins du SHOM (Figure 5.3).



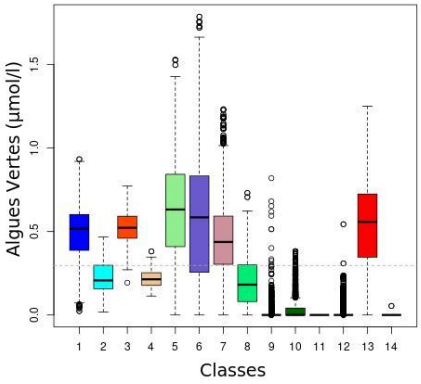
(a) PM5 vs cl1



(b) Température (°C)



(c) Salinité (PSU)

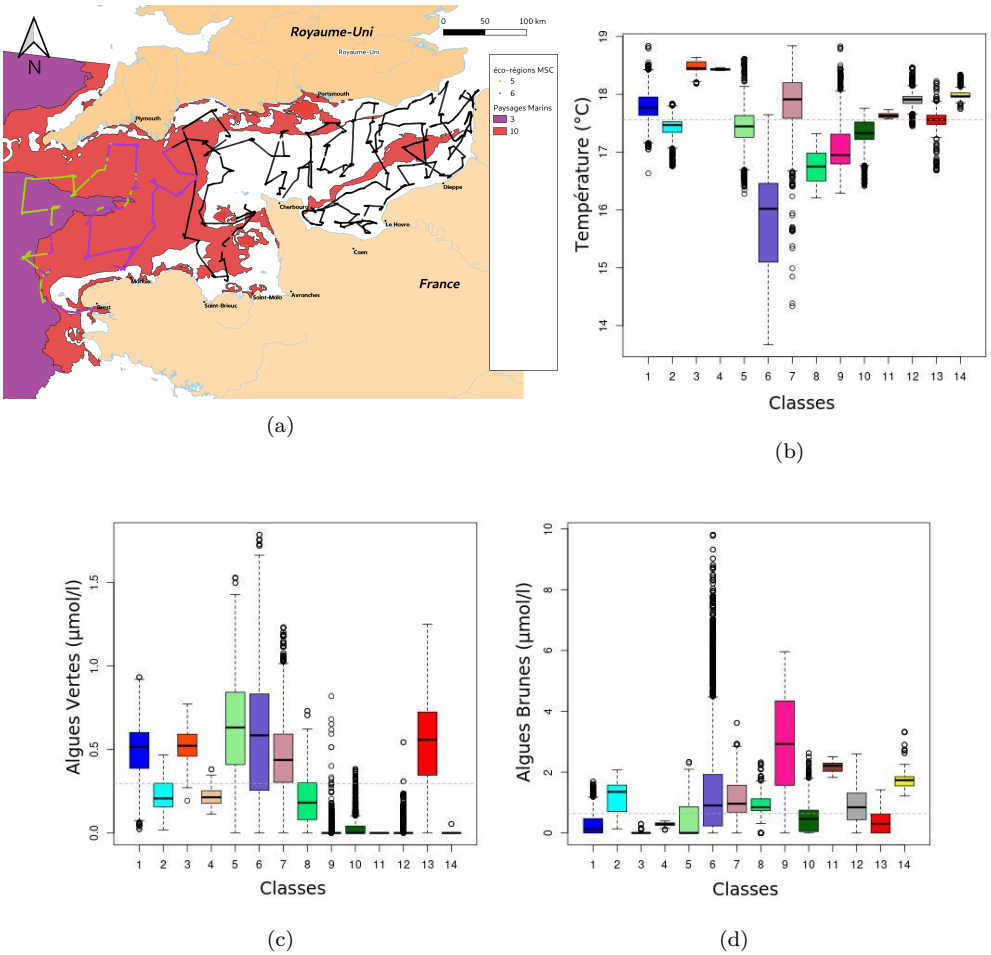


(d) Algues Vertes (µg l^{-1})

FIGURE 5.4 – Classification au niveau M-SC 4 des données de la campagne CGFS 2018 en Manche sur la période de septembre à octobre. (a) répartition spatiale de la classe 1 (cl1) et du paysage marin 5 (PM5 - eaux mélangées sous influence tidale) ; Boîte de dispersion par classe (b) de la température (°C), (c) de la salinité (PSU) et (d) de la fluorescence AOA - algues vertes (µg l^{-1}). La ligne grise en pointillés représente la valeur moyenne de la variable.

Par exemple, la classe 1 (cl1) peut être associée au paysage marin 5 (PM5 - eaux mélangées sous influence tidale). Elle est répartie sur la quasi-totalité de l'aire du PM5. Seuls les trajets proches des zones côtières autour de la Bretagne ne sont pas affectés à cl1 (Figure 5.4a) et sont définis par d'autres classes. Mais le trajet du navire ne passe que très peu dans cette zone. cl1

est caractérisée par des températures (Figure 5.4b) et des salinités (Figure 5.4c) globalement au dessus de la moyenne et, par la présence importante d'algues vertes (Figure 5.4d). Ces propriétés biogéochimiques sont concordantes avec les caractéristiques des eaux mélangées de la Manche centrale.



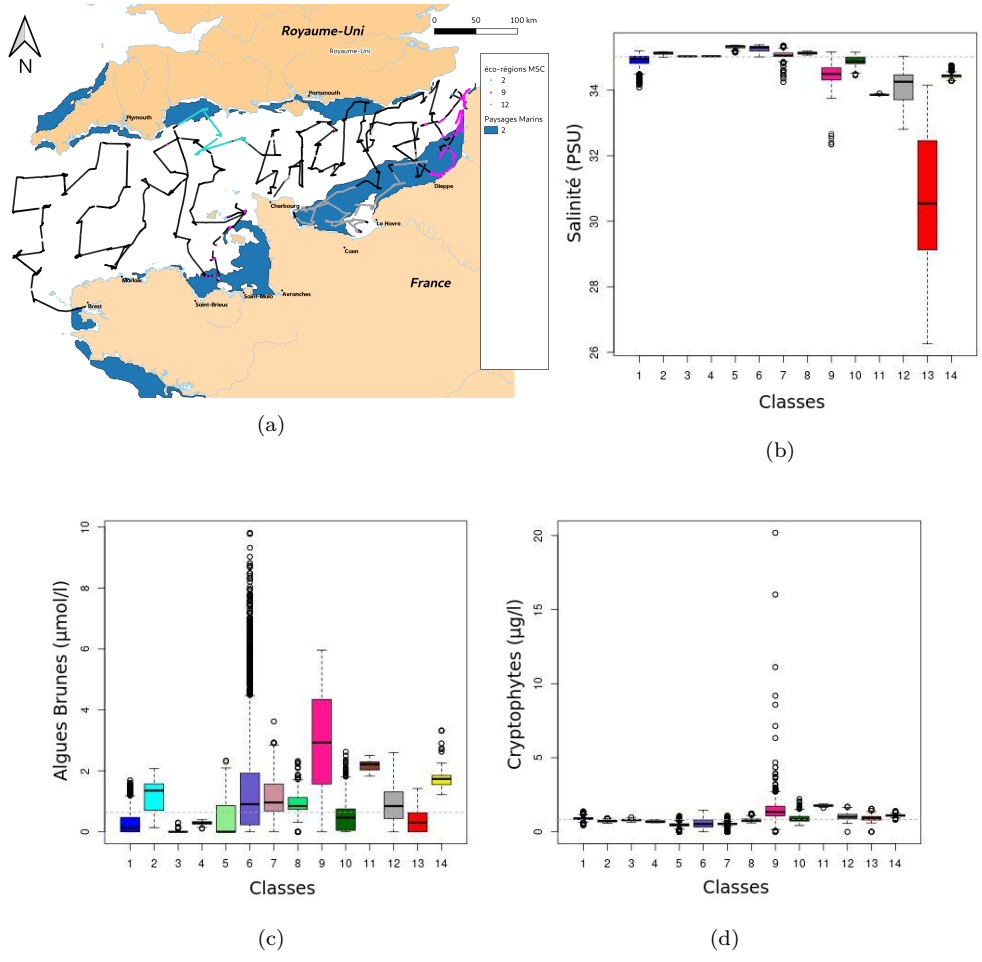
(a) (b) (c) (d)

FIGURE 5.5 – Classification au niveau M-SC 4 des données de la campagne CGFS 2018 en Manche sur la période de septembre à octobre. (a) répartition spatiale de la classe 5 (cl5) et 6 (cl6) et des paysages marins 3 (PM3- eaux du large avec stratification saisonnière) et 10 (PM10 - eaux mélangées) ; Boîte de dispersion par classe (b) de la température (°C), (c) de la fluorescence AOA - algues vertes (éq. $\mu g\,l^{-1}$) et (d) de la fluorescence AOA - algues brunes (éq. $\mu g\,l^{-1}$). La ligne grise en pointillés représente la valeur moyenne de la variable.

Une autre concordance peut être faîte pour la classe 5 (cl5) et la classe 6 (cl6). Elles se rapportent respectivement aux paysages marins 3 (PM3 - eaux du large avec stratification saisonnière) et 10 (PM10 - eaux mélangées). Elles sont toutes deux caractérisées par une dominance d'algues vertes et d'algues brunes (Figure 5.5c et 5.5d). Cet assemblage est représentatif des zones du large. La distinction entre les deux classes se fait principalement au niveau de la température, cl6 étant bien plus froide que cl5 (Figure 5.5b). Le front thermique de Ouessant, localisé à la

limite Ouest du PM10 et à la limite Est du PM3, peut expliquer cette différence de température. Les eaux plus froides de cl6 sont bien caractéristiques de la zone située avant le front thermique (zone plus froide et plus mélangée) et les eaux de surface plus chaudes de cl5 définissent bien une zone de stratification plus élevée comme pour le PM3. La classe cl6 est aussi plus chargée en algues brunes, ce qui coïncide avec une entrée dans le bassin occidental et la présence de masses d'eaux plus côtières.

De plus, il est intéressant de noter que les paysages marins ne sont pas définis avec une frontière "linéaire". Ici, une partie du PM3 est imbriquée dans le PM10. Cette distinction est aussi faite par la classification *M-SC*, où l'imbrication entre cl5 et cl6 est bien visible sur le parcours de la CGFS (Figure 5.5a).



(a) (b) (c) (d)

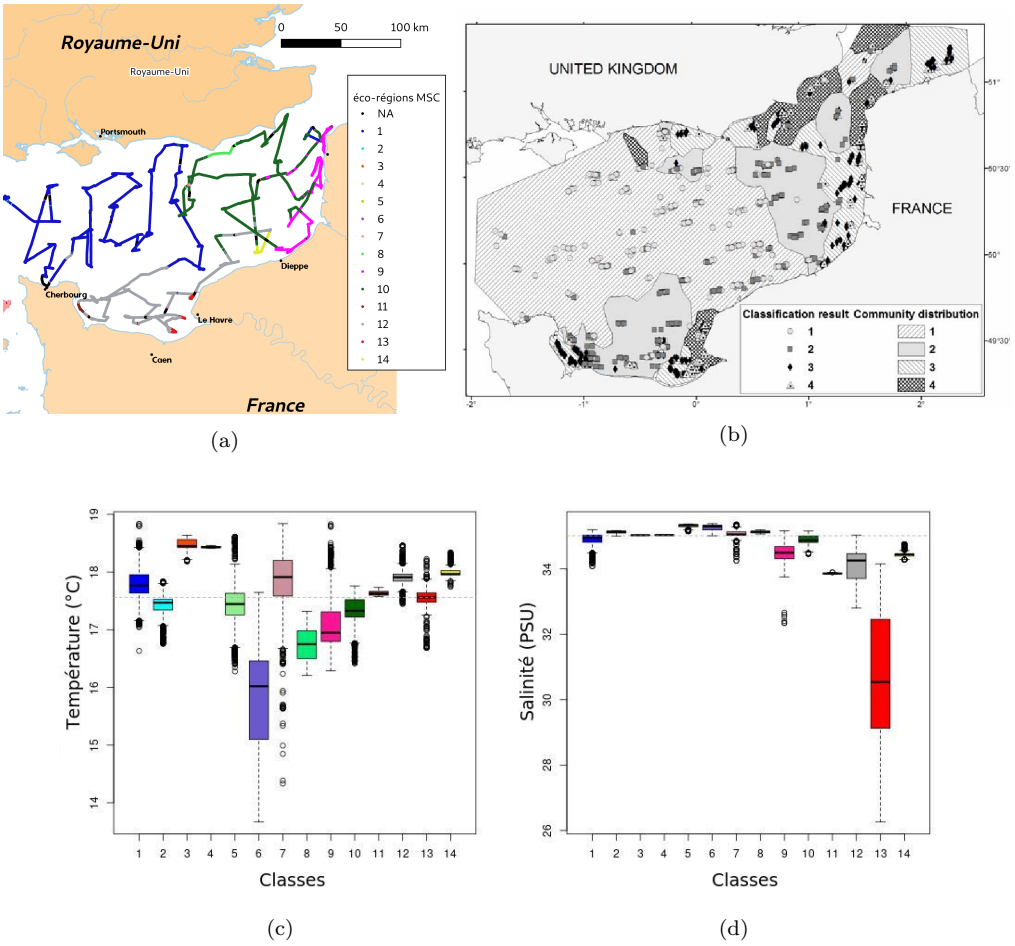
FIGURE 5.6 – Classification au niveau M-SC 4 des données de la campagne CGFS 2018 en Manche sur la période de septembre à octobre. (a) répartition spatiale des classes 2 (cl2), 9 (cl9) et 12 (cl12) et du paysage marin 2 (PM2 - eaux côtières et peu profondes sous l'influence des eaux douces) ; Boîte de dispersion (b) de la salinité (PSU), (c) de la fluorescence AOA - algues brunes (éq. $\mu g\,l^{-1}$) et (d) de la fluorescence AOA - cryptophytes (éq. $\mu g\,l^{-1}$). La ligne grise en pointillés représente la valeur moyenne de la variable.

Les classes 2 (cl2), 9 (cl9) et 12 (cl12) sont toutes apparentées aux paysages marins 2 (PM2 - eaux côtières et peu profondes sous l'influence des eaux douces) (Figure 5.6a). Elles sont caractérisées par de faibles salinités, en particulier cl9 et cl12 (Figure 5.6b), ce qui concorde avec des zones sous l'influence d'apports en eau douce. cl2 est identifiée plus au large, elle est donc plus salée. Elles sont aussi toutes caractérisées par de fortes concentrations en algues brunes. cl12 et cl9 comptent aussi de fortes concentrations en cryptophytes, relativement aux autres classes. En revanche cl2, située au niveau des côtes anglaises, n'a pas une forte concentration en cryptophytes. La présence de ces deux groupes correspond bien à la zone côtière où la diversité phytoplanctonique est plus importante. Dans ce cas de figure, la classification *M-SC* est plus restrictive que le découpage par paysage marin. Elle permet de faire une distinction au sein de la zone côtière (PM2) et d'identifier 3 éco-régions côtières : les côtes anglaises (cl2), la baie de Seine (cl9) et la baie de Somme (cl12).

5.1.3.2 Validation des résultats : Exemple des assemblages de poissons

La classification par *M-SC* a aussi été mise en relation avec une classification des communautés de poissons. Dans une démarche similaire à la notre (multivariées et multi-spécifiques), l'étude de Vaz et al. 2007 définit des assemblages halieutiques et détermine une bio-régionalisation de la Manche orientale (Figure 5.7b). Elle met en évidence la relation entre la structure des assemblages halieutiques et de l'environnement de la zone. Dans ces travaux, quatre types d'assemblages principaux à méso-échelle ont été définis : « Le groupe 1 est caractéristique d'une communauté benthique associée à des sédiments durs et de galets, une hydrologie océanique (salinités et températures élevées en octobre), de forts courants de marées et des eaux relativement profondes. Le groupe 2 est caractéristique d'une communauté benthique, associée à des sédiments de galets et de sables grossiers, avec une hydrologie et une bathymétrie intermédiaires entre le large et la côte. Le groupe 3, la communauté benthique, est associé à des sédiments de sable fin, à l'hydrologie et à la bathymétrie côtières (faible salinité et température en octobre, eaux peu profondes, courants plus faibles). Le groupe 4 est caractérisé par des types de sédiments hétérogènes, des boues aux sables grossiers, et les divers types de communautés benthiques associés, ainsi que par l'hydrologie et la bathymétrie côtière »[Vaz et al. 2007]

Pour la Manche orientale, la classification par *M-SC* met en évidence 6 classes et 4 d'entre elles peuvent facilement être associées aux 4 assemblages halieutiques. cl1 est relatif au groupe 1. Elle présente, tout comme le groupe 1, des caractéristiques hydrologiques et biologiques océaniques : avec une salinité et une température au dessus de la moyenne (Figure 5.7c et 5.7d). cl12 coïncide avec le groupe 2. Elle est également caractérisée par une hydrologie intermédiaire entre la côte et le large avec des salinités faibles et des températures relativement élevées (Figure 5.7d et 5.7c). cl12 a une aire de répartition plus importante que celle définie pour le groupe 2. Mais la séparation nette entre les groupes 2 et 4 se retrouve aussi entre la cl12 et cl13. cl13 est structurée par des salinités très faibles. Elle est représentée dans les zones sous fortes influences des apports d'eau douce. Elle est donc cohérente avec le groupe 4. Enfin, cl9 peut être rattachée au groupe 3. Elle présente aussi des caractéristiques hydrologiques côtières avec des salinités faibles et des températures plus faibles (en octobre) que la moyenne des autres classes.



(a) (b)

(c) (d)

FIGURE 5.7 – Classification au niveau M-SC 4 des données de la campagne CGFS 2018 en Manche Est sur la période de septembre à octobre. (a) répartition spatiale des classes ; (b) répartition des 4 assemblages de communautés benthiques. ; Boîte de dispersion par classe (c) de la température (°C) et (d) la salinité (PSU).

5.1.4 Conclusions et perspectives

Avec cette application, nous avons pu démontrer que l'approche *M-SC* est tout aussi performante sur des cas d'études spatialisés. Elle est capable de fournir un découpage cohérent à haute définition spatiale et temporelle. Cette approche réussit à identifier des masses d'eau (paysages marins, éco-régions) avec des répartitions hétérogènes et détecte aussi bien de grandes structures comme les masses d'eau au large, que des structures fines au niveau côtier. Elle permet donc de prendre en considération la variabilité des différents paramètres, de les intégrer de manière multivariée et donc d'identifier des processus. L'intégration des données de fluorescences (AOA) ajoute une dimension supplémentaire à la classification et permet la définition des relations entre les communautés et les propriétés hydrologiques et bio-géochimiques de l'environnement.

L'approche multivariée et multi-échelle de *M-SC* permet d'introduire de nouveaux concepts et de tendre vers une approche intégrée pouvant renseigner plusieurs compartiments écologiques et différents niveaux trophiques. Elle permet ainsi de proposer des évaluations globales ou plus locales

qui peuvent aider à des prises de décisions dans le cadre de projets qui visent à établir des bilans de l'état du milieu marin (DCSMM, OSPAR) ou dans le cadre de la structuration des réseaux d'observation (projet H2020 JERICO-S3). Elle pourrait, par exemple, aider à valider et à améliorer la définition des structures hydrologiques particulières liées ou associées aux paysages marins. La mise en place d'une interface utilisateur associée aux données FerryBox pourrait également contribuer à la définition et à l'adaptation en temps *quasi-réel* des stratégies d'échantillonnages lors des campagnes en mer. Un échantillonnage supplémentaire ou la délocalisation d'un point de mesure pourrait être fait lorsque la méthode détecte une zone d'intérêt ou un changement des caractéristiques des masses d'eaux traversées.

5.2 Approche météorologique : Application de M-SC à des données de précipitations

5.2.1 Contexte général

Au cours de la dernière décennie, les événements extrêmes (tempêtes, inondations, sécheresses) sont de plus en plus fréquents et/ou intensifs et 85 % de la population mondiale est touchée par ces événements. Selon les Nations Unies, 70 % de l'économie mondiale dépend des activités dites météo sensibles. Or 75 % des pays les plus vulnérables ne disposent pas ou peu d'informations météorologiques fiables, précises et efficaces. Selon l'Organisation Météorologique Mondiale (*OMM*) seul un tiers des pays disposent de moyens matériels et/ou de logistique pour offrir des informations météorologiques hautes résolutions à leur population comme par exemple des modèles météorologiques à haute résolution, des réseaux de stations météorologiques, ou des données de satellites. De même, certains Services Météorologiques Nationaux (*SMN*) ont des difficultés à déployer et à maintenir une infrastructure opérationnelle telle qu'un enregistreur pluviométrique. Or les acteurs économiques et institutionnels œuvrant dans ces régions peu instrumentées ont besoin de ces informations pour éclairer leurs décisions en réponse aux risques et aux opportunités pour leurs activités. Dans ce but, WeatherForce (financeur majoritaire de cette thèse) a pour mission de fournir à ses clients une information précise, opérationnelle, contextuelle et fiable sous forme d'indicateurs pour apporter des solutions.

Dans ce contexte, la Sodexam (*SMN* de Côte d'Ivoire) s'intéresse à l'interprétation et à la prévision des événements pluvieux sur la base des séries chronologiques issues de modèles de prévision. Il s'agit d'une information fondamentale pour préserver la culture hors sol, les infrastructures ou la production agricole locale comme le cacao. Sodexam a défini un critère de pluviométrie basé sur 4 seuils : pas de pluie, pluie faible, pluie modérée et pluie forte sur la base d'un seuil du cumul journalier de pluie (en mm). Sur la base de ce critère, la problématique de cette étude est d'obtenir un modèle de classification et de prédiction permettant de relier des données facilement accessibles et fiables aux futures observations, en se concentrant sur la caractérisation de la saison des pluies.

Les méthodes d'identification des dates de début et de fin de la saison des pluies peuvent être divisées en deux catégories : les méthodes basées sur la répartition des précipitations (seuil ou bilan hydrique) et celles prenant en compte la dynamique atmosphérique. L'approche par seuil a été donnée initialement par Stern et al. 1981 avec pour définition du début des pluies : la première apparition d'une quantité de pluie déterminée au cours de deux jours successifs. Ce concept a été décliné sous de multiples formes. Wang et LinHo 2002 ont défini le déclenchement de la saison des pluies comme la différence entre la moyenne des précipitations de la pentade (c'est-à-dire 5 jours) et la moyenne du mois le plus sec du climat au cours d'une année spécifique doit dépasser 5 mm. Pour l'Afrique de l'Ouest, Sanogo et al. 2015 ont adapté le seuil à 2,4 mm afin d'obtenir des résultats comparables à un critère local. Toutes ces approches requièrent de paramétrer judicieusement des seuils de pluviométrie. Les autres approches sont fondées sur la circulation atmosphérique d'altitude [Joseph et Moritz 1994 ; Omotosho 1992 ; Hendon et Liebmann 1989]. Elles peuvent être robustes mais nécessitent des données aérologiques pas toujours disponibles. De plus, les prévisions faites par la station Sodexam sont très peu diffusées et souvent incomplètes.

Pour comprendre la dynamique des événements pluvieux, l'approche par *Machine Learning* est une solution alternative. Les méthodes supervisées semblent être un choix rationnel puisque nous disposons d'une base de données labellisées. Cependant, il est difficile d'obtenir un résultat satisfaisant en raison d'une grande confusion entre les pluies faibles et modérées qui se produisent dans les mêmes conditions environnementales. Les fortes pluies sont également difficiles à identifier,

car ce groupe est sous-représenté dans l'ensemble des données. En conséquence, nous proposons une méthode d'identification et de prédiction par le biais de la méthode de classification spectrale non supervisée (*M-SC*) et d'apprentissage semi-supervisé (Random Forest). L'objectif est d'extraire des modèles guidés par la géométrie des données.

Dans cette étude, nous souhaitons obtenir une prévision à $j+1$ jours, $j+7$ jours et $j+30$ jours des niveaux de pluie (Forte pluie, pluie moyenne, faible pluie, pas de pluie) en nous concentrant principalement sur les fortes pluies, catégorie la plus difficile à prédire. Nous appliquons la même méthodologie sur des données de réanalyse. En première partie,les choix de bases de données et de méthodes de traitement sélectionnés pour répondre à cette problématique sont exposés. Ensuite, les résultats de classification et de prédiction sont présentés.

5.2.2 Présentation des bases de données

Les prévisionnistes au sein des *SMN* peuvent se baser sur deux informations météorologiques : les données futures (Figure 5.8) et les données passées (Figure 5.9).

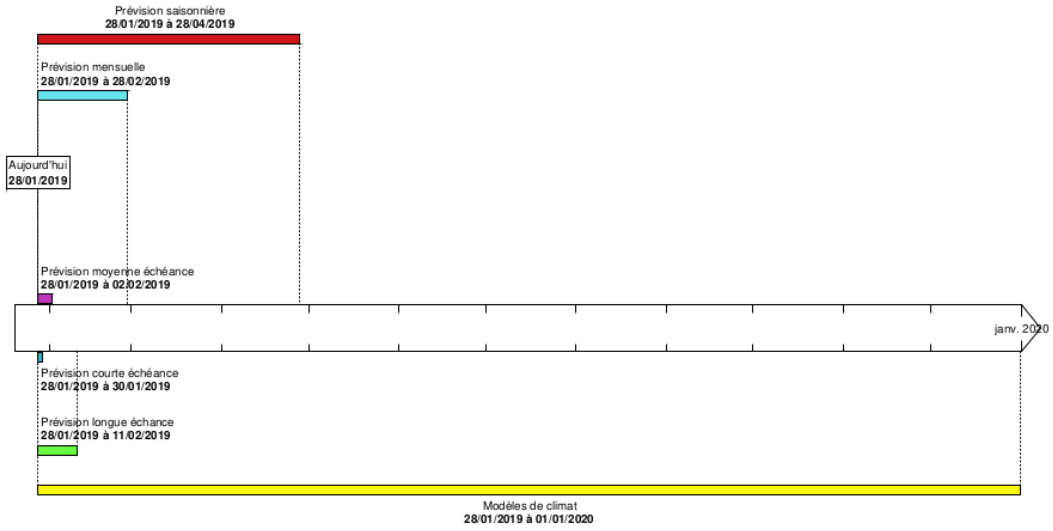


FIGURE 5.8 – Échelle temporelle de chacune des méthodes de prévisions : Données futures.

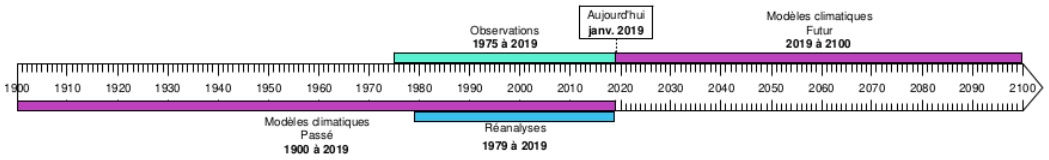


FIGURE 5.9 – Échelle temporelle de chacune des méthodes d'observations : Données passées.

Pour des besoins de prévisions, il semble logique de se tourner vers les données futures. Néanmoins, ces données présentent un certain nombre de contraintes. Il existe plusieurs échéances

de prévisions (Figure 5.8). Les prévisions courtes échéances ($j+2$) sont souvent sur une petite zone du globe et tournent sur des modèles localisés. Les prévisions moyennes ($j+3$ à $j+7$) et longues ($j+15$) échéances sont issues de modèles globaux avec des mailles plus larges et moins précises. Les prévisions mensuelles ($j+30$) permettent d'avoir une première impression du mois à venir, souvent avec des tendances et une idée globale du temps (*i.e.* passage de beaucoup de perturbations, probable vague de froid ...). Pour les prévisions saisonnières, seules les tendances moyennes, *i.e.* température plus chaude ou plus froide que la normale, sont annoncées. Ainsi, plus l'on regarde une prévision pour une date lointaine, moins cette prévision est fiable.

Pour les données du passé (Figure 5.9), il existe trois grands types : les données provenant de modèles de climat, les observations et les réanalyse issues de modèles de prévisions à plus ou moins courtes échéances. Les données de modèle climatique donnent, dans une zone du globe (ou sur le globe), des tendances et non une valeur précise, à différentes échelles : il y a 100 ans, 1000 ans, ...Les observations, à ne pas confondre avec les réanalyses, sont les mesures *in-situ* au point de la station. Les observations sont donc des mesures très locales qui dépendent fortement de l'environnement d'implantation (relief, constructions, ...). Si l'on souhaite avoir une information dans un périmètre autour de la station ou dans une zone sans station, il est donc conseillé d'utiliser les réanalyses. En effet, ces données sont issues de modèles globaux à partir desquels les paramètres sont extraits pour une zone précise (mailles), à une date précise. À la différence des modèles climatiques, c'est une donnée précise et non une tendance. Les réanalyses sont le moyen le plus efficace d'avoir une information constante et sur une zone d'études ou les stations météorologiques sont faiblement représentées. Dans notre cas, c'est l'étude des données passées pour prédire le futur qui nous intéresse. Les données de réanalyse semblent être les plus adaptées. Toutefois, selon la taille de la maille du modèle global, les données peuvent être plus ou moins précises (Figure 5.10). C'est pourquoi il est important de bien choisir le modèle.

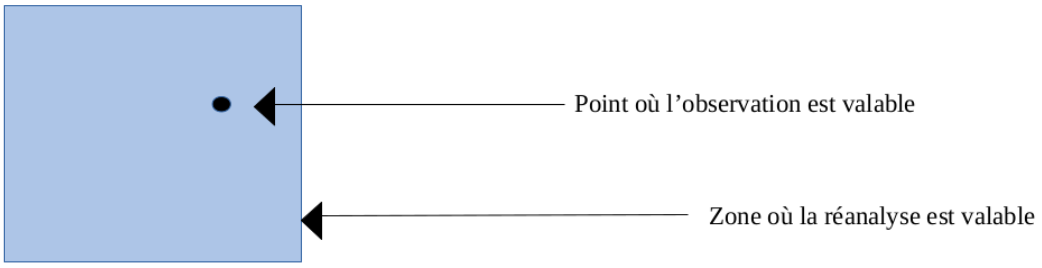


FIGURE 5.10 – Champs d'action d'une observation et d'une réanalyse.

Au niveau Européen, il est possible d'avoir accès, via le Centre Européen de Prévision (CEP-ECMWF), au modèle :

- ERA-Interim est basé sur l'ancien modèle du CPE, avec un maillage de 80 km, un temps de latence de 3 mois pour l'accès aux données, et des données depuis 1979.
- ERA5 est basé sur le nouveau modèle du CPE, avec un maillage de 25 km, avec un temps de latence de 3 mois et des données depuis 1979. 0.25 *degrees* de longitude
- ERA5 T est la nouvelle version de ERA5. Il a été mis en service depuis 2019 et permet d'avoir un temps de latence plus court (7 jours).

Ensuite *National Center for Environmental Prediction* fournit aussi des réanalyses :

- Le modèle CFS avec une maille de 30 km, une disponibilité le lendemain et des données depuis 1979.

Dans notre cas, les réanalyses de ERA5 sont sélectionnées. De ces réanalyses sont extraites 9 mailles de grilles autour d'Abidjan ce qui couvre une surface d'environ 180 km, sur une période de 3 ans (2016 à 2018) avec une mesure toutes les heures (Figure 5.11). Le jeu de données est composé de 15 variables dont 9, définies par un expert en météorologie, sont utilisées comme variables contributives (Tableau 5.3). Toutes les analyses seront réalisées sur un signal défini par concaténation des cumuls de pluie journalière 9 mailles de grilles. La série temporelle est ainsi plus grande et les différences de localisation nous assurent une plus grande variabilité dans le signal.

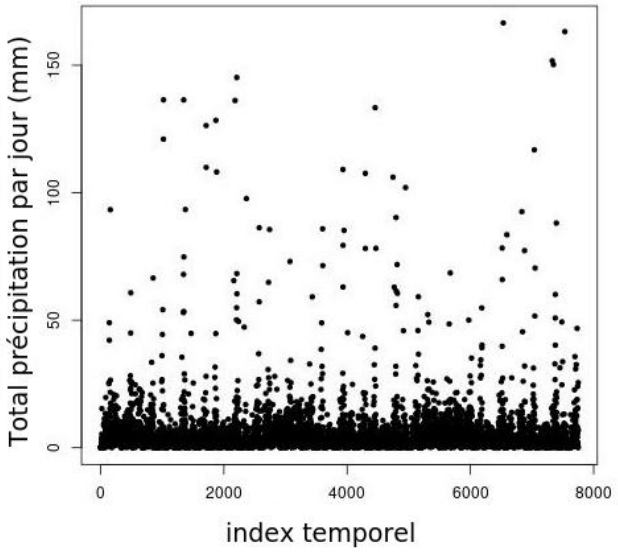


FIGURE 5.11 – Série temporelle définie par concaténation des cumuls de pluie journalière autour d'Abidjan pour les 9 mailles de grille du modèle ERA5. Chaque maille étant mesurée pendant 3 ans, la série temporelle compte donc environ 8 000 observations après concaténation.

TABLEAU 5.3 – Variables du jeu de données ERA5. En rouge les variables contributives (utilisées pour la classification).

Variable	Signification
tp	Cumul des précipitations
tp_std	Ecart-type de l'ensemble de membres du cumul des précipitations
tp_mean	Moyenne de l'ensemble de membres du cumul des précipitations
ctl_tp	Membre de contrôle du cumul des précipitations (Prévision non perturbée)
cape	Energie potentielle convective dont dispose une particule d'air lors de son ascension à partir du niveau de convection libre
cin	Energie d'inhibition convective qu'il faut fournir à une particulepour qu'elle atteigne le niveau de convection libre
gh500	géopotentiel à une pression de 500 hPa
msl	Pression au niveau de la mer
u300	Vent de surface zonal à une pression de 300 hPa
v300	Vent de surface méridional à une pression de 300 hPa
wbpt500	Température potentielle du thermomètre mouillé à 500 hPa
wbpt850	Température potentielle du thermomètre mouillé à 850 hPa
Dates	dates en JJ/MM/AAAA
lat	Latitude du pixel considéré
lon	Longitude du pixel considéré

Ce jeu de données est labellisé en 4 états (Figure 5.12). Pour chaque donnée journalière est attribuée un label de 1 à 4 suivant les critères de pluies fournis par la *SMN* de Côte d'Ivoire. Ces critères définissent 4 intensités en fonction du cumul de pluie journalier de pluie (tp en mm) :

1. Pas de pluie :$tp <= 5$
2. Faible pluie : $5 < tp <= 10$
3. Pluie moyenne : $10 < tp <= 50$
4. Forte pluie : $tp > 50$

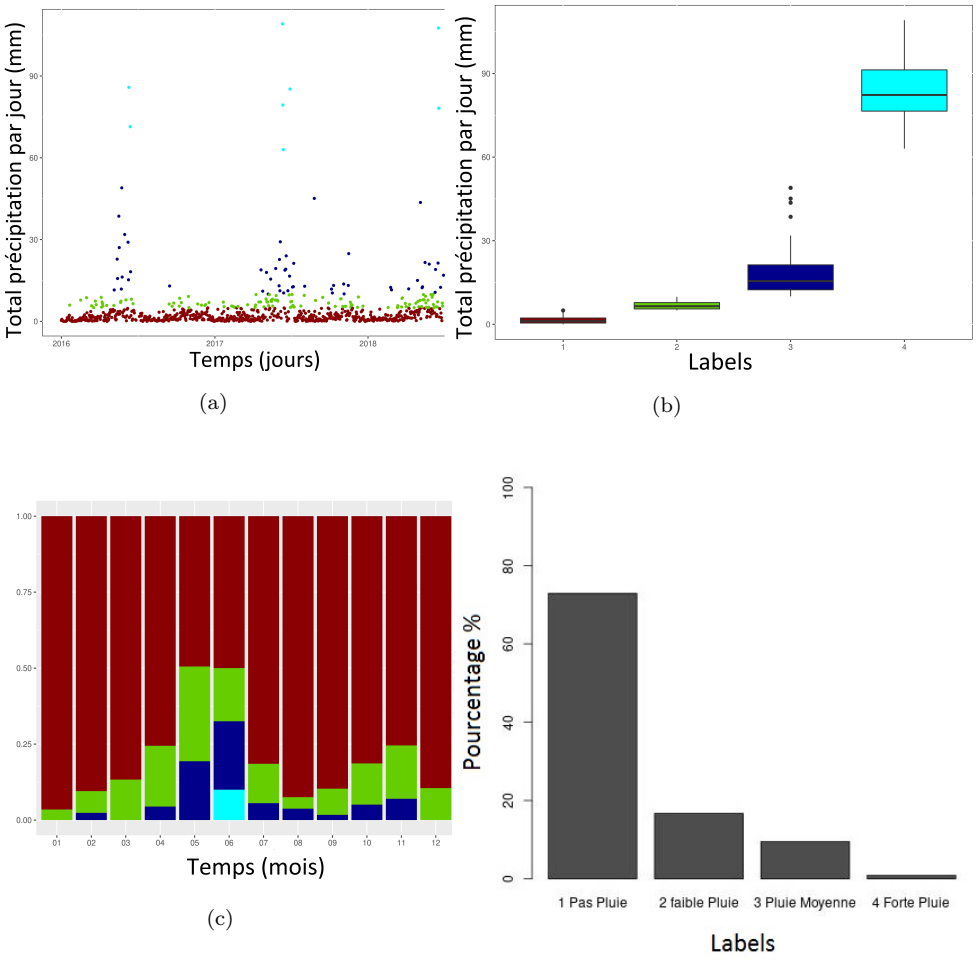


FIGURE 5.12 – Labellisation des données ERA5 sur la période 2016-2018. (a) État de pluie reporté sur le signal de pluie (mm/j), (b) Dispersion du signal de pluie par classe (boîte de dispersion), (c) Fréquence d'occurrence (c) de chaque état par mois, (d) Répartition de chaque état en pourcentage.

Les 4 états ne sont pas équirépartis (Figure 5.12d). Même si la série temporelle a été multipliée par 9 lors de la concaténation des 9 mailles de grille, le label 4 (Forte pluie) reste sous-représenté dans l'ensemble de la base. Ce label est celui que nous souhaitons prédire en priorité. Il est donc nécessaire de prendre en considération cette distribution particulière lors de notre analyse.

5.2.3 Calibration de la méthode de classification M-SC

La méthode M-SC a été calibrée pour répondre au mieux à la problématique de détection de la saison des pluies et notamment, aux problématiques liées à la difficulté de détecter des événements rares, extrêmes et de courtes durées. Dans ce cas d'étude, c'est le label 4 (Forte pluie) qui sera le plus difficile à identifier et à caractériser. En effet, le label 4 est sous-représenté par rapport aux autres (Figure 5.12d). De plus, les événements qui le composent, sont de courtes durées, d'environ un jour à trois jours maximum. Ils comptent donc un faible nombre de points.

Pour mieux considérer ces caractéristiques sur la distribution et la taille des classes, nous avons tout d'abord choisi de ne plus effectuer la mise à zéro de la diagonale de la matrice de similarité W (ligne 5 algorithme 9 : $w_i i = 0$). Le passage à zéro de la diagonale a pour but d'empêcher qu'une classe soit constituée seulement du point lui-même et évite donc la caractérisation de classes isolées. Or il est possible, pour le cas du label 4, de rencontrer des événements isolés. L'étape est donc supprimée. Nous nous assurons ainsi que les événements isolés ne seront pas éliminés de la classification.

Ensuite, pour les mêmes raisons, le critère de silhouette (sil.min) (Section 3.2.3.2) est augmenté de 0,7 à 0,9. L'augmentation du critère de silhouette induit la recherche de classes plus connexes et donc plus compactes. Cela a pour conséquences indirectes d'augmenter la segmentation et de permettre d'accepter des classes de tailles plus petites.

Enfin, le critère de coupe (crit) choisi est la méthode du GAP. Pour rappel, le critère de coupe permet d'estimer automatiquement un nombre de classes (Section 3.2.3.1). La technique du GAP est préférée dans ce cas de figure, car elle est moins restrictive que la méthode PEV et permet d'augmenter légèrement le nombre de classes. L'étude étant exploratoire et la saison des pluies étant plus rare d'un point de vue dynamique, il est plus judicieux d'élargir le champ de la recherche.

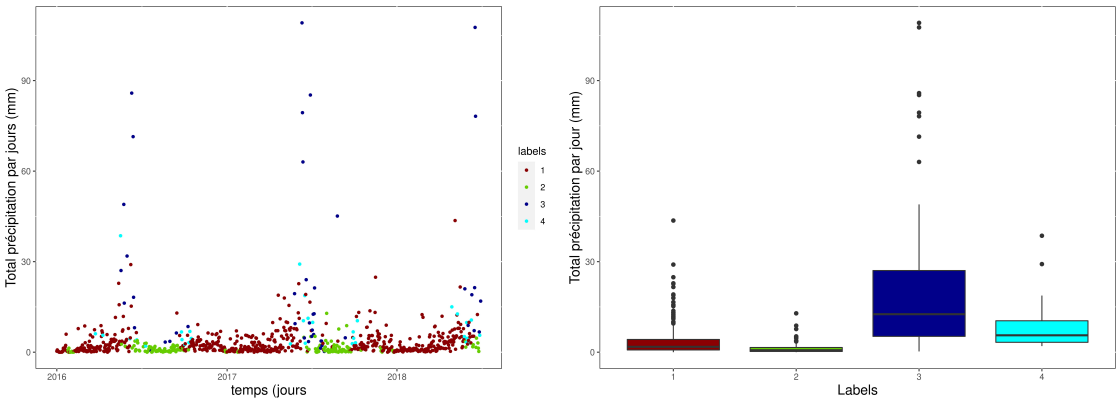
5.2.4 Résultats de la classification

Dès les premiers niveaux de classification, une bonne séparation des classes "avec et sans fortes pluies" se fait. Au niveau 3 (Figure 5.13), le label "4-Forte pluie" est isolé en grande partie dans le classe 3 (cl3, en bleu) (Figure 5.13b). Toutefois, si nous comparons les 4 états de pluies et les classes issues de la classification par *M-SC* sur les 9 mailles de grille (Tableau 5.4), il est possible de constater une confusion encore importante. En effet, sur les 64 points labellisés "Forte pluie" 46 sont identifiés dans cl3, mais 15 sont identifiés dans la classe 1 (cl1, en rouge) et 7 dans la classe 4 (cl4, en cyan). De plus, si l'on considère cl3 comme la classe correspondant aux fortes pluies, elle contient un nombre relativement conséquent de faux positifs et notamment un grand nombre de points définis comme "Pluie moyenne". Néanmoins, cette nouvelle classification reste intéressante. En effet, ce nombre de faux positifs est dû principalement à la différence entre les deux manières de classer les données. En réalité, la labellisation via le cumul journalier de pluie uniquement est une approche "seuillée" où la limitation des classes se fait via un seuil fixe. Alors que dans le cas de *M-SC*, l'approche est multivariée et spectrale, la rendant indépendante de la forme de la distribution des paramètres. Ainsi, la méthode n'offre pas seulement une hiérarchisation des classes par valeurs (hautes ou faibles) mais permet de définir un dynamique complète. Pour la classe cl3, ce sont non pas des observations dont les valeurs de pluviométrie sont hautes mais des pics entiers détectés : début de croissance, maintien et fin (Figure 5.13a). Cette dynamique se retrouve aussi temporellement. Avec l'approche seuillée, les fortes pluies sont identifiées seulement au mois de juin et tous les autres états sont identifiés pour tous les mois de l'année de manière proportionnelle à leur pourcentage d'apparition (Figure 5.12c et 5.12d). Avec l'approche *M-SC*, une dynamique temporelle est mise en évidence : (i) des évènements de type "moyenne et forte

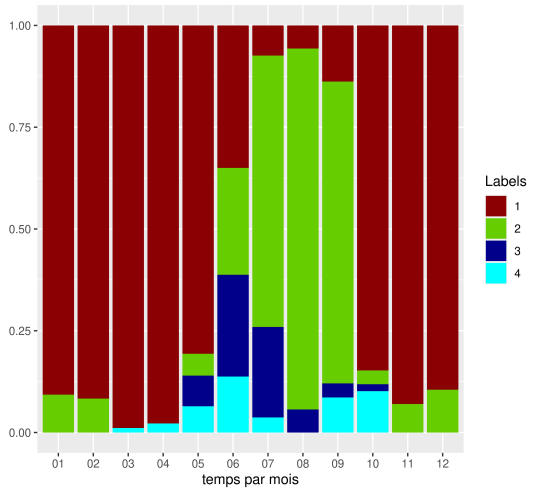
pluie" (cl3 et cl4) sont identifiés entre le mois de mai et d'août et (ii) un contraste "pas de pluie et faible pluie" (cl1 et cl2) est identifiés pour le reste de la période (Figure 5.13c).

TABLEAU 5.4 – Tableau de contingence entre les états définis manuellement à partir des seuils de cumuls de pluie et les classes obtenues par classification spectrale (M-SC) au niveau 3

Label Sodexam vs label MSC	1	2	3	4
1 : Pas de pluie	3885	1641	44	81
2 : Faible pluie	983	163	78	70
3 : Pluie moyenne	459	74	111	92
4 : Forte pluie	15	0	46	7



(a) (b)



(c)

FIGURE 5.13 – Classification au niveau M-SC 3 des données ERA5 sur la période 2016-2018 (a) reportée sur le signal de pluie (mm/j). (b) Dispersion du signal de pluie par classe (boîte de dispersion). (c) Fréquence d'occurrence de chaque état par mois.

Des niveaux de classification plus profonds sont calculés de façon à augmenter le niveau de détails et à obtenir une meilleure caractérisation des 4 états de pluies. Au niveaux 6, 32 classes sont définies. À ce niveau, la confusion et surtout le nombre de faux positifs par classe ont très nettement diminué (Tableau 5.5). La diminution de la confusion permet d'isoler par vote majoritaire 4 classes contre 1 ou 2 classes isolées au niveau 3 (Tableau 5.6). Pour rappel dans ce cas de figure, le vote majoritaire consiste à attribuer pour chaque classe un des 4 états en fonction du nombre le plus grand défini par la table de confusion. À partir de ce principe, il est donc possible pour chaque classe d'établir une correspondance avec les 4 états de pluie (Forte, moyenne, faible, ou nulle) définis au début de l'étude (Tableau 5.7). Cette correspondance est facilement identifiable sur la figure 5.14. Ainsi, les classes 19, 20 et 21 correspondent à l'état "4-Forte pluie". Les classes de 17 à 32 (sans 19, 20 et 21) sont caractéristiques de l'état "3-Moyenne pluie", les classes 6 à 9 de l'état "2-Faible pluie" et les autres classes de l'état "1-Pas de pluie".

Tableau 5.5 – Tableau de contingence entre les états définis manuellement à partir des seuils de cumuls de pluie et les classes obtenues par classification spectrale (M-SC) au niveau 6.

True vs MSC	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
1 : Pas de pluie	1113	18	83	9	9	1156	695	428	374	366	254	225	489	253	45		
2 : Faible pluie	160	0	0	0	0	164	297	213	149	107	45	0	10	1	0		
3 : Pluie moyenne	59	0	0	0	0	30	147	152	71	37	36	0	1	0	0		
4 : Forte pluie	0	0	0	0	0	0	0	10	5	0	0	0	0	0	0		
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1 : Pas de pluie	9	4	15	2	0	1	8	6	8	2	2	18	16	15	18	0	10
2 : Faible pluie	0	33	12	4	2	5	10	12	0	1	10	5	8	14	11	9	12
3 : Pluie moyenne	0	47	25	12	6	5	8	8	0	7	6	19	26	17	5	7	5
4 : Forte pluie	0	6	8	18	7	5	1	1	0	0	0	2	2	1	0	2	0

Tableau 5.6 – Tableau de contingence entre les états définis manuellement à partir des seuils de cumuls de pluie et les classes obtenues par classification spectrale (M-SC) au niveau 6 après vote majoritaire.

Label Sodexam vs Label MSC	1	2	3	4
1 : Pas de pluie	5552	27	70	2
2 : Faible pluie	1157	58	73	6
3 : Pluie moyenne	538	39	141	18
4 : Forte pluie	15	9	19	25

Tableau 5.7 – Tableau de correspondance entre les états définis manuellement à partir des seuils de cumuls de pluie et les classes obtenues par classification spectrale (M-SC) au niveau 6.

Etats de pluie	Classes M-SC
1 : Pas de pluie	1 à 6 et 10 à 17
2 : Faible pluie	17, 18 et 22 à 32
3 : Pluie moyenne	6,7,8,9
4 : Forte pluie	19, 20, 21

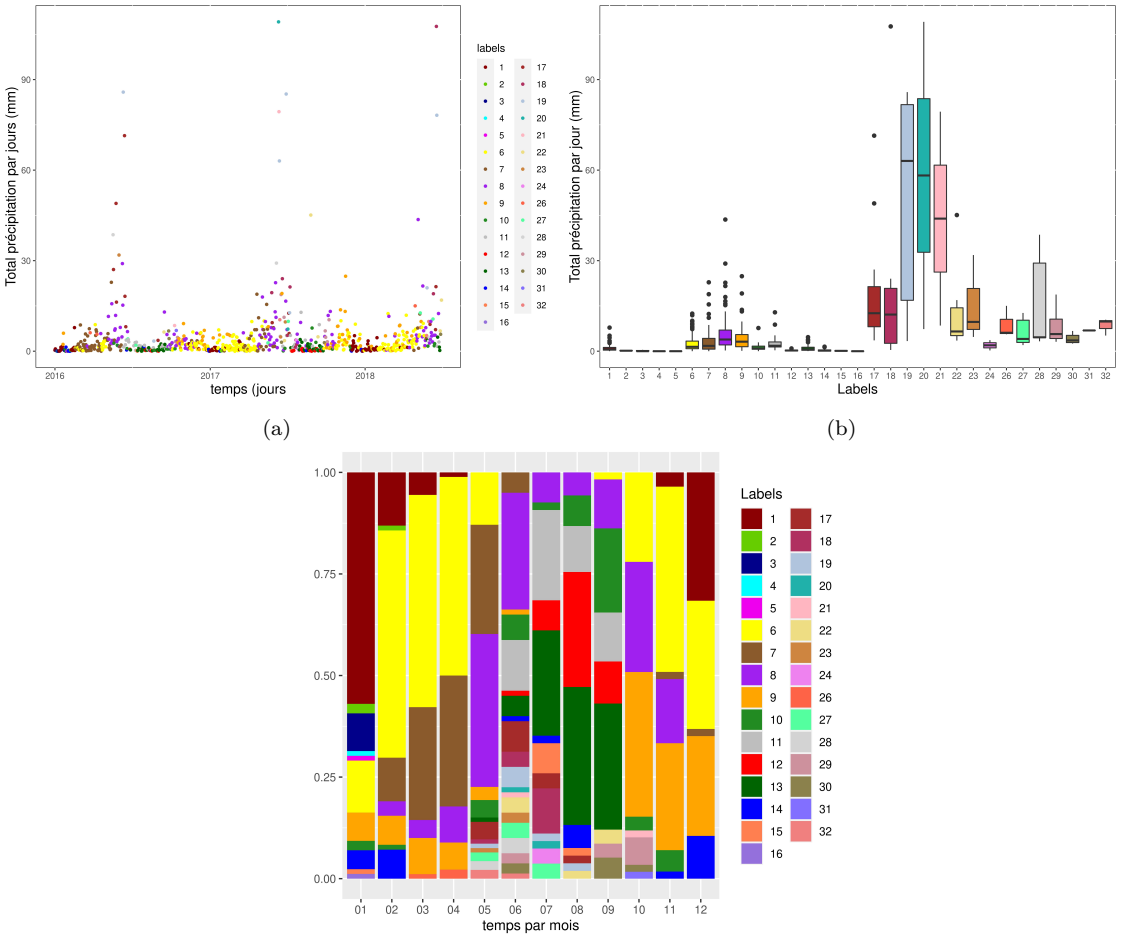


FIGURE 5.14 – Classification au niveau M-SC 6 des données ERA5 sur la période 2016-2018 (a) reportées sur le signal de pluie (mm/j). (b) Dispersion du signal de pluie par classe (boîte de dispersion). (c) Fréquence d'occurrence de chaque état par mois.

5.2.5 Paramétrage du modèle d'apprentissage

La méthode d'apprentissage *Random Forest* [BREIMAN 1999] est sélectionnée pour répondre à la problématique de prédiction des états de pluie. Elle a été choisie car c'est une méthode robuste, rapide et efficace lorsque le jeu de données est petit, comme dans notre cas de figure.

Pour rappel Random Forest est une méthode supervisée basée sur la construction d'une multitude d'arbres de décision ("Forêt d'arbre") qui combinés, forment un modèle robuste. Chaque arbre de la forêt est construit sur une fraction ("*in bag*") des données d'entraînements originales ce qui génère de petits sous-ensembles de données constitués de la nouvelle fraction et du restant ("*out of bag*"). Chaque arbre est entraîné sur ces échantillons. Le résultat final du modèle d'ensemble est déterminé par un vote majoritaire de tous les arbres de décision.

Le modèle d'apprentissage est construit sur la base de 100 arbres ($ntree = 100$) d'une profondeur de 5 niveaux ($mtry = 5$). Le nombre d'arbres est choisi afin de minimiser le taux

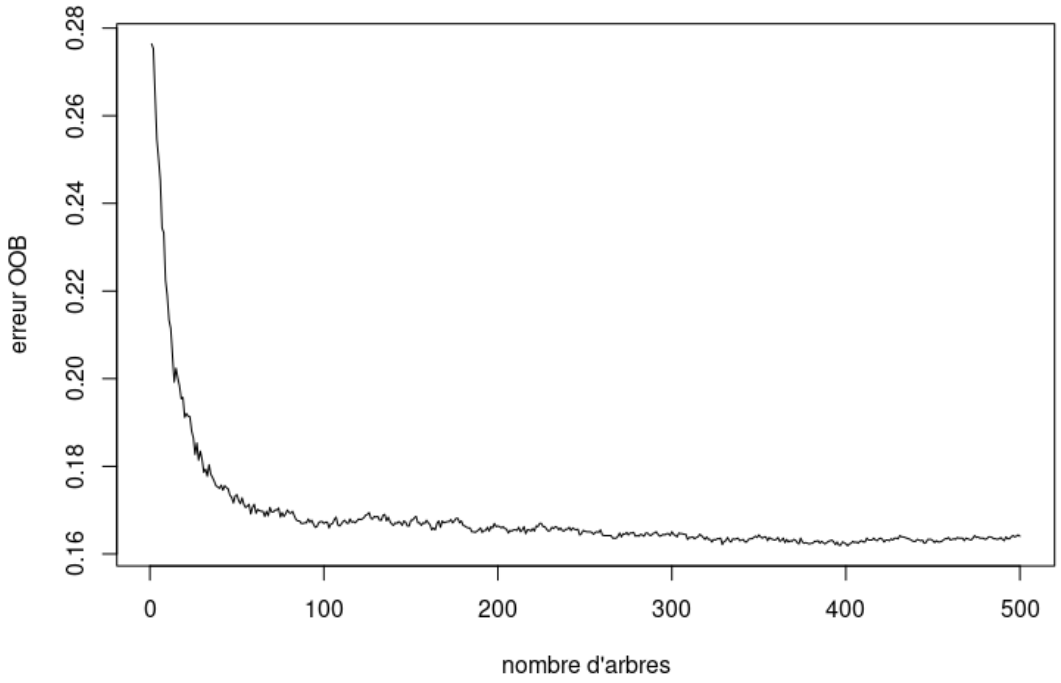


FIGURE 5.15 – Taux d'erreur Out Of Bag (OOB) en fonction du nombre d'arbres de décision.

d'erreur *Out Of Bag* (OOB). L' OBB est l'estimation acceptable du taux d'erreur théorique obtenu grâce au bootstrap. 100 est une valeur d'arbres où l'erreur est faible et commence à se stabiliser (Figure 5.15). C'est aussi une valeur assez petite, ce qui permet de limiter le temps de calcul. La profondeur de chaque arbre est choisie supérieure à la racine carrée du nombre de variables soit supérieure à $\sqrt{9}$.

5.2.6 Construction de la base d'entraînement

La base d'apprentissage est construite à partir des mêmes variables que celles utilisées pour la classification et des classes issues de *M-SC* au niveau 6. Le système d'apprentissage supervisé crée une relation entre les variables d'entrées et de sorties souhaitées, ici les classes *M-SC* ou les états de pluies rattachés. Ainsi le système d'apprentissage sera capable d'attribuer les dites classes à un nouveau fichier : entrée non classée. Or nos classes sont relatives à un état à l'instant j et nous souhaitons prédire cet instant plusieurs jours à l'avance. Cela signifie que nous souhaitons connaître pour une certaine distribution des variables explicatives du modèle à l'instant j : quelle sera la classe à $j+x$ jours. Pour répondre à cette interrogation, un décalage temporel est opéré sur le vecteur de classes. Ainsi, nous avons créé les trois nouveaux vecteurs de classes ($Classej+1$, $Classej+7$ et $Classej+30$) avec un décalage temporel de $x=1,7$ et 30 jours (Figure 5.16). Ainsi, pour une observation (i), la $Classej_i$ correspond à la classe définie par *M-SC* pour l'instant j, soit la classe à laquelle elle appartient, et la $Classej+30_i$ correspond à la classe définie par *M-SC* pour l'observation $i+30$, soit la classe à laquelle appartient l'observation $i+30$ (Décalage

temporel en vert sur la figure 5.16). De même les $Classej+1_i$ et $Classej+7_i$ correspondent à la classe de l'observation $i+1$ et $i+7$ (Décalage temporel en rouge et en bleu sur la figure 5.16). Ces trois nouveaux vecteurs permettent d'effectuer une prédiction à un jour, une semaine et un mois.

var1	var2	var3	...	Classe J	Classe J+1	Classe J+7	Classe J+30
2 408,79	72,81	0,36		1	3	3	15
825,14	64,74	0,12		3	3	3	14
721,48	66,36	0,22		3	3	3	14
263,36	63,75	0,08		3	3	1	14
676,40	68,68	0,03		3	3	1	14
0,00	60,73	0,00		3	4	1	14
0,00	62,76	0,00		4	3	1	14
0,00	63,09	0,00		3	3	1	14
254,49	60,83	0,00		3	3	1	10
922,96	68,86	0,01		3	1	1	14
1 253,94	76,50	0,04		1	1	1	14
2 239,25	80,38	0,47		1	1	1	1
2 583,98	82,55	0,78		1	1	7	1
3 281,97	80,24	2,18		1	1	1	1
4 456,98	83,57	1,87		1	1	1	1
3 820,51	81,66	3,61		1	1	3	1
4 191,99	82,55	3,73		1	1	3	1
3 887,98	82,35	5,32		1	1	14	1
2 502,49	82,44	1,80		1	7	12	1
3 970,49	85,54	1,43		7	1	14	1
3 476,28	85,11	15,36		1	1	3	1
2 784,61	77,34	0,39		1	3	5	1
1 397,36	57,44	0,06		3	3	16	7
1 719,62	61,80	0,21		3	14	15	1
1 878,15	74,50	0,56		14	12	14	1
2 173,51	68,23	0,03		12	14	14	1
1 213,26	65,34	0,27		14	3	14	1
412,99	65,58	0,02		3	5	14	7
8,26	56,47	0,00		5	16	14	1
0,00	62,43	0,00		16	15	14	7
17,12	65,09	0,01		15	14	14	7
				...	...	...	...

FIGURE 5.16 – Extrait du fichier d'apprentissage sur 31 jours avec les variables explicatives (var1, var2, var3) et les vecteurs de sortie (Classe J, Classe J+1, Classe J+7, Classe J+3). Les cases rouges représentent le décalage effectué à J+1, les cases vertes à J+7 et les bleues à J+30.

Ensuite la base d'apprentissage est découpée en deux parties : une base d'apprentissage et une base de test. Dans cette étude, elles ne sont pas construites de manière aléatoire comme c'est généralement le cas. Le modèle d'apprentissage est entraîné sur une base d'apprentissage originale composée des 8 mailles de grilles extérieures. Il est ensuite appliqué sur la base test composée de la maille centrale (Figure 5.17).

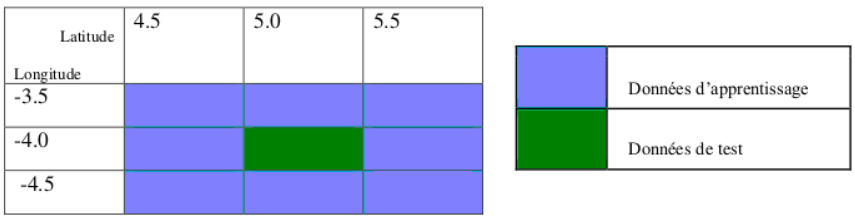


FIGURE 5.17 – Schéma de la répartition des données d'apprentissage (en cyan) et des données de test (en vert) selon la longitude et la latitude.

5.2.7 Résultats de prédiction

Un modèle de prédiction est donc construit à partir de la base d'apprentissage. L'approche usuelle d'évaluation consiste à estimer, dans un premier temps, les performances du modèle réalisées à partir de la base d'entraînement puis de déterminer l'efficacité de prédiction sur la base test. Pour cela, nous utilisons 4 indices de performance qui confrontent les classes prédites et les classes cibles ("Vrai") (Section 2.4.4.2).

5.2.7.1 Évaluation du modèle d'apprentissage

Le taux de reconnaissance des trois modèles de prédiction créés à partir des bases d'entraînement avec les classes cibles : $Classe j+1$ ou $Classe j+7$ ou $Classe j+30$ jours, est de 1. Nos modèles sont donc capables de prédire 100 % des données. Un score de 100 % indique que le modèle prédictif s'adapte bien à la base de données d'entraînement, même un peu trop. En effet, dans ce cas, le modèle prédictif prend en compte toutes les variations (corrélations généralistes et bruits) de la base d'entraînement. Le modèle se généralise mal, c'est-à-dire, qu'il ne sera performant que sur des jeux de données qu'il connaît déjà ; par exemple, les mêmes mailles de grilles l'année d'après, mais qu'il aura plus de mal à faire une prédiction sur une nouvelle base de données comme une autre zone géographique. Ce score peut s'expliquer en partie par le nombre de classes importantes et la façon dont sont construites ces classes. En effet, la classification par *M-SC* au niveau 6 est déjà performante et les scores de précision sont forts, ce qui facilite la prédiction. Dans notre cas, l'objet de l'étude est très localisé, cela n'est donc pas vraiment dérangeant, mais c'est un critère à prendre en considération si le modèle veut être appliqué ailleurs.

5.2.7.2 Évaluations de la prédiction des classes

Le taux de reconnaissance des modèles de prédiction, réalisés sur la base de données test, est supérieur à 0,90 pour les trois modèles. Les tableaux 5.11, 5.12 et 5.13 permettent de mesurer respectivement la qualité des systèmes de prédiction pour $j+1$, $j+7$ et $j+30$ jours. Chaque ligne correspond à une classe cible, par exemple $Classe j+1$ et chaque colonne correspond à une classe prédite par *Randum Forest*. Pour les trois simulations, il y a très peu de confusion entre les classes, le système de prédiction parvient donc à classifier correctement les classes. Ces résultats se retrouvent aussi au niveau des indices de performance (Tableau 5.8, 5.9, 5.10). Par exemple, les classes 19 à 21, qui correspondent à l'état de forte pluie, sont très bien prédites. Seule la classe 19 présente quelques taux négatifs. Il en est de même pour la majorité des classes. Toutefois, la classe 25 est mal ou pas prédite à $j+1$ comme à $j+30$. Mais cette classe ne contient que 2 points, ce qui peut la rendre difficile à prédire.

Que ce soient pour les classes $j+1$ ou les classes $j+7$ et $j+30$, il n'y a pas de baisse de performance du modèle de prédiction. Un décalage d'un mois peut sembler grand, mais le modèle est très stable, avec pour rappel un taux de reconnaissance de 100 % pour la base d'entraînement. Ainsi, le modèle est capable de fournir une prédiction correcte des classes *M-SC* à 1 mois. Mais ces classes ne sont pas les états de pluies recherchés initialement. C'est pourquoi un deuxième modèle de prédiction a été construit à partir des correspondances entre les états de pluie et les classes du niveau 6 (Tableau 5.7).

Tableau 5.8 – Indices de performance pour chaque classe du système de prédiction à $j+1$. Precision est relatif au taux de faux positifs. Recall est le rapport entre les vrais positifs correctement prédits et toutes les observations de la classe cible. F1 est la moyenne pondérée de Precision et Recall.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Precision	0,99	1,00	0,89	1,00	1,00	0,94	0,97	0,88	0,90	0,86	0,82	1,00	0,98	1,00	1,00	1,00
Recall	0,98	1,00	1,00	1,00	1,00	0,98	0,90	0,84	0,93	0,94	0,91	0,96	0,95	1,00	1,00	1,00
F1	0,98	1,00	0,94	1,00	1,00	0,96	0,93	0,86	0,91	0,90	0,86	0,98	0,96	1,00	1,00	1,00

	17	18	19	20	21	22	23	24	26	27	28	29	30	31	32
Precision	0,92	0,82	1,00	1,00	0,67	1,00	0,67	1,00	1,00	1,00	0,83	0,75	1,00	1,00	0,60
Recall	0,85	0,90	0,71	1,00	1,00	0,50	0,67	1,00	0,67	0,71	1,00	0,75	0,83	1,00	1,00
F1	0,88	0,86	0,83	1,00	0,80	0,67	0,67	1,00	0,80	0,83	0,91	0,75	0,91	1,00	0,75

Tableau 5.9 – Indices de performance pour chaque classe du système de prédiction à $j+7$. Precision est relatif au taux de faux positifs. Recall est le rapport entre les vrais positifs correctement prédits et toutes les observations de la classe cible. F1 est la moyenne pondérée de Precision et Recall.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Precision	0,99	1,00	0,89	1,00	1,00	0,96	0,97	0,89	0,90	0,85	0,75	0,92	0,94	1,00	1,00	1,00
Recall	0,98	1,00	1,00	1,00	1,00	0,99	0,95	0,87	0,88	0,97	0,86	0,92	0,91	0,91	1,00	1,00
F1	0,98	1,00	0,94	1,00	1,00	0,98	0,96	0,88	0,89	0,90	0,80	0,92	0,93	0,95	1,00	1,00

	17	18	19	20	21	22	23	24	26	27	28	29	30	31	32
Precision	0,83	1,00	0,71	1,00	1,00	0,80	0,67	0,67	1,00	1,00	0,71	0,86	1,00	1,00	0,50
Recall	0,77	0,70	0,71	1,00	1,00	0,67	0,67	1,00	0,67	0,71	1,00	0,75	0,83	1,00	1,00
F1	0,80	0,82	0,71	1,00	1,00	0,73	0,67	0,80	0,80	0,83	0,83	0,80	0,91	1,00	0,67

Tableau 5.10 – Indices de performance pour chaque classe du système de prédiction à $j+7$. Precision est relatif au taux de faux positifs. Recall est le rapport entre les vrais positifs correctement prédits et toutes les observations de la classe cible. F1 est la moyenne pondérée de Precision et Recall.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Precision	0,99	1,00	0,89	1,00	1,00	0,96	0,96	0,91	0,94	0,89	0,80	0,92	0,96	0,96	1,00	1,00
Recall	0,98	0,67	1,00	1,00	1,00	0,99	0,90	0,88	0,91	1,00	0,94	0,92	0,89	1,00	1,00	1,00
F1	0,98	0,80	0,94	1,00	1,00	0,97	0,93	0,90	0,92	0,94	0,87	0,92	0,92	0,98	1,00	1,00

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Precision	0,92	1,00	1,00	1,00	1,00	1,00	0,75	1,00	0,00	1,00	0,71	0,83	0,88	1,00	1,00	0,40
Recall	0,85	0,80	0,86	1,00	1,00	0,83	1,00	1,00	NA	0,67	0,71	1,00	0,88	0,83	1,00	0,67
F1	0,88	0,89	0,92	1,00	1,00	0,91	0,86	1,00	NA	0,80	0,71	0,91	0,88	0,91	1,00	0,50

TABLEAU 5.11 – Tableau de contingence pour les résultats à $j+1$. Chaque ligne correspond à une classe cible déterminée par M-SC et décalée d'un jour ($Classe j+1$), et chaque colonne correspond à une classe prédite par Randum Forest.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	84	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	215	1	0	2	0	0	0	0	0	0	0
7	0	0	0	0	0	6	85	2	1	0	0	0	0	0	0	0
8	0	0	0	0	0	3	2	94	5	1	6	0	0	0	0	0
9	0	0	0	0	0	4	0	1	74	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	32	1	0	0	0	0	0
11	0	0	0	0	0	0	0	1	0	2	32	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	25	1	0	0	0
13	0	0	0	0	0	0	0	1	0	2	0	0	52	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	11	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
18	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0
23	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2
27	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Tableau 5.12 – Tableau de contingence pour les résultats à $j+7$. Chaque ligne correspond à une classe cible déterminée par M-SC et décalée d'une semaine ($Classe j+7$), et chaque colonne correspond à une classe prédite par Randum Forest.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	84	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	216	1	0	2	0	0	0	0	0	0	0
7	0	0	0	0	0	2	89	2	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	2	97	4	0	6	0	0	0	0	0
9	1	0	0	0	0	5	0	4	70	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	33	1	0	0	0	0	0
11	0	0	0	0	0	0	0	1	0	4	30	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	24	2	0	0	0
13	0	0	0	0	0	0	0	0	0	2	2	1	50	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	1	20	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	10	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0
18	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	0	5	0	0	1	0	0	0	0	0	0	0	0	0	0
20	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2
27	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

TABLEAU 5.13 – Tableau de contingence pour les résultats à $j+30$. Chaque ligne correspond à une classe cible déterminée par M-SC et décalée d'un mois ($Classej+30$), et chaque colonne correspond à une classe prédite par Randum Forest.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	84	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	216	1	0	1	0	0	0	0	0	0	0
7	0	0	0	0	0	4	85	4	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	3	99	3	0	5	0	0	0	0	0
9	0	0	0	0	0	4	0	1	73	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0
11	0	0	0	0	0	0	0	1	0	1	33	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	24	2	0	0	0
13	0	0	0	0	0	0	0	0	0	2	2	2	49	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	11	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
18	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	1	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	5	1	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2
27	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
31	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2

5.2.7.3 Évaluation de la prédiction des états de pluie

À partir de la classification *M-SC*, nous avons pu établir une correspondance entre les classes et les états de pluie (Forte, moyenne, faible, ou pas de pluie) définis au début de l'étude. Ces correspondances sont décrites dans le tableau 5.7 (section 5.2.4). Afin de faire un comparatif cohérent entre les états de pluies et les résultats de prédiction, nous avons remplacé les classes du niveau 6 prédites par leurs correspondances.

Les résultats à $j+1$ (Tableau 5.14), $j+7$ (Tableau 5.15) et $j+30$ (Tableau 5.16) sont assez similaires. Ceci démontre que le protocole de labellisation, via la correspondance des classes *M-SC* et de prédiction, via un *Randum Forest* est aussi stable. Tous comme la prédiction des classes du niveau 6, elle permet une prédiction aussi performante pour 1 jour que pour 30 jours. La confusion pour les classes "2- faible pluie", "3-moyenne pluie" et "4-forte pluie" reste assez élevée, mais elle est plus faible que la confusion faite par *M-SC* au niveau 3 (Section 5.2.4). La méthode M-SC à 6 niveaux, puis la mise en correspondance, ne faussent donc pas les résultats et permettent même de les améliorer. La classe "2-faible pluie", avec un Recall de 0,04 et un score F1 de 0,08 est la plus mal prédite, elle est souvent confondue avec la classe "1-Pas de pluie". Cette confusion s'explique notamment car les classes prédites, issues de classification par *M-SC*, sont comparées avec les états de pluie déterminés par approche "seuillée". Comme nous avions pu déjà le constater à la section 5.2.4, cela peut biaiser le résultat. De plus, pour les correspondances se sont les labels "1-pas de pluie" et "2-faible pluie" qui comptent le plus grand nombre de classes issues de la classification *M-SC* au niveau 6, ce qui augmente aussi cette confusion. Certaines de ces classes peuvent être proches et donc mal prédites. Au regard des résultats de détection (Section 5.2.4), le label "4-forte pluie" avec un Recall de 0,33 et un score F1 de 0,39 est relativement bien prédit. En effet, ce label reste le plus dur à détecter, il est donc sensément plus difficile à prédire.

Tableau 5.14 – Prédiction à $j+1$ des états de pluies. (A) Tableau de contingence entre les états de pluie définis manuellement à partir des seuils de cumuls de pluie et les correspondances issues de la classification M-SC au niveau 6. (B) Indices de performance pour chaque classe. Precision est relatif au taux de faux positifs. Recall est le rapport entre les vrais positifs correctement prédits et toutes les observations de la classe cible. F1 est la moyenne pondérée de Precision et Recall.

(A)

Label Sodexam vs Label MSC	1	2	3	4
1	4893	24	54	1
2	1055	49	65	5
3	507	34	124	16
4	15	8	17	20

(B)

	1	2	3	4
Precision	0,76	0,43	0,48	0,48
Recall	0,98	0,04	0,18	0,33
F1	0,86	0,08	0,26	0,39

Tableau 5.15 – Prédiction à $j+7$ des états de pluies. (A) Tableau de contingence entre les états de pluie définis manuellement à partir des seuils de cumuls de pluie et les correspondances issues de la classification M-SC au niveau 6. (B) Indices de performance pour chaque classe. Precision est relatif au taux de faux positifs. Recall est le rapport entre les vrais positifs correctement prédits et toutes les observations de la classe cible. F1 est la moyenne pondérée de Precision et Recall.

(A)

Label Sodexam vs Label MSC	1	2	3	4
1	4887	24	54	1
2	1055	49	65	5
3	507	34	124	16
4	15	8	17	20

(B)

	1	2	3	4
Precision	0,76	0,43	0,48	0,48
Recall	0,98	0,04	0,18	0,33
F1	0,86	0,08	0,26	0,39

TABLEAU 5.16 – Prédiction à $j+30$ des états de pluies. (A) Tableau de contingence entre les états de pluie définis manuellement à partir des seuils de cumuls de pluie et les correspondances issues de la classification M-SC au niveau 6. (B) Indices de performance pour chaque classe. Precision est relatif au taux de faux positifs. Recall est le rapport entre les vrais positifs correctement prédits et toutes les observations de la classe cible. F1 est la moyenne pondérée de Precision et Recall.

(A)

Label Sodexam vs Label MSC	1	2	3	4
1	4865	24	54	1
2	1055	49	65	5
3	506	34	124	16
4	15	8	17	20

(B)

	1	2	3	4
Precision	0,76	0,43	0,48	0,48
Recall	0,98	0,04	0,18	0,33
F1	0,85	0,08	0,26	0,39

5.2.8 Conclusions et perspectives

La classification spectrale non supervisée *M-SC* extrait un modèle de classification guidé par la géométrie des données. Elle permet d'identifier des états caractéristiques indépendants des seuils de précipitations et d'aborder la problématique de détection des pluies avec une approche plus phénologique. Ainsi, dès les premiers niveaux, une classification cohérente avec la mousson et la saison des pluies est observée. Chaque classe peut être associée à un état de pluie et constitue un phénomène d'ensemble (début de croissance, maintien et fin).

Toutefois, l'inégalité de distribution entre les états rend la détection et la prédiction des fortes pluies difficile. Les niveaux plus profonds augmentent la précision de détection de l'état "4-Forte pluie", mais aussi le nombre de classes. Mais, elle reste assez pertinente pour fournir un jeu de données labellisé pour faire de la prédiction. La combinaison d'une méthode de classification supervisée et du décalage temporel après classification non supervisée est une approche non conventionnelle, mais les résultats ont démontré que c'était une méthode stable qui semble fournir des prévisions correctes avec un mois de décalage temporel.

Pour approfondir et confirmer cette solution, il aurait été intéressant d'essayer de faire une prédiction à partir du modèle construit par Random Forest et de la comparer aux observations et/ou de modèle de prévision mensuelle et ainsi, calculer pourcentage de réussite de prédiction à 30 jours. Malheureusement, nous n'avions pas les données à disposition. De plus, pour améliorer les performances de classification et donc de prévision, il aurait été intéressant de traiter les données horaires, mais l'objectif de l'étude était de fournir une prévision journalière et nous n'avions a disposition que les états de pluie relatifs aux cumuls de pluie journaliers. Les états de pluie horaires fournis par des experts auraient permis une comparaison plus haute résolution, ce qui aurait diminué l'impact de l'inégalité de distribution entre les états. Enfin, la méthode Random Forest, choisie initialement pour sa rapidité et son efficacité, semble efficace dans le cadre de cette étude, mais d'autres méthodes de classification supervisées auraient pu être efficaces. Une première phase de test avec des réseaux de neurones, qui montre des résultats similaires, a été réalisée dans le cadre d'un encadrement de projet d'étude. Leurs travaux ont montré des résultats similaires aux nôtres, mais pas plus performants.

Chapitre 6

Synthèses, conclusions et perspectives

Les observations de la zone côtière sont un point clef pour améliorer nos connaissances des écosystèmes (processus biotiques et abiotiques) sur les impacts des activités humaines et les conséquences sociétales. La construction d'un système d'observation mondial est essentiel pour décrire les variables clés et prédire le fonctionnement des écosystèmes. Un des défis majeurs pour la communauté scientifique qui effectue des observations du milieu marin côtier est d'intégrer les observations des variables océaniques essentielles pour les processus physiques, biogéochimiques et biologiques à des échelles spatiales et temporelles appropriées, et d'une manière soutenue et scientifiquement fondée [Farcy et al. 2019].

Le compartiment phytoplanctonique tient une place centrale au sein des écosystèmes marins, notamment au niveau des zones côtières. Il entre en jeu dans de nombreux processus chimiques et biologiques et a un rôle prépondérant dans le cycle de la majorité des éléments océaniques. Le cycle de vie du phytoplancton est conditionné et limité par les propriétés des masses d'eau. Ainsi, les efflorescences phytoplanctoniques sont des événements qui répondent à des changements liées à des forçages environnementaux et anthropiques. Ces raisons en font un indicateur à haute réactivité face aux changements de l'environnement et la qualité de l'eau. De plus, certains organismes planctoniques intégrés à la catégorie des *Harmful Algal Blooms* (HABs) ont des effets toxiques ou nocifs qui peuvent causer des dysfonctionnements de l'écosystème et/ou engendrer des maladies et de la mortalité à plusieurs niveaux trophiques (chez l'homme, les poissons, les crustacés ...) et notamment au niveau des organismes commerciaux. Compte tenu de ces caractéristiques, le *Global Ocean Observing System* (GOOS) et le *Framework for Ocean Observing* (FOO), ont identifié la biomasse et la diversité du phytoplanctonique comme des variables océaniques essentielles (EOVs) [Lindstrom et al. 2012].

Ainsi, comprendre comment la dynamique planctonique et des états environnementaux associées varient, de l'échelle locale à l'échelle mondiale, est essentiel pour aborder la sécurité alimentaire, les cycles biogéochimiques et le bon état écologique des masses d'eau [Lombard et al. 2019], notamment au niveau des zones côtières qui sont des zones socio-économiques très importantes.

De nombreuses initiatives sont mises en place afin de répondre à ces défis, telles que la structuration de la communauté scientifique dans le cadre du projet H2020 JERICO-NEXT ou encore le consortium COAST-HF de l'IR-ILICO. Les principaux moteurs de ces initiatives sont la directive-cadre sur l'eau (DCE) [DCE 2000/60/CE], la Directive cadre Stratégie pour le milieu marin (DCSMM) [DCSMM 2008/56/CE], ainsi que les conventions sur des mers régionales telle que OSPAR [OSPAR 1992], la convention de Barcelone. Le programme JERICO-NEXT a défini

plusieurs points clefs de la stratégie scientifique côtière. Parmi ces points, le programme exprime un besoin d'examen des menaces environnementales et des lacunes des programmes de surveillance actuels des États membres de l'Union Européenne. Il est notamment mis en évidence la nécessité d'améliorer la surveillance des pressions ou des menaces environnementales et il est indiqué des insuffisances dans les échelles spatiales et/ou temporelles auxquelles la surveillance a lieu, ainsi qu'une surveillance inadéquate des paramètres [Farcy et al. 2019]. De plus, Planque et al. 2011 souligne que l'incapacité d'étudier le changement de l'écosystème à des échelles spatio-temporelles appropriées sous-estimera les incertitudes des prévisions futures ce qui limitera la capacité à réagir aux changements de façons appropriées. Un autre point lié à la compréhension de ses écosystèmes dynamiques, résident dans la qualification et la quantification de l'environnement et des observations rattachées et dans la nécessité d'approches multidisciplinaires. Kupschus et al. 2016 expliquent que pour répondre aux besoins de l'approche écosystémique, la surveillance doit permettre d'établir un lien causal entre les effets des pressions anthropiques et de la variabilité environnementale de l'écosystème, tout en tenant compte de la complexité de ces relations.

Ainsi, le développement des outils d'aide à la caractérisation et à la prédiction - identification de périodes clefs, échantillonnages adaptatifs, prévisions d'événements - devient primordial. La compréhension de la structure des données et l'extraction des informations écologiques multivariées et multi-échelles afin d'améliorer les connaissances et émettre des hypothèses sur le fonctionnement de l'environnement nécessite des méthodes et outils adaptés. La question est donc d'identifier et de mettre en place des outils pertinents, les plus généralistes et automatiques possibles en fonction de ces problématiques écologiques. Cependant, la complexité des ensembles de données environnementales, *i.e.* non linéairement séparables, avec une forte connexité locale, un nombre de données manquantes important, une taille importante ..., peut rendre cette tâche difficile. Les méthodes de *Machine Learning* comptent parmi les stratégies adaptées à ces problématiques écologiques et méthodologiques. Dans une première étude de Rousseeuw et al. 2015b, les auteurs ont eu recours à une méthode de classification spectrale pour segmenter les données. Ils démontrent, sur des cas simples (données linéairement séparables) et des cas plus complexes (données convexes), l'efficacité de la segmentation par classification, en comparaison à des méthodes usuelles en biologie tels que *K-means* [Hartigan et Wong 1979] et *HC* [Borcard et al. 2011]. Ils concluent entre autres que la classification spectrale est plus adaptée pour les bases de données volumineuses multivariées non linéaires, comme c'est le cas dans les études environnementales. Cette méthode a donc été sélectionnée comme base méthodologique lors de cette thèse. Les méthodes de classification spectrale ne tolèrent pas les données manquantes (*NA*), ainsi Rousseeuw et al. 2015b évoque des pistes afin de palier à cette caractéristique, telle que la complétion élastique *DTW*. Cette méthode non conventionnelle est basée sur une hypothèse de récurrence du processus tel un schéma saisonnier. Ensuite, développée et validée par T.T.H Phan [Phan et al. 2017], cette méthode permet d'obtenir une complétion qui préserve la dynamique des signaux complexes dans des séries avec une quantité de données manquantes consécutives importante. C'est la raison pour laquelle, elle a été intégrée dans le cadre de cette thèse.

Dans ce contexte, cette thèse a été consacrée à la mise en œuvre, l'adaptation et la validation d'outils de traitement de données HF (objectif M1) et multicritères (objectifs M2), basés sur les méthodes de classification supervisée et non supervisée, qui permettent de définir des schémas de fonctionnement des efflorescences (facteurs de contrôles, conditions d'initiation et de terminaison et de contrôles de l'amplitude) (Objectif E1). Ces outils seront utilisés pour ensuite effectuer une comparaison des facteurs d'influences (Objectifs E2) et ainsi, comprendre le rôle des événements sur la dynamique et la structuration du phytoplancton dans l'écosystème (Objectifs E3) (Section 3). À partir des méthodes supervisées et d'une base de données labellisée via des outils non supervisés, ont été développés des modèles de reconnaissance (Agent de classification et de prédiction) (Objectif E4) afin de détecter et de prédire automatiquement des événements. La

caractérisation et la prédiction de ces évènements, via le protocole de classification puis de labellisation et enfin de prédiction, permettent de réduire le temps de labellisation d'autres bases et aident à la comparaison des réponses de chaque écosystème à l'identification de changements futurs possibles à court et à long terme.

6.1 Détermination des états environnementaux par classification spectrale multi-niveau

6.1.1 Récapitulatif des ajouts méthodologiques pour la classification non-supervisée

Sur la base de plusieurs méthodes de pré-traitement et traitement des données HF, plusieurs ajouts et développements méthodologiques ont été faits afin (i) d'améliorer le protocole de classification non supervisée dans le cadre de bases de données ayant une forme complexe et une forte connexité locale et (ii) d'obtenir une caractérisation de schémas de fonctionnements, globaux et/ou extrêmes, des écosystèmes côtiers (Schémas de synthèse figure 6.1). Ces ajouts ont été fait à différentes étapes du protocole :

- **L'automatisation et la généralisation du protocole de pré-traitement** ; Avant toutes étapes de classification, il est fondamental de pré-traiter les données afin de bénéficier des séries de données les plus exploitables possibles. Dans notre cas, quatre étapes sont nécessaires : l'alignement, la correction, la complétion et la normalisation des données.
  - **L'alignement** facilite l'étude de la dynamique temporelle et simplifie les comparaisons entre les variables, les calculs d'indices et des statistiques temporelles relatives à chaque état (durée, dates de début et de fin ...). Un calcul automatique du pas de temps et une proposition d'un pas de temps d'alignement idéal est faite afin de fournir une série régulière.
  - **La correction de gamme** est mise en place afin de contrôler la phase de modification et de suppression des données aberrantes ou hors gammes, tout en restant au plus proche des données brutes. Pour cela une gamme "capteur" et une gamme "experte" ont été définies au préalable en fonction de la zone d'étude et permettent d'écarter de manière fiable une grande partie des possibles valeurs aberrantes.
  - Pour **la complétion**, une méthode non conventionnelle a été choisie : la complétion élastique *DTW* [Phan et al. 2017]. À partir de cette méthode, deux paramètres (smallHole et acceptHole) ont été ajoutés afin de permettre d'adapter la taille des fenêtres de complétion. Ceux-ci garantissent une complétion en adéquation avec les processus étudiés et évitent la complétion de périodes trop longues et donc la création "artificielle" de cycles complets de variations. Puis, trois étapes de complétion ont été définies en fonction de la taille de la fenêtre à compléter : (i) par moyenne directe pour une valeur de capteur absente *NA* isolée, (ii) par moyenne mobile pour les petites fenêtres de données manquantes et (iii) par *DTW* pour les grandes fenêtres afin d'optimiser cette phase de complétion et de limiter la phase calculatoire de l'algorithme *DTW*.
  - **L'étape de normalisation** est effectuée pour standardiser les données et permettre l'obtention des données indépendantes de l'unité ou de l'échelle choisie.

- **La classification non supervisée :** Une nouvelle méthode de classification non supervisée : la classification spectrale multi-niveau (*MultiLevel Spectral clustering - M-SC*) a été dévelopée afin de définir des états environnementaux multicritères (combinaison de

paramètres physico-chimiques et biologiques).
Sur la base de la méthode de Rousseeuw et al. 2015b, j'ai contribué à améliorer l'algorithme existant en y ajoutant :

— **Une architecture profonde** qui a pour but de hiérarchiser l'information en partant des changements globaux jusqu'aux événements extrêmes. L'ajout d'un critère nivMax permet de définir plusieurs niveaux de classification, ce qui augmente le niveau de détails et d'aller vers une détection d'évènements plus spécifiques et de plus courtes durées.
— **Un critère de définition automatique du nombre de classes (crit)** plus restrictif a été ajouté pour permettre une classification entièrement non supervisée dans les couches plus profondes. Cet algorithme basé sur une analyse des amplitudes des valeurs propres limite le nombre de classes dans les couches profondes, contrairement au critère GAP initialement utilisé.
— **Un critère de limitation de la sur-segmentation (sil.min)** est mis en place pour stopper la segmentation d'un cluster lorsque celui-ci est bien isolé. Ce critère est basé sur la notion de Silhouette et stoppe la classification dans les niveaux plus profonds lorsque un ensemble d'observations est suffisamment connexe et bien isolé.

Ainsi, au travers de ces ajouts méthodologiques, le nouvel algorithme *M-SC* combine de manière optimisée, trois aspects des méthodes de *Machine Learning*, (i) une approche spectrale, sur la base de l'algorithme *NJW-SC* [Ng et al. 2001] avec (ii) une approche hiérarchique, comme pour les algorithmes tels que *HC* [Borcard et al. 2011] et *H-SC* [Sanchez-Garcia, Fennelly, Norris et al. 2014], et (iii) l'approche fondée sur la densité, comme pour les algorithmes tels que *DBSCAN* [Ester et al. 1996] et *HDBSCAN*.

- **Validation des ajouts méthodologiques** La nouvelle méthode a ensuite été replacée dans un contexte global par le biais d'une analyse comparative. La performance de la méthode *M-SC*, ainsi que celles d'une dizaine de méthodes usuelles, ont été évaluées à partir de 8 jeux de données artificielles et expérimentales, ayant un caractère spatial ou temporel et des formes difficiles à segmenter. L'approche *M-SC* a également été testée sur des données hautes fréquences de la station MAREL-Carnot, afin d'évaluer les réponses de l'algorithme sur un cas pratique dans le contexte de l'observation des eaux côtières. La calibration NivMAx=3, crit=PEV et sil.min=0,7 a été validée comme configuration par défaut.

6.1.2 La méthode M-SC : Outil de détection des états environnementaux

Pour répondre à nos objectifs méthodologiques (M1 et M2), le nouvel algorithme *M-SC* qui combine actuellement, une approche spectrale avec une architecture profonde, une détermination automatique du nombre de classes aux travers de l'analyse des amplitudes des valeurs propres et de la densité des classes, a été développé.

L'architecture profonde offre la possibilité de segmenter les données multivariées et de hiérarchiser les événements en allant des modèles généraux vers les événements extrêmes, en fonction du niveau (Section 3.4.2). Ceci permet un traitement sans perdre les informations clefs pour la détection d'événements extrêmes. En effet, cette hiérarchisation permet dès les premiers niveaux d'identifier les contributions fortes des variables les plus structurantes, telles que la température, liée aux tendances et aux variations grandes échelles (section 1.2.3), et d'observer ainsi des états environnementaux plus spécifiques lors des niveaux plus profonds, tout en conservant la totalité

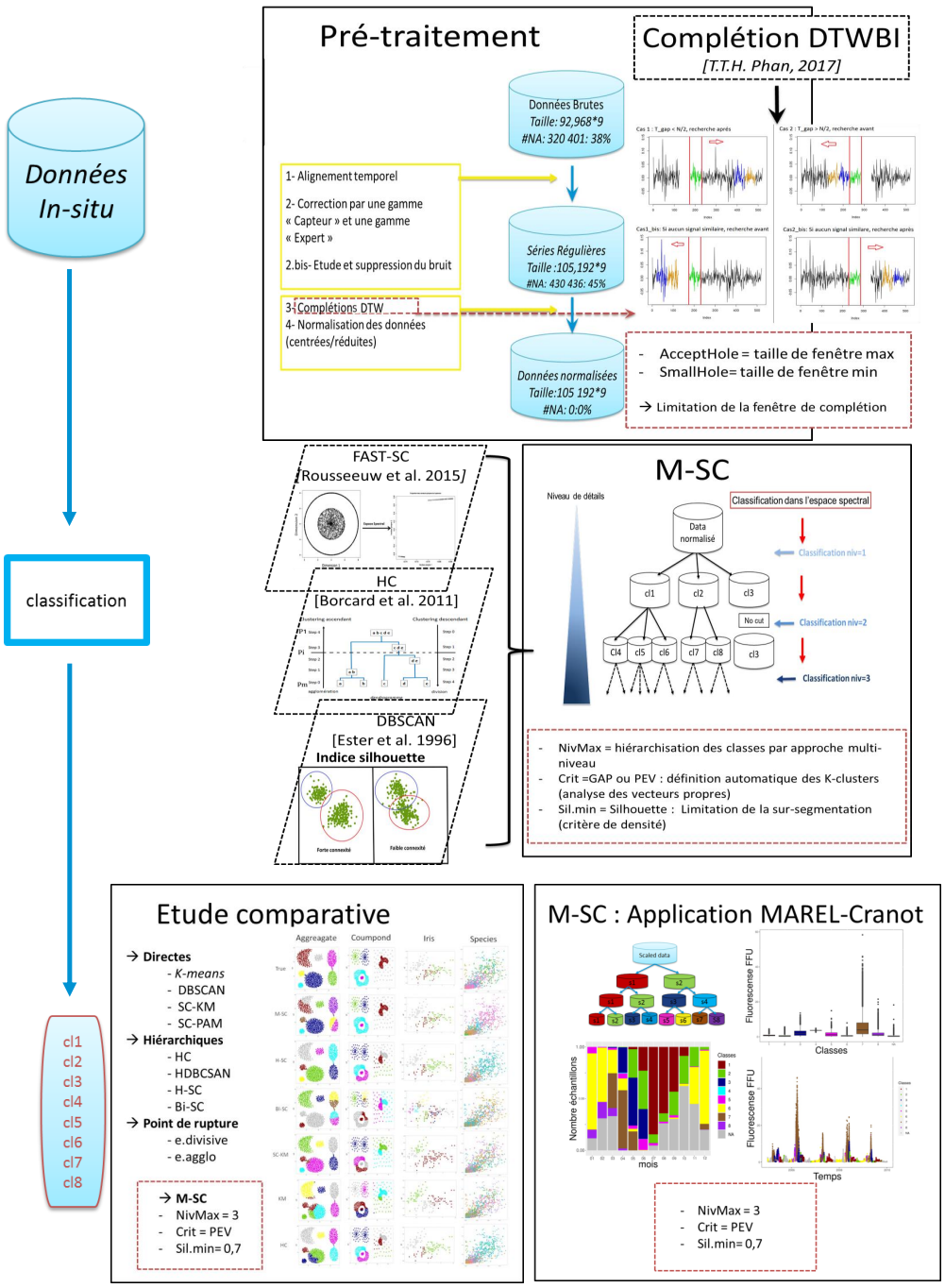


FIGURE 6.1 – Démarche illustrée de la phase de classification non supervisée : du prétraitement à la validation de l'outil M-SC. Trois points clefs sont illustrés : (1) L'automatisation du protocole de prétraitement, (2) le développement de la méthode de classification non supervisée M-SC et (3) le protocole de validation sur plusieurs jeux de données test et sur un cas appliqué (MAREL-Carnot).

des données et des variables explicatives. Cela permet une meilleure identification des effets causaux des variables structurantes, sans réduire la variabilité des données et évite une perte de contraste des variables dépendantes et indépendantes. Ainsi, la méthode M-SC permet une hiérarchisation plus nette des différents facteurs d'influences. De plus, la mise en place d'un calcul automatique du nombre de K-classe à chaque niveau, via le critère de coupe (crit) (section 3.2.3), facilite cette hiérarchisation et améliore la détection des événements extrêmes auparavant limitée par le nombre de classes recherché par l'expert. Ensuite, l'ajout du critère d'arrêt (sil.min) basé sur les indices de densité et de connexité simplifie le regroupement des classes réparties inégalement ; cela permet l'identification de classes de tailles et de durées différentes. C'est un avantage significatif pour l'étude des processus dont la phénologie est fortement variable, comme pour le phytoplancton. Ainsi, les experts peuvent utiliser ce critère pour trouver un compromis entre la sur-segmentation et le nombre de classes pour leur tâche de labellisation.

Ces différents aspects ne sont pas présents ou seulement partiellement abordés par les autres algorithmes de *clustering*. Les tests effectués sur les différents jeux de données, artificiels et expérimentaux, dont les signaux sont non stationnaires et de formes convexes et non linéairement séparables, ont montré que l'architecture *M-SC* apporte une valeur ajoutée pour segmenter les différents types de formes, contrairement aux algorithmes usuels (Section 2.4.4). Les résultats obtenus montrent que *M-SC* fait toujours partie des méthodes les plus efficaces pour détecter des structures cohérentes quelle que soit la base de données. Cela démontre une bonne faculté d'adaptation et le caractère générique de la méthode *M-SC*. Ainsi cette méthode offre un potentiel d'application large et une base stable d'inter-comparaisons entre ces applications.

Ces résultats révèlent également, dès les premiers niveaux, une bonne capacité de labellisation et proposent un nombre de classes correct avec des structures cohérentes. La classification spectrale multi-niveaux de *M-SC* permet donc de mettre en œuvre des approches imbriquées et d'optimiser l'extraction des connaissances lors de l'examen de données couvrant différentes échelles (temporelle, fréquentielle ou spatiale). L'approche *M-SC* semble donc bien adaptée pour proposer une segmentation des séries temporelles ou des données spatiales selon des modèles cohérents, qui pourraient aussi bien apparaître plusieurs fois qu'une seule fois.

En revanche, l'approche spectrale exige toujours un ensemble complet de données à l'entrée de l'algorithme, c'est-à-dire des observations sans valeur manquante (*NA*). Grâce à l'intégration de méthode de complétion optimisée, cette limite de l'approche est minimisée. En effet, elle permet d'être plus proche de la dynamique du signal et limite le lissage des variations. Ainsi, elle est plus efficace et cohérente pour la complétion des grandes séquences de données manquantes. De plus, l'ajout des deux critères de définition de la taille de fenêtre min et max de complétion permet la limitation de la taille acceptable de la période à compléter. Cette taille de fenêtre est un point non négligeable, car cela garantit que la complétion est en accord avec les échelles des processus étudiés. Cette ajout méthodologique, limite le risque de corrélations aléatoires qui réduit la variabilité potentiellement à des tendances trop monotones, comme le ferait le lissage par moyenne mobile. Cela augmente donc la capacité d'étudier la dynamique des processus à plus haute variabilité à des échelles spatio-temporelles appropriées.

6.2 Labellisation des événements

6.2.1 Bilan : Identification des états environnementaux et labellisation

La classification par *M-SC* et les analyses complémentaires ont permis de labelliser les classes (cl) en états environnementaux (l) en fonction des variables structurantes et des assemblages phytoplanctoniques. Les classes identifiées suivent le schéma saisonnier typique des zones tempérées. Chaque saison est représentée par une ou plusieurs classes, chacune étant reliée à des facteurs

d'influences qui peuvent impacter les proliférations (Figure 6.2). Pour le site MAREL-Carnot, la labellation suivantes a été établie :

- **Période hivernale (classe cl6) :**
  - cl6 est labellisée comme état hivernal avec une faible productivité. Cet état est l'un des plus diversifié. Il se compose de 17 taxons dominant cohérent avec la période. La majeure partie des taxons sont des stratèges-R (R : espèces Rudéales) dont certains sont euryhalines tel que *Skeletonema costatum* ou *Asterionellopsis*. C'est une classe rattachée à de forte concentration en sels nutritifs et des dessalures importantes.

- **Période printanière (classes cl8, cl7, cl3 et cl2.1) :**
  - cl8 est labellisée comme un état pré-printanier correspondant à une amorce avant l'efflorescence printanière principale. Elle est représentée par un assemblage transitoire, composé de plusieurs taxons stratèges-R tels que *Pseudo-nitzschia, Chaetoceros, Thalassiosira*. Cet état est associé à une augmentation de l'éclairement et un rapport N/P élevé avec un excès d'azote par rapport au phosphate.
  - cl7 est labellisée comme état d'efflorescence favorable à la prolifération de la Prymnésiophycée *Phaeocystis globosa*. Cet état représenté principalement par le taxon *Phaeocystis* est marqué par la mise en place une amélioration des conditions d'éclairement favorisées par une faible turbidité et une augmentation de la lumière, et par conséquent une augmentation de la zone euphotique. C'est aussi un état caractérisé par une forte production et par conséquent une chute des nutriments. Le taxon *Phaeocystis* regroupe tous les phases du cycle de vie du genre. En fonction de celle-ci des stratégies de vie C (C : espèces colonialistes-invasives) et S (S : tolérance au Stress) seront observées.
  - cl3 est labellisée comme état d'efflorescence potentiellement toxique où l'ont retrouve la combinaison *Phaeocystis* et *Pseudo-nitzschia*, de stratégie R. Il est corrélé à des conditions de fin d'efflorescence printanière avec un limitation forte en nitrate et en silicate.
  - cl2.1 est labellisée comme un état de transition correspondant à la fin de l'efflorescence. Associer à des conditions environnementales automnales, il présente un assemblage plus diversifié que cl3 et cl7 avec des taxons de stratégie intermédiaire entre S et R, tel que *Rhizosolenia*. Il est déclenché par une ré-augmentation des apports de nutriments dans le milieu.

- **Période estivale (classe cl1) :**
  - cl1 est labellisée comme état estival avec une faible productivité : caractérisée par une augmentation de la température et un éclairement fort. Cet état est représenté par un assemblage taxonomique varié, cohérent avec la période et dominé par les *Chaetoceros*.

- **Période automnale (classe cl2.2) :**
  - cl2.2 est labellisée comme état d'efflorescence secondaire automnal. Composé par un mélange de taxons estivaux et automnaux, il correspond à un état dont les conditions d'éclairement et les apports en sels nutritifs sont favorables au développement phytoplanctonique.

- **Deux états définis comme extrêmes (classes cl5 et cl4) :**
  - cl5 correspond à une augmentation rapide des nutriments dans le milieu. Cette augmentation est propice à la croissance phytoplanctonique et peut générer des efflorescences spontanées.

- cl4 n'est pas labellisée à partir d'assemblage taxonomique, mais il correspond à un apport de phosphate conséquent qui s'est produit deux années de suite (Probablement du à une période de crue exceptionnelle ; la confirmation de cette hypothèse nécessite l'ajout de données non accecibles dans le cadre de cette thèse).

6.2.2 Vers une caractérisation multivariée des processus environnementaux

La méthode *M-SC*, combinée à deux approches complémentaire à savoir une détermination des assemblages taxonomiques et un diagramme de Margalef [Margalef 1978] étendu à 6 variables, a été appliquée pour labelliser le jeu de données MAREL-Carnot afin de définir des schémas de fonctionnement des efflorescences du phytoplancton (objectif E1) et comprendre et hiérarchiser le rôle respectif de chaque état et les effets des potentielles perturbations (Objectifs E2, E3). L'ajout de variables complémentaires (EOVs) augmente la précision de la labellisation et permet de distinguer les différents facteurs d'influence de chaque classe définie par *M-SC* et de rattacher ces classes à d'autres facteurs pression ou réponse liées à l'état identifié. Ainsi, la labellisation en 8 états environnementaux a permis d'estimer et de valider la capacité de la méthode *M-SC* à détecter des schémas de fonctionnements cohérents avec la dynamique phytoplanctonique de la zone.

Au travers de ce travail de labellisation, a été mis en évidence deux aspects intéressants de la méthode de classification *M-SC* : (i) la capacité de détections des évènements extrêmes et (ii) la mise en évidence d'une succession cohérente des états environnmentaux.

En effet, la méthode *M-SC* propose différentes échelles d'interprétation. Lors de l'étape de classification au niveau 3, les classes cl4 et cl5 sont identifiées comme ayant une dynamique caractéristique d'un évènement extrême. Pour rappel, nous considérons un évènement extrême comme un évènement de courte durée et/ou un évènement particulier par rapport au cadre général. Le caractère sporadique et spontané de ces évènements rend leur détection et leur compréhension complexe. Dans ce cas, *M-SC* se présente comme un outil pertinent d'aide à leur identification. Il peut aussi bien être utilisé pour les facteurs de déclenchement, que pour la détermination des conditions caractéristiques de ces évènements. Néanmoins, cette détermination est limitée par le nombre de variables utilisées en entrée de la méthode, mais aussi par le pas d'échantillonnage spatial et temporel.

En effet, il est important de noter que dans une optique de détection d'évènements extrêmes, le calcul des classes à un niveau plus profond (4 ou même 5 ou 6), aurait été plus intéressant. Cependant, la mise en correspondance avec les assemblages taxonomiques aurait nécessité un échantillonnage et/ou un dénombrement des espèces phytoplanctonique plus fréquent. La méthode de classification, via les données HF, permet d'identifier des classes au niveau horaire, un prélèvement bi-mensuel est trop faible pour aller jusqu'à ce niveau d'interprétation.

Les résultats obtenus sur le jeu de données issues de la campagne en mer CGFS (Channel Ground Fish Survey) (Section 5.1.3) ont démontré que l'ajout de données issues de mesure fluorimétrique telle que l'AOA ajoute une dimension supplémentaire à la classification et permet la définition des relations entre les groupes phytoplanctonique et les propriétés hydrologiques et biogéochimiques de l'environnement. Ainsi, ils sont à considérer comme une valeur complémentaire, plus haute fréquence, aux données par microscopie.

De plus, une chronologie cohérente au niveau physico-chimique et biologique est mise en évidence. Une succession des évènements printaniers définis par cl8, cl3 et cl7 a pu être identifiée. Ces résultats sont d'autant plus intéressants sachant que la classification est réalisée indépendamment du temps et que rien ne présageait de la capacité de la méthode à réorganiser les états avec un cohérence temporelle, conforme à ce qui est connu pour un tel écosystème. Ainsi, il est possible d'identifier les caractéristiques propres à chaque état et distinguer les différents

facteurs d'influences lors de cette succession. Par exemple, il est possible de dissocier les efflorescences nuisibles de *Phaeocystis*, de celles combinant *Phaeocystis* et *Pseudo-nitzschia* dont les conséquences pourraient être plus importantes, puisque cumulant le développement d'un taxon nuisible à forte biomasse et d'un taxon potentiellement toxique. Ainsi, cette approche est un outil d'aide conséquent pour tout expert souhaitant identifier et mieux comprendre les relations pressions/impacts. Elle renforce les capacités de prédiction de certains événements comme les efflorescences nuisibles du phytoplancton.

De plus, il est intéressant de noter que la méthode permet de se focaliser sur les périodes clefs et fait donc une distinction intra-saisonnière de la période productive, alors que les périodes non productives (plus stable vis-à-vis de la dynamique phytoplanctonique) sont caractérisées seulement par une classe. Ce résultat est lié à l'approche hiérarchique de la méthode et au critère d'arrêt basé sur la densité des classes. Il démontre que la méthode ne sur-segmente pas les séries et qu'elle est bien capable de créer des classes de taille inégale conforme à l'hétérogénéité temporelle attendue des processus étudiées. Ainsi, la combinaison de ces deux approches entraîne une segmentation structurée et cohérente avec la dynamique variable et non linéaire du phytoplancton.

Ces résultats démontrent le potentiel de la méthode *M-SC* pour aider à améliorer les connaissances sur la structure et le fonctionnement de états environnementaux. La méthode peut permettre de proposer des évaluations globales ou locales et pourrait aider, par exemple, à la prise de décision dans le cadre du projet qui vise à établir des bilans de l'état du milieu marin tel que la DCSMM [DCSMM 2008/56/CE]. Suite aux travaux réalisés sur les données de la campagne en mer CGFS (Chapitre 5, Section 5.1.3), l'outil de classification spectrale a déjà été proposé comme complément aux méthodes d'évaluation de la DCSMM dans le but de valider à partir de données physico-chimiques et biologiques *in situ* la délimitation des paysages marins utilisés comme unités d'évaluation du Bon État Écologique [Lefebvre et Devreker 2019c]. De plus, son potentiel d'identification n'est pas limité aux données marines. Les travaux réalisés sur les données de précipitations, montre qu'il est possible d'élargir le champ de la recherche d'autres secteurs dont les problématiques de détection et de prévision également importantes, telles que la météorologie.

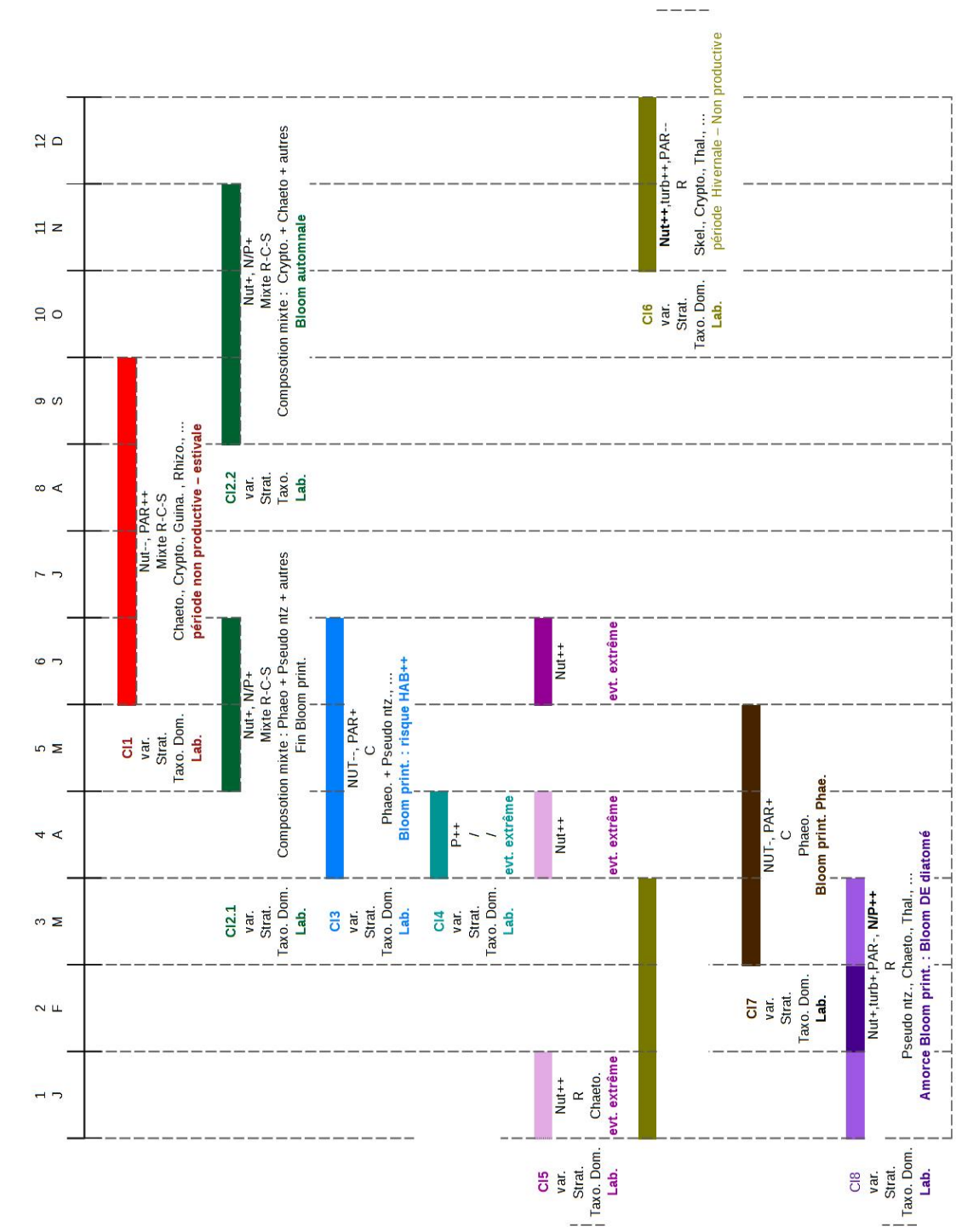


FIGURE 6.2 – Classification et labellisation des différents états environnementaux en Manche Orientale. Schéma récapitulatif des différentes classes (cl_i), de leurs distributions mensuelles, des taxons dominants (Taxo. Dom.) et des potentiels stratégies de vie rencontrée (Strat.), des variables structurantes (Var.) (les symboles $+$,$++$,$-$ et $--$ indiquant leurs importances relatives dans le milieu) et du label (l_i) qui leur a été attribuée par statistique.

6.3 Apprentissage et reconnaissance des évènements

6.3.1 Protocoles d'identification et de prédiction automatiquement des évènements

Les phases de classification par *M-SC* et de labellisation après expertise ont permis d'établir une typologie des évènements pour la station MAREL-Carnot entre 2005 et 2009. À partir de cette analyse, il est possible de construire une base de données d'entraînement qui permettra, par le biais de méthodes de *Machine Learning* supervisées, de développer des agents/modèles de classification afin d'identifier et de prédire automatiquement des évènements. Différents modèles de prédiction ont été testés afin d'observer la capacité de détections des schémas d'efflorescences du phytoplancton sur le site d'étude MAREL-Carnot et dans d'autres écosystèmes (MAREL-Iroise et MesuRho) (Schémas de synthèse figure 6.3).

- **Bilan de la phase de construction du modèle.**
  - Méthodes d'apprentissages : 3 méthodes d'apprentissage, *k-ppv*, *SVM*, *RF* avec pour chacun 2 paramétrisations différentes, ont été testées.
  - Configurations : Chaque méthode est entraînée sur les jeux d'entraînement MAREL-Carnot suivant 5 configurations d'apprentissage (1 à 5, Tableau 4.1)
  - Indices de performance : la précision, le rappel (Recall), le score F1, le taux de reconnaissance global (accuracy - acc.) et #Iso, le nombre de classes bien isolées.
  - Capacité d'apprentissage : Les 3 systèmes de reconnaissance ont tous une bonne capacité d'apprentissage.
  - Capacité de généralisation : Les résultats obtenus en généralisation restent inférieurs aux résultats d'apprentissages. Les labels l4 et l5, sont les moins bien prédits. Ces états sont des évènements extrêmes comprenant peu d'exemples dans le jeu d'entraînement et qui sont donc plus difficiles à prédire.

- **Prédiction sur d'autres jeux de données.** Nous avons cherché à estimer la capacité de généralisation du modèle construit à partir de la base labellisée de MAREL-Carnot, à reconnaître un type d'événement sur d'autres sites d'études plus ou moins contrastés, ici MAREL-Iroise et MesuRho, c'est-à-dire un site dans l'Atlantique et un site en Méditerranée.
  - Modèle de classification : Le choix d'explorer des solutions simples, rapides en temps de mise en œuvre et calculatoire, et ne nécessitant pas un grand nombre d'observations a été fait, c'est pourquoi seul le classifieur K-ppv à un voisin a été selectionné.
  - Configurations : 4 configurations de généralisations ont été testées : la base MAREL-Iroise pour les années 2006, puis 2010-2014 et MesuRho pour 2009 et 2014-2016.
  - Indices de performance : La capacité de généralisation est évaluée à partir des critères de comparaison du partitionnement : le Rand Index (RI) et l'Adjusted Rand Index (ARI) et d'un critère temporel, en observant la répartition des classes par mois et une projection des résultats sur le signal de fluorescence.
  - Capacité de généralisation : Quelle que soit la base de test utilisée et le niveau *M-SC* d'interprétation, le score RI est supérieure à 50 %. De plus, même s'il est faible le taux ARI est positif. Ces résultats démontrent qu'il est pertinent de proposer une première labellisation de cette base à partir de MAREL-Carnot et que le partitionnement obtenu n'est pas aléatoire. De manière générale, les labels l4 et l5 (évènements extrêmes) sont peu représentés dans les prédictions MAREL-Iroise et MesuRho ; ils ne correspondent pas à des périodes identiques à celle de MAREL-Carnot. Le modèle avait déjà du mal à les prédire sur les jeux test MAREL-Carnot. De plus, ils sont structurés par des

variables non incluses dans le jeu de données d'apprentissage, restreint aux variables communes aux différents systèmes de masures HF.

- **Analyses et comparaisons des dynamiques prédites.**
  - Pour la station MAREL-Iroise, une dynamique proche de celle des classes de MAREL-Carnot a été retrouvée, avec quelques différences plus prononcées au niveau 3. À ce niveau, une similitude de structure des principales phases de bloom entre les classes (cl_i) et les labels prédits (l_i) est clairement identifiée. Les autres classes sont aussi identifiées, mais avec plus de confusion. Les mêmes conclusions se retrouvent d'un point de vue temporel, avec une prédiction cohérente des états printaniers et des périodes qui diffèrent de MAREL-Carnot pour les autres états.
  - Pour la station MesuRho, une différence de structuration des variables plus importante entre les jeux de données MesuRho et MAREL-Carnot est constatée. Une répartition inégale des classes est observable dès le premier niveau. Mais les prédictions restent concordantes avec les caractéristiques du site. De même au niveau temporel, il est difficile de retrouver un schéma équivalent à celui de MAREL-Carnot.

  Ces résultats démontrent que le modèle est capable de prédire des états en cohérence avec la zone d'études. Toutefois, les processus biogéochimiques et donc les états peuvent varier d'une région à une autre et le modèle peut ne pas être adapté à la nouvelle variabilité du jeu de données. Cette conclusion peut paraître triviale, mais elle permet d'insister sur l'hétérogénéité des écosystèmes étudiés et le besoin de baser nos observations, et par conséquent d'émettre nos hypothèses et résultats, à partir de modèles locaux à régionaux, en fonction du questionnement scientifique considéré.

6.3.2 Vers un outil d'aide à la reconnaissance et à la prédiction des états environnementaux

Une des étapes de cette thèse a consisté à étendre l'approche *M-SC* totalement non supervisée à une approche semi-supervisée. Cette approche a pour but de développer des modèles de classification supervisées afin d'aller vers de la prédiction automatique des différents schémas de fonctionnement (objectif E4) définis après expertise (labels *M-SC*). Les connaissances *a priori* sont alors utilisées comme informations clefs dans la phase d'apprentissage. Elles permettent d'améliorer les connaissances sur la dynamique et la variabilité du jeu de données et par conséquent, permettent de mieux définir les limites des états et enrichissent le modèle de prédictions.

L'évaluation des modèles - *k-ppv*, *SVM*, *RF*- conduit sur la base des données MAREL-Carnot, démontre qu'il est possible avec cette approche de construire un modèle pertinent de reconnaissance automatique des états environnementaux. Ce modèle offre la possibilité de proposer un système d'aide à la labellisation et à la prédiction sans connaissance *a priori* sur un nouvel ensemble de variables identiques ou restreintes issues du même site d'étude. Cette méthode permet une optimisation du temps de labellisation et fournit une première estimation des états caractéristiques de la zone nouvelle zone étudiée. Cette première labellisation d'un nouveau jeu de données permet d'identifier rapidement les similitudes et les différences de dynamique entre les sites étudiés et permet ainsi rapidement d'envisager l'ajout de nouvelle variables et l'ajustement de la stratégie d'échantillonage.

À partir de ce modèle, nous avons évalué la capacité de cette approche à définir des schémas de fonctionnement dans d'autres écosystèmes qui ont des caractéristiques environnementales différentes. Des schémas cohérents temporellement, vis-à-vis des labels prédits et de leurs dynamiques environnementales, ont pu être mis en évidence. Ainsi, le modèle peut être employé comme outil d'identification et de comparaison des caractéristiques environnementales d'une

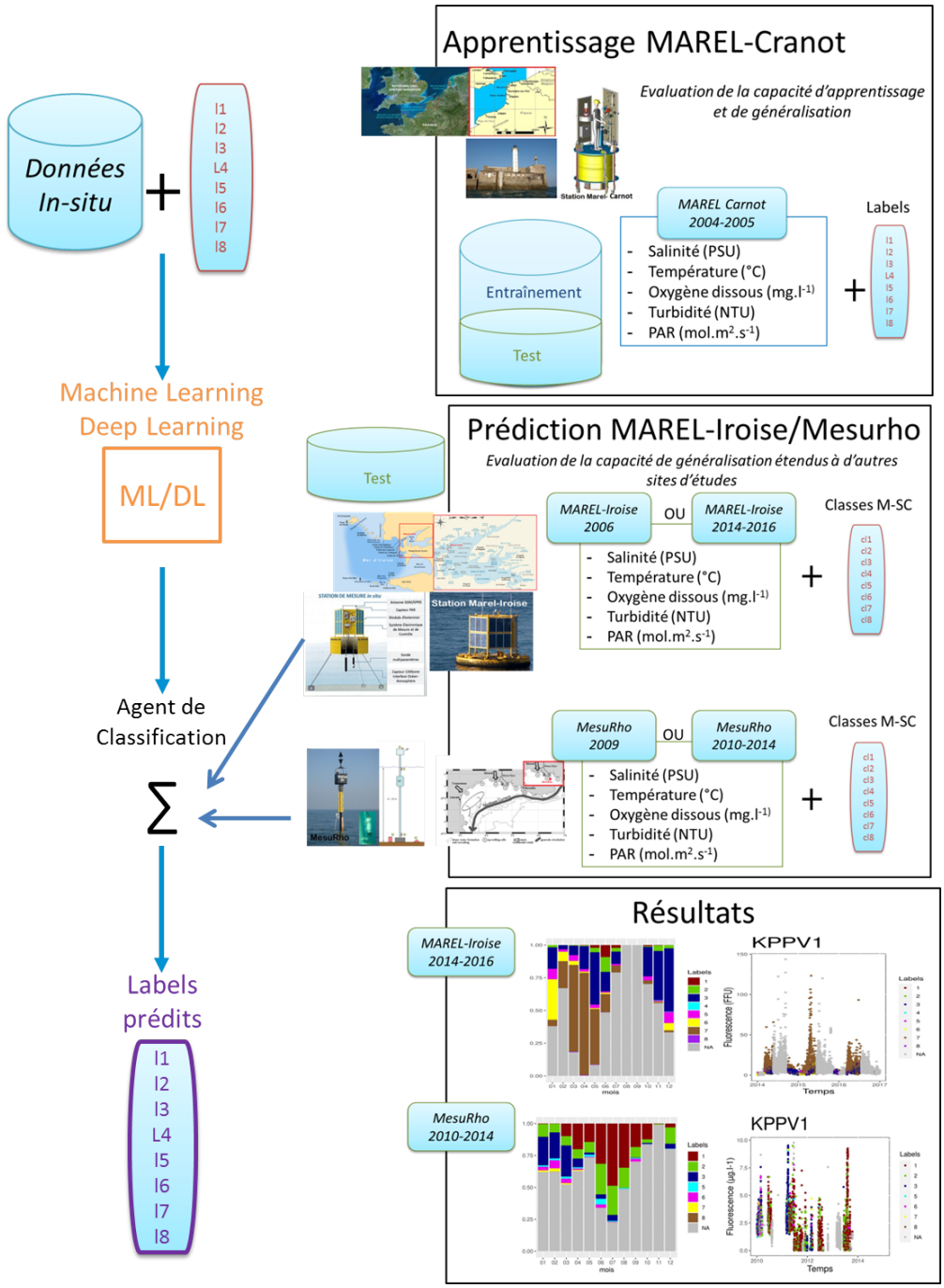


FIGURE 6.3 – Construction du système de reconnaissance d'états environnementaux en trois étapes : 1 apprentissage et validation sur le jeu MAREL-Carnot, 2 généralisation sur de nouveau jeux de données MAREL-Iroise et MesuRho, 3 analyse de la dynamique des labels prédits.

nouvelle base de données. Il offre la possibilité d'identifier les caractéristiques communes entre un jeu d'apprentissage et un jeu de données non labellisé issus d'une autre base de données. Il est ainsi possible de coupler les schémas obtenus par le système de reconnaissance à différentes échelles, spatiales d'une part en utilisant une base de données localisée dans une zone différente et temporelle d'autre part, au moyen des niveaux de classification proposés par *M-SC*. Au regard des questions scientifiques,des besoins d'identification, de comparaisons et de prédiction des états environnementaux, à l'échelle nationale ou même de l'océan globale, il est possible de détecter et cibler, par le biais du modèle de reconnaissance MAREL-Carnot, des événements communs liés aux forçages globaux par rapport à une dynamique de système tempéré. Il aussi possible de différencier des événements particuliers propres à chaque site d'études liés à des forçages plus locaux. Toutefois, pour une détection à plus fine échelle, le modèle peut ne pas être adapté à la nouvelle variabilité des états et donc du jeu de données. Pour que la méthode soit performante à petite échelle, nous considérons que le modèle doit être construit dans un cadre régionalisé, c'est-à-dire à partir d'une base de données labellisée issue d'une zone géographique à proximité. Ainsi, l'utilisation du modèle appris faciliterait la caractérisation des jeux de données issues d'une autre bouée fixe, d'une campagne périodique ou encore lors d'un déploiement ponctuel.

De plus, cette comparaison à grande ou à petite échelle est limitée par le nombre de variables EOVs utilisées. Dans notre cas, les modèles montrent des résultats intéressants avec un nombre de variables restreintes. Néanmoins, l'ajout de variables structurantes lors de la classification spectrale, comme les concentrations en Nitrate, Phosphate et Silicate ou encore le signal de fluorescences en FFU conduiront à une compréhension plus approfondie de la dynamyque du plancton et de sa relation avec l'environnement, ce qui permettra d'améliorer la qualité de la prédiction. Toutefois, même si le nombre de nouveaux capteurs, instruments, plates-formes et de variables mesurées se multiplient, l'utilisation relativement diversifiée et hétérogène des méthodes et techniques rend difficile la découverte de modèles.

6.4 Axes futurs de recherche - Perspectives

Un certain nombre de solutions fondamentales et applicatives ont été proposées au cours de cette thèse. Cependant, plusieurs points peuvent être améliorés.

Enrichissement des bases de données

Bien que le volet multi-sources ait été mis en avant dans ce travail au travers des bases de données complémentaires et des analyses multi-échelles et multi-sites, il n'a pas été exploité dans son entièreté. En effet, l'étude des états environnementaux et de leur détection à des échelles appropriées, soit spatiales, *i.e.* écosystèmes contrastés, soit temporelles, *i.e.* inter-annuelle, intra-journalière . . . , par le biais de l'approche semi-supervisée, a plusieurs fois été limitée par le nombre d'années et de variables communes. Les différences de paramètres mesurés, de leurs unités de mesure et de leurs accessibilités rendent les comparaisons difficiles. Les directives et conventions internationales telles que DCE, DCSMM [DCSMM 2008/56/CE], OSPAR [OSPAR 1992], les projets européens et aussi les infrastructure de recherche comme l'IR-ILICO [Farcy et al. 2019], identifient tous un besoin d'homogénéisation et de mise en commun des données au niveau international. La stratégie scientifique idéale définie par le projet JERICO-FP7 propose : (1) une future stratégie globale pour les observations de l'océan côtier européen [D1.11 Morin et al. 2015] et (2) une définition d'une stratégie pour le suivi de la biodiversité marine [D1.9, Wijnhoven 2014], avec la nécessité d'un « examen des variables essentielles de l'océan et de la biodiversité (contributions du GOOS et de GEO BON) ». Mais dans la continuité de l'identification des EOVs,

il y a encore actuellement un besoin d'harmonisation des protocoles de mesure au sein des réseaux et au sein de la communauté pour un grand nombre de variables.

Dans un autre cadre applicatif, avec des objectif d'étude à grandes échelles des changements globaux, il existe au niveau des réseaux BF des séries de données complètes et plus longues avec potentiellement autant voir plus de variables mesurées. Une application de l'approche, si la série est suffisamment importante, pourrait apporter un nouveau regard sur l'étude.

Si l'on considère l'acquisition d'une base de données complète, complémentaire et homogène, la méthode *M-SC* est applicable et réplicable sur la majeure partie des jeux de données. Ainsi, aux mêmes titres que MAREL-Carnot, d'autres bases de données (satellitaires, météorologiques, HF ou BF, ...), auraient pu être classées, labellisées et apprises. Plusieurs modèles de classification pourront être construits, en fonction de la base d'entraînement mise en entrée et/ou de la méthode de *Machine Learning* utilisée. À la fin de cette phase, un système multi-agents pourrait être créé, puis en fonction du nouveau jeu de données mis en entrée et des demandes de l'utilisateur, le système pourra utiliser le modèle le plus performant. Ainsi, l'utilisateur disposerait d'une "bibliothèque" de modèles adaptés à la base de données et/ou au site d'étude, ce qui assurera une meilleure prédiction. Dans le cadre des approches et systèmes d'observation intégrés, la méthode *M-SC* fournirait un outil collaboratif de visualiser et de classer les données et permettrait de partager les données résultantes d'une manière homogène. Un premier pas vers cette idée a été fait et plusieurs bases de données ont été rassemblées et analysées. Toutefois, la majorité des données disponibles étaient, soit des variables redondantes, soit des variables complémentaires qui se sont avérées difficilement exploitables.

De plus, la mise en commun et l'exploitation de plusieurs bases de données peuvent être limitées par les échelles d'échantillonnage. Même si un système multi-agents peut permettre l'utilisation de base de données avec plusieurs fréquences de mesures, l'utilisation de différentes variables comme entrées d'un classifieur nécessite une homogénéisation des échelles de celles-ci. Cela peut donc induire une perte d'information et/ou limiter le degré de caractérisation de chaque classe en états environnementaux. Comme exemple direct, l'identification de communautés phytoplanctoniques et la labellisation des classes au niveau horaire n'ont pas pu être réalisées. Les données issues de mesure spectrale telle que l'AOA [Beutler, Wilthshire et al. 2002] ou le cytomètre en flux [Gwatkin 1995] peuvent être considérées comme variables complémentaires, plus haute fréquence, par rapport aux données par microscopie. La mise en place de ce type de système est d'ailleurs envisagé dans le cadre de la jouvence MAREL-Carnot engagée dans le CPER Marco 2014-2020 et poursuivie par une intégration des systèmes de mesures dans le nouveau CPER IDEAL. L'objectif est d'effectuer une transition au niveau instrumentation qui devrait permettre de passer de MAREL-Carnot - Génération 1.0 à MAREL-Carnot - Génération 2.0 d'ici 2021. Cette dernière permettant d'élargir le spectre des mesures disponibles. L'intégration de ces données ajouterait une dimension supplémentaire à la classification et permettrait la définition des relations entre les communautés et les propriétés hydrologiques et bio-géochimiques de l'écosystème.

6.4.1 Amélioration de l'approche semi-supervisée

Les tests sur les nombreuses bases de données et les optimisations apportées à la méthode ont permis de s'assurer de la fiabilité de la mise en application de la méthode. Les protocoles de prétraitement et de classification *M-SC* ont été poussés à un haut degré de paramétrisation et d'automatisation. Un des points clefs de la méthode est l'estimation du nombre de classes. Les

méthodes d'analyses des amplitudes des valeurs propres (GAP et PEV) sont bien adaptées, mais les méthodes EM pourraient proposer un partitionnement alternatif ; elles restent toutefois plus coûteuses en terme de mise en œuvre et en temps de calcul. Un autre point clef de la méthode est la complétion des valeurs manquantes. Dans notre cas, elle a été effectuée par *DTWBI* sur des séries de manières indépendantes. L'utilisation d'une méthode de complétion multivariée tel que *DTWBI* (extension multivariée de *DTWBI*) améliorerait cette étape, mais risque de réduire encore davantage le nombre de correspondances et le nombre d'observations communes pour une date donnée) et donc le nombre de fenêtres complétées. L'émergence des DNN - Deep Neural Network ou réseau avec apprentissage réduit (EML) sont des pistes à creuser, elles sont aussi basées sur une fenêtre d'observation pour prédire une nouvelle fenêtre ou données. Cependant les premiers études ne démontrent pas que ces classifieurs aient une efficacité supérieure sur des données à forte variabilité, notamment pour préserver la forme des données. D'autre part, le choix du calcul de la similarité et de l'opérateur Laplacien n'a pas été étudié ici. Le choix d'une matrice de similarité W ou de la définition Laplacienne L différente pourrait affecter les résultats et être plus approprié en fonction de l'application. La version Fastspectral est déjà un exemple de calcul mieux adapté pour les grands jeux de données.

En ce qui concerne l'apprentissage, le choix a été fait d'explorer des solutions simples, rapides en temps de mise en œuvre et calculatoire, et ne nécessitant pas un grand nombre d'observations pour aboutir à des résultats satisfaisants. L'exploitation de méthodes neuronales peu profondes a montré de premiers résultats prometteurs avec des performances équivalentes aux autres méthodes, mais l'approche neuronale plus profonde n'a pas pu être explorée. La quantité actuelle d'observations ne permet pas de paramétrer simplement ces techniques. Des tests sur une base de données plus grande seraient un point à approfondir, premièrement sur une base de données artificielles, puis sur une base de données environnementale plus longue.

Un apprentissage par renforcement serait aussi envisageable afin d'augmenter la quantité d'informations traitée et dans le même temps élargir le nombre d'observations de labels sous-représentées. L'apprentissage par renforcement permet de bénéficier d'un nombre d'informations plus importants et permet de se reposer sur l'utilisation de données indirectement étiquetées par des récompenses. L'utilisation de ce principe n'a pas été faite ici, car le nombre de variables communes aux trois stations en entrée du modèle d'apprentissage était trop faible et peu informatif dans le cadre des évènements extrêmes, mais l'enrichissement des bases de données peut être une solution pour la mise en œuvre de cette approche. Une autre solution serait d'enrichir le modèle d'apprentissage par apprentissage dynamique en y intégrant les nouvelles données prédites. Ces approches permettent en effet de réadapter le modèle en fonction des nouvelles données acquises et en restant optimisée.

Cette approche semi-supervisée permet d'identifier et prédire des schémas de fonctionnement à partir de données complexes, de formes convexes, non linéaires, non stationnaires avec des fréquences d'échantillonnages multiples. Elle peut donc être appliquée pour un grand nombres de jeux de données et de cas d'études par la communauté scientifique. Toutefois, elle peut sembler complexe et peu abordable pour certain utilisateur. Une interface utilisateur, comme celle mise en place dans le package uHMM : :interface [Rousseeuw et al. 2015b] faciliterait la diffusion et l'utilisation de la méthode.

6.4.2 Application environnementales

Dans le contexte actuel où de nombreux systèmes d'observations intégrés, génèrent un nombre croissant de données multivariées, multi-sources et multi-échelles et sont mis en place, la capacité

de l'approche semi-supervisée à traiter de nombreux types de données permet une grande comparabilité des résultats. Au-delà de l'amélioration des connaissances scientifiques, cette comparabilité est cruciale pour aborder de façon multidisciplinaire, la compréhension de l'évolution de notre océan et de ses habitats. Muniz Piniella et al. 2018 et Kupschus et al. 2016 mettent en évidence le besoin de mettre en œuvre des stratégies flexibles qui permettent d'isoler les effets des changements de méthodes et des changements dans l'écosystème. Ainsi, son potentiel de détection des évènements multivariés globaux et extrêmes à haute définition spatiale et temporelle peut être un outil permettant l'inter-comparaison des systèmes d'observation et des écosystèmes marins. La méthode pourrait, par exemple, être utilisée comme point de comparaison inter-sites et fournirait un moyen de comparaison entre les dispositifs d'échantillonnage existants. Dans le cadre du COAST-HF (IR-ILICO), il pourrait permettre la comparaison et l'étalonnage, avec des normes identiques, de l'ensemble des stations HF et d'identifier certaines « contraintes et opportunités technologiques » comme spécifier dans les objectifs stratégiques [Farcy et al. 2019]. Ceci faciliterait l'intégration de ces systèmes d'observation dans des projets nationaux ou européens, à partir d'une stratégie scientifique harmonisée et commune à l'échelle.

De plus, l'approche semi-supervisée peut aussi être envisagée comme outils d'aide aux développements d'indicateurs communs. Dans Kupschus et al. 2016, il est défini que la précision maximale de la mesure d'indicateurs individuels n'équivaut pas à la précision d'une estimation de l'état global de l'écosystème. Il indique ensuite que pour maximiser cette précision il faut déterminer la relation entre les indicateurs. La méthode *M-SC* caractérise et améliore les connaissances quant à la structure et la dynamique du phytoplancton en réponse à des facteurs environnementaux multi-critères. Cette caractérisation permet l'identification des rôles multi-factoriels des mécanismes qui régissent cette dynamique. Comme exposé dans l'introduction de ce chapitre, c'est un point fondamental pour aborder les différentes problématiques environnementales liées aux changements régionaux et mondiaux de la zone côtière [Lombard et al. 2019]. Cette approche offre de nouvelles perspectives d'interprétation des processus et de développement de nouveaux indicateurs complémentaires via des informations intégrées. Ainsi, la méthode pourrait, par exemple, être appliquée en complément des indicateurs BF déjà mis en place par les différents consortiums scientifiques réunis par la DCE et la DCSMM. En plus d'apporter une vision multi-factoriels, son traitement adapté aux données hautes fréquences ajoute aussi une nouvelle dimension plus proche de la variabilité de certains processus

Par ailleurs, si l'on se focalise sur une problématique côtière plus ciblée, les efflorescences algales nuisibles (HABs) produisent des impacts locaux dans presque tous les systèmes d'eaux douces et marines. Il s'agit d'un problème qui survient à l'échelle mondiale et qui nécessite une compréhension scientifique intégrée et coordonnée, conduisant à des réponses et des solutions régionales [C. R. Anderson et al. 2019]. La méthode *M-SC* a la capacité de faire des distinctions inter-blooms et inter-spécifiques et permet l'identification des rôles multi-factoriels des processus environnementaux à l'échelle de la formation des blooms. Ainsi, elle pourrait être utilisée comme système d'alerte des HABs et aider à la gestion de ceux-ci. Avec une base de données plus précises qui intégrerait des variables rattachées directement aux communautés phytoplanctoniques (ex : identifications microscopiques, AOA, cytométrie) ou aux apports anthropiques (concentration de nutriment aux embouchures, débit fluvial...), cette approche pourrait même permettre une distinction entre les périodes caractéristiques, « naturelles » ou « liées à un facteur anthropique » des blooms de HABs. Dans le cadre de la zone Manche proche de MAREL-Carnot, cette perspective est totalement envisageable puisque les groupes *Phaeocystis* et *Phaeocystis* combinés à *Pseudo-nitzschia* ont été distingués par la méthode *M-SC* dans des états environnementaux séparés.

Enfin, ses capacités d'analyse et de prédiction d'états environnementaux en temps *quasi* réels peuvent contribuer à la mise en place d'une stratégie adaptative de l'échantillonnage pour l'observation, la surveillance, la recherche et la gestion de l'environnement. Comme cela a été démontré dans certain travaux [Lefebvre et Devreker 2019c , Lefebvre et Poisson-Caillault 2019], ainsi que dans le chapitre 5 de cette thèse, les méthodes de classification spectrale peuvent être un appui fiable pour la détermination d'éco-régions marines. La mise en place d'un outil de traitement et de valorisation les flux de données par le biais de cette approche est donc imaginable. Cela fournirait un découpage sur la base de paramètres physiques, hydrodynamiques et biologiques et chimiques, ce qui favoriserait l'étude des sites et la comparaison des stratégies d'observation établies pour un site de mesure ou lors d'une campagne (Ajustement de la fréquence d'échantillonnage, adaptation de la zone et de la période de mesure, ...). Il serait, ainsi possible de détecter les zones d'intérêts ou d'observer les évolutions à venir et ainsi faire des choix stratégiques, comme un échantillonnage supplémentaire ou la délocalisation d'un point de mesures, adapté aux processus. Il serait intéressant et possible de mettre en place cette perspective, lors d'une campagne écosystémique comme la CGFS ou l'IBTS en Manche mer du Nord (ainsi que des campagnes similaires dans d'autres régions) où des données FerryBox sont disponibles. Ces campagnes étant effectuées chaque année sur une période identique, il est possible de classer, de labelliser et d'apprendre les différentes éco-régions à partir des données passées et de vérifier la capacité du modèle à prédire en temps réel, à partir des nouvelles données FerryBox en direct lors de la campagne, le passage de tel à tel environnement afin d'optimiser l'échantillonnage.

DEFINITIONS

Indices de validation

RI : *l' indice de rand mesure la consistance (le taux d'accord) entre deux partitions,* ***plus sa valeur est proche de 1 plus le partitionnement est satisfaisant.***

ARI : *l' indice de rand ajusté est la version corrigée du hasard de l'indice de rand (RI),* ***plus sa valeur est proche de 1 plus le partitionnement est satisfaisant.*** *ARI peut générer des valeurs négatives si l'indice est inférieur à l'indice attendu.*

Dunn : *l'indice Dunn est le rapport de la plus petite distance entre les observations ne faisant pas partie du même groupe et de la plus grande distance intra-groupe. L'indice Dunn a une valeur comprise* ***entre zéro et l'infini et doit être maximisé.*** *Il reflète la séparabilité des données et des groupes. Un indice de Dunn proche de zéro indique des données très liées, tandis qu'un score élevé indique des données facilement séparables.*

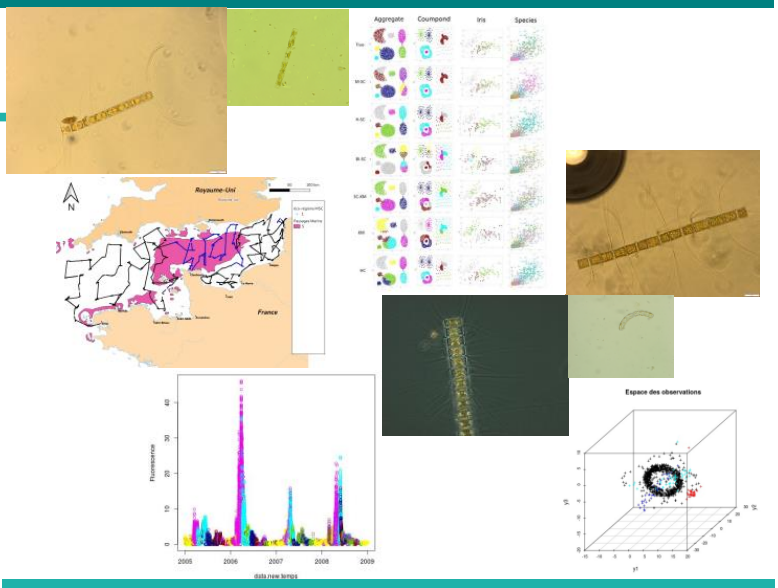
La Silhouette *est la mesure de la similitude d'un objet avec son propre cluster (cohésion) par rapport à d'autres clusters (séparation).* ***Une grande silhouette (presque 1) est très bien séparée,*** *tandis qu'une petite silhouette (environ 0) signifie que les observations sont connectées entre deux groupes.*

Le taux de précision *(Accuracy) est le rapport entre l'observation correctement prédite/classée et l'ensemble des observations. Elle reflète le pourcentage d'étiquettes bien reconnues selon une vérité terrain donnée, plus la précision est* ***proche de 1 plus notre modèle est précis.***

La précision *est le rapport entre les observations positives correctement prédites et le total des observations positives prédites.* ***Un résultat proche de 1 est un bon résultat.***

Le recall (Rappel ou sensibilité) *est le rapport entre les observations positives correctement prédites et toutes les observations de la classe réelle. Il répond à la question : combien de classe ont bien été prédites ?* ***Un Recall supérieur à 0,5 est considéré comme un bon score.***

Le F1 Score *est la moyenne pondérée de la précision et du rappel. Par conséquent, ce score tient compte à la fois des faux positifs et des faux négatifs. le F1 est généralement plus utile que la précision, surtout pour une distribution de classe inégale.* ***Un résultat proche de 1 est un bon résultat.***



Terminologies méthodologiques

L'apprentissage artificiel *(Machine Learning) englobe toutes les méthodes permettant de construire un modèle de la réalité à partir de données ; soit en améliorant un modèle préexistant, partiel ou moins général (Transfert Learning), soit en réant complètement le modèle.*

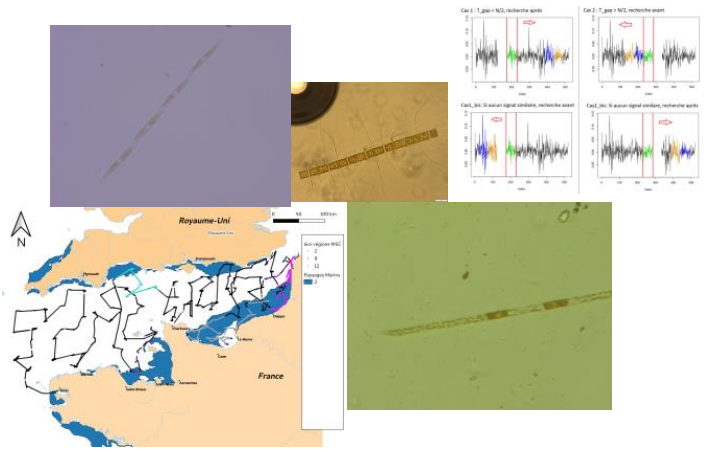
Classification et classe *seront employées pour désigner la phase de partitionnement des données. Les classes (notées c i pour la ième classe) désignent les groupes d'objets issus de la classification non supervisée ; Le mot classe est utilisé comme terme analogue à clusters.*

Labellisation et label *seront employés pour désigner la phase de caractérisation des classes par un expert scientifique. Le label (notés l i pour le ième label) désigne les classes après expertise.*

La classification non supervisée *(Clustering) consiste, à contrario, à identifier des groupes d'objets (classes) tels que les objets (ici les observations) les plus similaires soient dans le même groupe.*

La classification supervisée *(Classification) désigne le regroupement des classes à partir de leurs descripteurs en fonction des informations sur les données (ici les labels).*

Le modèle de reconnaissance et/ou de prédiction *est un modèle d'apprentissage construit à partir d'une méthode de classification non supervisée qui permet la détection, la classification et la prédiction d'un jeu de données non classé/labellisé.*



Terminologies biologiques

L'état environnemental *est la somme des conditions écologiques définissant un schéma de fonctionnement d'un écosystème.*

La phénologie *est l'étude des variations des phénomènes périodiques.*

Phénologique *est la modification des cycles de vie déterminés par des variations/perturbations du milieu.*

La fluorescence *est la mesure de la concentration en chlorophylle. Utilisée comme proxies de la biomasse phytoplanctonique.*

Essential Ocean Variables (EOVs) : *variables d'intérêts standardisées.*

Taxon : *entité d'êtres vivants regroupés parce qu'ils possèdent des caractères en commun du fait de leur parenté.*

L'eutrophisation *est un déséquilibre du milieu provoqué par l'augmentation de la concentration d'azote et de phosphore dans le milieu qui conduit à une prolifération importante d'algues, prolifération qui peut provoquer des phénomènes d'hypoxie, voire d'anoxie.*

HABs (Harmful Algal Blooms) : *efflorescences d'algues toxiques.*

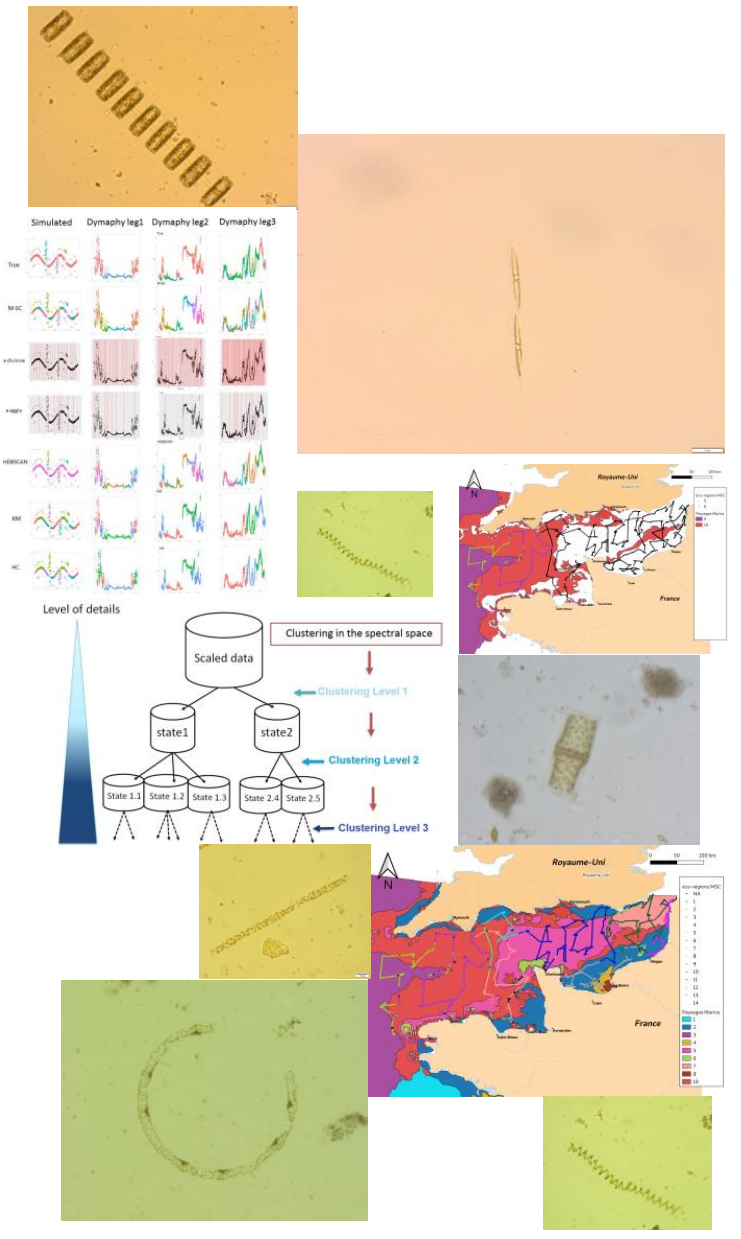
Hautes Fréquences (HF) : *mesures infra-horaires voir infra-journalières.*

Basses Fréquences (BF): *mesures bimensuelles, mensuelles voir quelquefois hebdomadaire.*

Notations

Nous désignerons pour la suite du document les notations suivantes

- X la matrice de données N × M , avec N le nombre d'objets de la base et M le nombre de dimensions formant l'espace de description des objets ;
- $\overrightarrow{x_n}$ = {x_{n1} }, . . . , x_{nM} pour n ∈ [1, N] le n-ième objet de la base ;
- K le nombre de groupes identifiés après segmentation automatique par classification non
- supervisée ;
- k ∈ 1, . . . , K le numéro du k-ième groupe ;
- cl_n ∈ 1, . . . , K le numéro de classe de l'objet n ;
- l_n le label de l'objet n après labellisation des K groupes par un expert qui sera utilisé pour un apprentissage supervisé.



Bibliographie

Albérola, C. et C. Millot (avr. 2003). « Circulation in the French mediterranean coastal zone near Marseilles : the influence of wind and the Northern Current ». en. In : *Continental Shelf Research* 23.6, p. 587-610. issn : 02784343. doi : 10.1016/S0278-4343(03)00002-5.

Albérola, C., C. Millot et J. Font (1995). « On the seasonal and mesoscale variabilities of the N orthem Current during Northem Current Ligurian Sea Flux V ariability Meanders the PRIM0-0 experiment in the western Mediterranean Sea ». In : *OCEANOLOGICA ACTA* 18.2, p. 163-192.

Allen, J. I. et L. Polimene (juil. 2011). « Linking physiology to ecology : towards a new generation of plankton models ». en. In : *Journal of Plankton Research* 33.7, p. 989-997. issn : 0142-7873, 1464-3774. doi : 10.1093/plankt/fbr032.

Amandine, S. (2010). « Impact du vent sur la circulation hydrodynamique dans le Golfe du Lion : modélisation haute résolution ». Thesis. FRANCE : Université du Sud Toulon-Var.

Anderson, C. R. et al. (mai 2019). « Scaling Up From Regional Case Studies to a Global Harmful Algal Bloom Observing System ». In : *Frontiers in Marine Science* 6, p. 250. issn : 2296-7745. doi : 10.3389/fmars.2019.00250.

Anderson, D. M., A. D. Cembella et G. M. Hallegraeff (1998). *Physiological ecology of harmul algal blooms*. Springer-Verlag. T. G41. NATO ASI. Springer-Verlag.

Annie, C. et al. (2015). « The Bay of Brest (France), a new risky site for toxic Alexandrium minutum blooms and PSP shellfish contamination ». en. In : *Harmful algae news* 51. Sous la dir. d'I. UNESCO, p. 4-5.

Belin, C., C. Beaucour, C. Bouchoucha et N. Ganzin (2012). *MÉDITERRANéE OCCIDENTALE : ÉTAT BIOLOGIQUE*
Caractéristiques biologiques - biocénoses Communautés du phytoplancton. DCE : CARACTÉRISTIQUES ET ÉTAT ÉCOLOGIQUE 17, p. 1-10.

Ben-Hur, A. et J. Weston (2010). « A User's Guide to Support Vector Machines ». In : *Data Mining Techniques for the Life Sciences*. Sous la dir. d'O. Carugo et F. Eisenhaber. T. 609. Series Title : Methods in Molecular Biology. Totowa, NJ : Humana Press, p. 223-239. isbn : 978-1-60327-240-7 978-1-60327-241-4. doi : 10.1007/978-1-60327-241-4_13.

Beucher, C. et al. (2004). « Production and dissolution of biosilica, and changing microphytoplankton dominance in the Bay of Brest (France) ». en. In : *Marine Ecology Progress Series* 267, p. 57-69. issn : 0171-8630, 1616-1599. doi : 10.3354/meps267057.

Beutler, M., K. Wilthshire et al. (2002). « fluorimetric method for the differentiation of algal populations in vivo and in situ. » In : *Photosynthesis Research* 72, p. 39-53.

Beutler, M., K. H. Wiltshire et al. (2002). « [No title found] ». In : *Photosynthesis Research* 72.1, p. 39-53. issn : 01668595. doi : 10.1023/A:1016026607048.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Advanced applications in pattern recognition. New York : Plenum Press. isbn : 9780306406713.

Billen, G. et J. Garnier (juil. 2007). « River basin nutrient delivery to the coastal sea : Assessing its potential to sustain new production of non-siliceous algae ». en. In : *Marine Chemistry* 106.1-2, p. 148-160. issn : 03044203. doi : 10.1016/j.marchem.2006.12.017.

Blanc, F., M. Leveau et K. H. Szekielda (1969). « Effets eutrophiques au débouché d'un grand fleuve (Grand Rhône) ». In : *Marine Biology* 3, p. 233-242.

Bojinski, S. et al. (jan. 2014). « The Concept of Essential Climate Variables in Support of Climate Research, Applications, and Policy ». In : *Bulletin of the American Meteorological Society* 95.9, p. 1431-1443. issn : 0003-0007. doi : 10.1175/BAMS-D-13-00047.1.

Borcard, D., F. Gillet et P. Legendre (2011). *Numerical Ecology with R*. en. Use R ! New York : Springer-Verlag. isbn : 9781441979766.

Bourrin, F. et X. Durrieu de Madron (déc. 2006). « Contribution to the study of coastal rivers and associated prodeltas to sediment supply in Gulf of Lions (N-W Mediterranean Sea) ». In : *Vie et Milieu* 56, p. 1-8.

Boyce, D. G., M. R. Lewis et B. Worm (juil. 2010). « Global phytoplankton decline over the past century ». en. In : *Nature* 466.7306, p. 591-596. issn : 0028-0836, 1476-4687. doi : 10.1038/nature09268.

Breiman, L. (1999). « Random forests ». In : *Machine learning* 45.1, p. 5-32.

Breiman, L. (oct. 2001). « Random Forests ». en. In : *Machine Learning* 45.1, p. 5-32. issn : 1573-0565. doi : 10.1023/A:1010933404324.

Breton, E., C. Brunet, S. Sautour et J. M. Brylinski (août 2000). « Annual variations of phytoplankton biomass in the Eastern English Channel : comparison by pigment signatures and microscopic counts ». In : *Journal of Plankton Research* 22.8, p. 1423-1440. issn : 14643774. doi : 10.1093/plankt/22.8.1423.

Breton, E., V. Rousseau et al. (mai 2006). « Hydroclimatic modulation of diatom *Phaeocystis* blooms in nutrient-enriched Belgian coastal waters (North Sea) ». en. In : *Limnology and Oceanography* 51.3, p. 1401-1409. issn : 00243590. doi : 10.4319/lo.2006.51.3.1401.

Broche, P. et al. (nov. 1998). « Experimental study of the Rhone plume. Part I : physics and dynamics ». en. In : *Oceanologica Acta* 21.6, p. 725-738. issn : 03991784. doi : 10.1016/S0399-1784(99)80002-4.

Brock, G., V. Pihur, S. Datta et S. Datta (2008). « clValid : An R Package for Cluster Validation ». In : *Journal of Statistical Software, Articles* 25.4, p. 1-22. issn : 1548-7660. doi : 10.18637/jss.v025.i04.

Caillault-Poisson, É. et A. Lefebvre (juin 2017). « Towards Chl-a bloom understanding by EM-based unsupervised event detection ». In : *OCEANS 2017 - Aberdeen*, p. 1-5. doi : 10.1109/OCEANSE.2017.8084597.

Chatzigeorgakidis, G., S. Karagiorgou, S. Athanasiou et S. Skiadopoulos (fév. 2018). « FML-kNN : scalable machine learning on Big Data using k-nearest neighbor joins ». In : *Journal of Big Data* 5.1. doi : 10.1186/s40537-018-0115-x.

Chauvaud, L., F. Jean, O. Ragueneau et G. Thouzeau (2000). « Long-term variation of the Bay of Brest ecosystem : benthic-pelagic coupling revisited ». In : *Marine Ecology Progress Series* 200, p. 35-48.

Chauvaud, L., G. Thouzeau et Y.-M. Paulet (sept. 1998). « Effects of environmental factors on the daily growth rate of Pecten maximus juveniles in the Bay of Brest (France) ». en. In : *Journal of Experimental Marine Biology and Ecology* 227.1, p. 83-111. issn : 00220981. doi : 10.1016/S0022-0981(97)00263-3.

Clabaut, T., D. Devreker et A. Lefebvre (2019). *Résultats de la mise en oeuvre des réseaux REPHY et SRN. Zones côtières de la Manche orientale et de la baie sud de la Mer du Nord. Bilan de l'année 2018*. Report (Scientific report). FRANCE.

Claquin, P. et al. (jan. 2010). « Effects of simulated benthic fluxes on phytoplankton dynamic and photosynthetic parameters in a mesocosm experiment (Bay of Brest, France) ». en. In : *Estuarine, Coastal and Shelf Science* 86.1, p. 93-101. issn : 02727714. doi : 10.1016/j.ecss.2009.10.017.

Cloern, J. E. et R. Dufford (2005). « Phytoplankton community ecology : principles applied in San Francisco Bay ». en. In : *Marine Ecology Progress Series* 285, p. 11-28. issn : 0171-8630, 1616-1599. doi : 10.3354/meps285011.

Cloern, J. E. (mai 1996). « Phytoplankton bloom dynamics in coastal ecosystems : A review with some general lessons from sustained investigation of San Francisco Bay, California ». In : *Reviews of Geophysics* 34.2, p. 127-168. issn : 87551209. doi : 10.1029/96RG00986.

Cloern, J. E., P. C. Abreu et al. (fév. 2016). « Human activities and climate variability drive fast-paced change across the world's estuarine-coastal ecosystems ». en. In : *Global Change Biology* 22.2, p. 513-529. issn : 13541013. doi : 10.1111/gcb.13059.

Conan, P., M. Pujo-pay, P. Raimbault et M. Leveau (1998). « Variabilite hydrologique et biologique du golfe du Lion. II. Productivitk sur le bord interne du courant ». In : *Oceanoligia Acta* 21.3, p. 767-782.

Coppin, F. (1988). « CGFS : CHANNEL GROUND FISH SURVEY ». In : doi : 10.18142/11.

Coppin, F. et M. Travers -Trolet (2017). *Rapport de campagne CGFS 2017*. Rapp. tech. 53516.

Cornuéjols, A. et L. Miclet (2010). *Apprentissage artificiel*. French. OCLC : 654315256. Paris : Eyrolles. isbn : 9782212124712.

Coste, B., J. Gostan et H. J. Minas (1972). « Influence des conditions hivernales sur les productions phyto- et zooplanctoniques en Méditerranée Nord-Occidentale. III. Caractérisation des eaux de surface au moyen de cultures d'algues ». In : *Marine Biology* 16, p. 320-348.

Courp, T. et A. Monaco (sept. 1990). « Sediment dispersal and accumulation on the continental margin of the Gulf of Lions : sedimentary budget ». en. In : *Continental Shelf Research* 10.9-11, p. 1063-1087. issn : 02784343. doi : 10.1016/0278-4343(90)90075-W.

Cover, T. et P. Hart (jan. 1967). « Nearest neighbor pattern classification ». In : *IEEE Transactions on Information Theory* 13.1, p. 21-27. issn : 0018-9448, 1557-9654. doi : 10.1109/TIT.1967.1053964.

Crenan, B. (2019). *N/O Thalassa, Guide d'utilisation et de gestion de la Ferrybox.NE-2019-591.* Rapp. tech.

Cullen, J., W. Ford Doolittle, S. Levin et W. Li (1er juin 2007). « Patterns and Prediction in Microbial Oceanography ». In : *Oceanography* 20.2, p. 34-46. issn : 10428275. doi : 10.5670/oceanog.2007.46.

Cullen, J. J., P. J. Franks, D. M. Karl et A. Longhurst (2002). « Physical influences on marine ecosystem dynamics. » In : 12, p. 298-336.

Cushing, D. H. (1996). « The seasonal variation in oceanic production as a problem in population dynamics. » In : 24, p. 455-464.

Dauvin, J.-C. (jan. 2008). « The main characteristics, problems, and prospects for Western European coastal seas ». en. In : *Marine Pollution Bulletin* 57.1-5, p. 22-40. issn : 0025326X. doi : 10.1016/j.marpolbul.2007.10.016.

Dauvin, J.-C. et O. Lozachmeur (déc. 2006). « Mer côtière à forte pression anthropique propice au développement d'une gestion intégrée :
exemple du bassin oriental de la Manche (Atlantique Nord-Est) ». fr. In : *VertigO - la revue électronique en sciences de l'environnement* Volume 7 Numéro 3. issn : 1492-8442. doi : 10.4000/vertigo.1914.

David, V. et al. (août 2012). « Spatial and long-term changes in the functional and structural phytoplankton communities along the French Atlantic coast ». en. In : *Estuarine, Coastal and Shelf Science* 108, p. 37-51. ISSN : 02727714. DOI : 10.1016/j.ecss.2012.02.017.

DCE (2000). *Directive 2000/60/CE du Parlement Européen et du Conseil*. Rapp. tech. L327. Journal officiel des Communautés européennes, p. 1-72.

DCSMM (2008). *Directive 2008/56/CE du Parlement Européen et du Conseil*. FR. Journal officiel de l'Union européenne L 164, p. 19-40.

Del Amo, Y. et al. (1997). « Impacts of high-nitrate freshwater inputs on macrotidal ecosystems. I. Seasonal evolution of nutrient limitation for the diatom-dominated phytoplankton of the Bay of Brest (France) ». en. In : *Marine Ecology Progress Series* 161, p. 213-224. ISSN : 0171-8630, 1616-1599. DOI : 10.3354/meps161213.

Delegrange, A. et al. (déc. 2018). « Pseudo-nitzschia sp. diversity and seasonality in the southern North Sea, domoic acid levels and associated phytoplankton communities ». en. In : *Estuarine, Coastal and Shelf Science* 214, p. 194-206. ISSN : 02727714. DOI : 10.1016/j.ecss.2018.09.030.

Delmas, R. et P. Tréguer (1983). « Evolution saisonnièe8re des nutriments dans un écosystème eutrophe d'Europe occidentale (la rade de Brest). Interactions marines et terrestres ». In : *OCEANOLOGICA ACTA* 6.4, p. 345-356.

Devreker, D. et A. Lefebvre (2020). « TTAinterfaceTrendAnalysis : An R GUI for routine temporal trend analysis and diagnostics ». In : *J. Oceanogr. Res. Data,* 1(7), p. 1-18.

Dias, J., J. Vermunt et S. Ramos (2015). « Clustering financial time series : New insights from an extended hidden Markov model ». In : *EJOR* 243.3, p. 852-864. ISSN : 0377-2217. DOI : https://doi.org/10.1016/j.ejor.2014.12.041.

Dickey, T. D. (avr. 2003a). « Emerging ocean observations for interdisciplinary data assimilation systems ». en. In : *Journal of Marine Systems* 40-41, p. 5-48. ISSN : 09247963. DOI : 10.1016/S0924-7963(03)00011-3.

— (avr. 2003b). « Emerging ocean observations for interdisciplinary data assimilation systems ». en. In : *Journal of Marine Systems* 40-41, p. 5-48. ISSN : 09247963. DOI : 10.1016/S0924-7963(03)00011-3.

Dua, D. et C. Graff (2017). *UCI Machine Learning Repository*.

Dunn, J. C. (jan. 1973). « A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters ». en. In : *Journal of Cybernetics* 3.3, p. 32-57. ISSN : 0022-0280. DOI : 10.1080/01969727308546046.

DYMAPHY Project (2020).

DYPHYMA (2012). *Continuous Phytoplankton measurements (DYPHYMA), au Springer 2012, en Eastern Channel*.

Edwards, M. et A. J. Richardson (août 2004). « Impact of climate change on marine pelagic phenology and trophic mismatch ». en. In : *Nature* 430.7002, p. 881-884. ISSN : 0028-0836, 1476-4687. DOI : 10.1038/nature02808.

Emonet, R., J. Varadarajan et J.-M. Odobez (2014). « Temporal Analysis of Motif Mixtures Using Dirichlet Processes ». In : *IEEE Trans. Pattern Anal. Mach. Intell.* 36.1, p. 140-156. DOI : 10.1109/TPAMI.2013.100.

Eppley, R. (1972). « Temperature and phytoplankton growth in the sea ». In : *Fisheries Bulletin* 70.4, p. 1063-1085.

Ester, M., H. Kriegel, J. Sander, X. Xu et al. (1996). « A density-based algorithm for discovering clusters in large spatial databases with noise. » In : *KDD-96 Proceedings*.

Farcy, P. et al. (sept. 2019). « Toward a European Coastal Observing Network to Provide Better Answers to Science and to Societal Challenges ; The JERICO Research Infrastructure ». In : *Frontiers in Marine Science* 6, p. 529. ISSN : 2296-7745. DOI : 10.3389/fmars.2019.00529.

Flexas, M. et al. (2002). « Flow variability in the Gulf of Lions during the MATER HFF experiment(March-May 1997). » In : *Journal of Marine Systems* 33-34, p. 197-214.

Fraysse, M. (2014). « Rôle du forçage physique sur l'écosystème à l'est du Golfe du Lion : modulation de l'impact des apports anthropiques en sels nutritifs et matière organique étudiée par modélisation 3D couplée physique et biogéochimique. » 2014AIXM4101. Thèse de doct.

Gailhard, I. et al. (2002). « Variability patterns of microphytoplankton communities along the French coasts ». en. In : *Marine Ecology Progress Series* 242, p. 39-50. issn : 0171-8630, 1616-1599. doi : 10.3354/meps242039.

Garmendia, M., A. Borja, J. Franco et M. Revilla (jan. 2013). « Phytoplankton composition indicators for the assessment of eutrophication in marine waters : Present state and challenges within the European directives ». en. In : *Marine Pollution Bulletin* 66.1-2, p. 7-16. issn : 0025326X. doi : 10.1016/j.marpolbul.2012.10.005.

Gattuso, J.-P., M. Frankignoulle et R. Wollast (1998). « CARBON AND CARBONATE METABOLISM IN COASTAL AQUATIC ECOSYSTEMS ». In : *Annual Review of Ecology and Systematics* 29.1, p. 405-434. doi : 10.1146/annurev.ecolsys.29.1.405.

Gentilhomme, V. et F. Lizon (1997). « [No title found] ». In : *Hydrobiologia* 361.1/3, p. 191-199. issn : 00188158. doi : 10.1023/A:1003134617808.

Gieskes, W. W. C. et al. (mai 2007). « Phaeocystis colony distribution in the North Atlantic Ocean since 1948, and interpretation of long-term changes in the Phaeocystis hotspot in the North Sea ». en. In : *Biogeochemistry* 83.1-3, p. 49-60. issn : 0168-2563, 1573-515X. doi : 10.1007/s10533-007-9082-6.

Glibert, P. M. (mai 2016). « Margalef revisited : A new phytoplankton mandala incorporating twelve dimensions, including nutritional physiology ». en. In : *Harmful Algae* 55, p. 25-30. issn : 15689883. doi : 10.1016/j.hal.2016.01.008.

Gomez, F. et S. Souissi (sept. 2008). « The impact of the 2003 summer heat wave and the 2005 late cold wave on the phytoplankton in the north-eastern English Channel ». en. In : *Comptes Rendus Biologies* 331.9, p. 678-685. issn : 16310691. doi : 10.1016/j.crvi.2008.06.005.

Grassi, K., E. Poisson-Caillault, A. Bigand et A. Lefebvre (sept. 2020). « Comparative Study of Clustering Approaches Applied to Spatial or Temporal Pattern Discovery ». en. In : *Journal of Marine Science and Engineering* 8.9, p. 713. issn : 2077-1312. doi : 10.3390/jmse8090713.

Grassi, K., E. Poisson-Caillault et A. Lefebvre (juin 2019). « Multilevel Spectral Clustering for extreme event characterization ». In : *OCEANS 2019 - Marseille*. Marseille, France : IEEE, p. 1-7. isbn : 9781728114507. doi : 10.1109/OCEANSE.2019.8867261.

Grossel, H. (2016). *Manuel d'observation et de dénombrement du phytoplancton marin. Document de méthode REPHY*.

Gwatkin, R. B. L. (août 1995). « Practical flow cytometry, by H.M. Shapiro, Wiley-Liss, New York, 3rd ed, 1994, 542 pp, $79.95 ». en. In : *Molecular Reproduction and Development* 41.4, p. 530-530. issn : 1040-452X, 1098-2795. doi : 10.1002/mrd.1080410419.

Hamerly, G. et C. Elkan (déc. 2003). « Learning the k in k-means ». In : *Proceedings of the 16th International Conference on Neural Information Processing Systems*. NIPS'03. Whistler, British Columbia, Canada : MIT Press, p. 281-288.

Hartigan, J. A. et M. A. Wong (1979). « Algorithm AS 136 : A K-Means Clustering Algorithm ». In : *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, p. 100-108. issn : 0035-9254. doi : 10.2307/2346830.

Hendon, H. H. et B. Liebmann (déc. 1989). « The Intraseasonal (30–50 day) Oscillation of the Australian Summer Monsoon ». In : *Journal of the Atmospheric Sciences* 47.24, p. 2909-2924. issn : 0022-4928. doi : 10.1175/1520-0469(1990)047<2909:TIDOOT>2.0.CO;2.

Hernández, F. T. et al. (juin 2014). « Temporal changes in the phytoplankton community along the French coast of the eastern English Channel and the southern Bight of the North Sea ». en. In : *ICES Journal of Marine Science* 71.4, p. 821-833. issn : 1095-9289, 1054-3139. doi : 10.1093/icesjms/fst192.

Herrmann, M. (2007). « Formation et devenir des masses d'eau en Méditerranée nord-occidentale : influence sur l'écosystème planctonique pélagique : variabilité inter-annuelle et changement climatique ». 2007TOU30322. Thèse de doct., 1 vol. (308 p.)

Hoek, C. v. d., D. G. Mann et H. M. Jahns (1995). *Algae : an introduction to phycology*. eng. Cambridge ; New York : Cambridge University Press. isbn : 9780521304191.

Huang, N. E. et al. (mar. 1998). « The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis ». In : *Proceedings of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences* 454.1971, p. 903-995. doi : 10.1098/rspa.1998.0193.

Ignatiades, L. (1994). « Species dominance and niche breadth in bloom and non-bloom phytoplankton populations ». In : *Oceanologica Acta* 17, p. 89-96.

Jain, A. K. (juin 2010). « Data Clustering : 50 Years Beyond K-means ». In : *Pattern Recogn. Lett.* 31.8, p. 651-666. issn : 0167-8655. doi : 10.1016/j.patrec.2009.09.011.

James, N. et D. Matteson (2015). « ecp : An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data ». In : *Journal of Statistical Software* 62.7, p. 1-25. issn : 1548-7660. doi : 10.18637/jss.v062.i07.

Joseph, L. et C. Moritz (1994). « Mitochondrial-Dna Phylogeography of Birds in Eastern Australian Rain-Forests - First Fragments ». en. In : *Australian Journal of Zoology* 42.3, p. 385. issn : 0004-959X. doi : 10.1071/ZO9940385.

Karasiewicz, S. et al. (fév. 2018). *Realized niche analysis of phytoplankton communities involving HAB :*
Phaeocystis spp. as a case study. en. T. 72, p. 1-13. doi : 10.1016/j.hal.2017.12.005.

Karl, D. M. et M. J. Church (avr. 2017). « Ecosystem Structure and Dynamics in the North Pacific Subtropical Gyre : New Views of an Old Ocean ». In : *Ecosystems* 20.3, p. 433-457. issn : 1435-0629. doi : 10.1007/s10021-017-0117-0.

Karydis, M. et D. Kitsiou (août 2012). « Eutrophication and environmental policy in the Mediterranean Sea : a review ». en. In : *Environmental Monitoring and Assessment* 184.8, p. 4931-4984. issn : 0167-6369, 1573-2959. doi : 10.1007/s10661-011-2313-2.

Kissling, W. D. et al. (fév. 2018). « Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale ». en. In : *Biological Reviews* 93.1, p. 600-625. issn : 1469-185X. doi : 10.1111/brv.12359.

Kong, W., S. Hu, J. Zhang et G. Dai (oct. 2013). « Robust and smart spectral clustering from normalized cut ». en. In : *Neural Computing and Applications* 23.5, p. 1503-1512. issn : 1433-3058. doi : 10.1007/s00521-012-1101-4.

Kupschus, S., M. Schratzberger et D. Righton (août 2016). « Practical implementation of ecosystem monitoring for the ecosystem approach to management ». en. In : *Journal of Applied Ecology* 53.4. Sous la dir. de J. Blanchard, p. 1236-1247. issn : 00218901. doi : 10.1111/1365-2664.12648.

Lagadeuc, Y., J. M. Brylinski et D. Aelbrecht (août 1997). « Temporal variability of the vertical stratification of a front in a tidal Region Of Freshwater Influence (ROFI) system ». en. In : *Journal of Marine Systems* 12.1-4, p. 147-155. issn : 09247963. doi : 10.1016/S0924-7963(96)00094-2.

Lancelot, C. et S. Mathot (1987). « Dynamics of a Phaeocystis-dominated spring bloom in Belgian coastal waters. I. Phytoplankton activities and related parameters. » In : *Marine Ecology Progress Series* 37, p. 239-248.

Längkvist, M., L. Karlsson et A. Loutfi (2014). « A review of unsupervised feature learning and deep learning for time-series modeling ». In : *Pattern Recognition Letters* 42, p. 11-24. ISSN : 0167-8655. DOI : https://doi.org/10.1016/j.patrec.2014.01.008.

Le Jehan, P. et P. Tréguer (1984). « Evolution saisonnière de composés organiques dissous dans un écosystème eutrophe d'Europe occidentale (rade de Brest) ». In : *OCEANOLOGICA ACTA* 7.2, p. 181-190.

Le Pape, O. et A. Menesguen (août 1997). « Hydrodynamic prevention of eutrophication in the Bay of Brest (France), a modelling approach ». en. In : *Journal of Marine Systems* 12.1-4, p. 171-186. ISSN : 09247963. DOI : 10.1016/S0924-7963(96)00096-6.

Le Pape, O. et al. (jan. 1996). « Resistance of a coastal ecosystem to increasing eutrophic conditions : the Bay of Brest (France), a semi-enclosed zone of Western Europe ». en. In : *Continental Shelf Research* 16.15, p. 1885-1907. ISSN : 02784343. DOI : 10.1016/0278-4343(95)00068-2.

Lefebvre, A. et D. Devreker (2019a). *Contributions des mesures automatisées à haute fréquence de type FerryBox pour les programmes thématiques Eutrophisation et Habitats Pélagiques de la DCSMM. Campagnes 2018*. Ifremer. DOI : 10.13155/70594.

Lefebvre, A. et E. Poisson-Caillault (jan. 2019). « High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering ». en. In : *Marine Ecology Progress Series* 608, p. 73-92. ISSN : 0171-8630, 1616-1599. DOI : 10.3354/meps12781.

Lefebvre, A. (2015). *MAREL Carnot data and metadata from Coriolis Data Centre. SEANOE.*

Lefebvre, A. et D. Devreker (2019b). « Ferry Box Data Manager. Enregistrement au répertoire IDDN de l'Agence pour la Protection des Programmes. Logiciel L35 - Interface données Ferrybox / Thalassa ». In :

Lefebvre, A. et D. Devreker (2019c). *Contributions des mesures automatisées à haute fréquence de type FerryBox pour les programmes thématiques Eutrophisation et Habitats Pélagiques de la DCSMM. Campagnes 2018*. Ifremer. DOI : 10.13155/70594.

Lefebvre, A., N. Guiselin, F. Barbet et F. L. Artigas (nov. 2011). « Long-term hydrological and phytoplankton monitoring (1992-2007) of three potentially eutrophic systems in the eastern English Channel and the Southern Bight of the North Sea ». en. In : *ICES Journal of Marine Science* 68.10, p. 2029-2043. ISSN : 1095-9289. DOI : 10.1093/icesjms/fsr149.

Lefevre, D. et al. (1997). « Review of gross community production, primary production, net community production and dark community respiration in the Gulf of Lions ». en. In : *Deep Sea Research Part II : Topical Studies in Oceanography* 44.3-4, p. 801-832. ISSN : 09670645. DOI : 10.1016/S0967-0645(96)00091-4.

Lewis, M. R. et al. (sept. 1984). « Turbulent motions may control phytoplankton photosynthesis in the upper ocean ». en. In : *Nature* 311.5981, p. 49-50. ISSN : 0028-0836, 1476-4687. DOI : 10.1038/311049a0.

Lindstrom, E. et al. (juin 2012). *A Framework for Ocean Observing*. Rapp. tech. European Space Agency. DOI : 10.5270/OceanObs09-FOO.

Lochet, F. et M. Leveau (nov. 1990). « Transfers between a eutrophic ecosystem, the river Rhône, and an oligotrophic ecosystem, the north-western Mediterranean Sea ». en. In : *Hydrobiologia* 207.1, p. 95-103. ISSN : 0018-8158, 1573-5117. DOI : 10.1007/BF00041445.

Lohrenz, S. E. et al. (mai 1988). « Interrelationships among primary production, chlorophyll, and environmental conditions in frontal regions of the western Mediterranean Sea ». en. In : *Deep Sea Research Part A. Oceanographic Research Papers* 35.5, p. 793-810. ISSN : 01980149. DOI : 10.1016/0198-0149(88)90031-3.

Lombard, F. et al. (avr. 2019). « Globally Consistent Quantitative Observations of Planktonic Ecosystems ». In : *Frontiers in Marine Science* 6, p. 196. ISSN : 2296-7745. DOI : 10.3389/fmars.2019.00196.

Longhurst, A. R. et W. Glen Harrison (jan. 1989). « The biological pump : Profiles of plankton production and consumption in the upper ocean ». en. In : *Progress in Oceanography* 22.1, p. 47-123. ISSN : 00796611. DOI : 10.1016/0079-6611(89)90010-4.

Lorthiois, T. (nov. 2012). « Dynamic of suspended matter within the Rhône river plume (western mediterranean sea) based on ocean colour remote sensing ». Theses. Université Pierre et Marie Curie - Paris VI.

Loureiro, S. et al. (jan. 2009). « The significance of organic nutrients in the nutrition of Pseudo-nitzschia delicatissima (Bacillariophyceae) ». en. In : *Journal of Plankton Research* 31.4, p. 399-410. ISSN : 0142-7873, 1464-3774. DOI : 10.1093/plankt/fbn122.

Ludwig, W., E. Dumont, M. Meybeck et S. Heussner (mar. 2009). « River discharges of water and nutrients to the Mediterranean and Black Sea : Major drivers for ecosystem changes during past and future decades ? » en. In : *Progress in Oceanography* 80.3-4, p. 199-217. ISSN : 00796611. DOI : 10.1016/j.pocean.2009.02.001.

Lundholm, N. (2018). *IOC-UNESCO Taxonomic Reference List of Harmful Micro Alga (HABs)*.

Luxburg, U. von (nov. 2007). « A Tutorial on Spectral Clusterin ». In : *arXiv :0711.0189 [cs]*. arXiv : 0711.0189.

Maechler, M. et al. (2018). *cluster : Cluster Analysis Basics and Extensions*.

Magdalena (août 2018). *Strengthening Europe's Capability in Biological Ocean Observations*. en-US.

Maillet, G. M. et al. (déc. 2006). « Morphological changes and sedimentary processes induced by the December 2003 flood event at the present mouth of the Grand Rhône River (southern France) ». en. In : *Marine Geology* 234.1-4, p. 159-177. ISSN : 00253227. DOI : 10.1016/j.margeo.2006.09.025.

Margalef, R. (jan. 1978). « Life-forms of phytoplankton as survival alternatives in an unstable environment ». en. In : *Oceanologica Acta*.

Margulis, L. et K. V. Schwartz (1998). *Five kingdoms : an illustrated guide to the phyla of life on earth*. 3rd ed. New York : W.H. Freeman. ISBN : 9780716730262.

Marsaleix, P., C. Estournel, V. Kondrachoff et R. Vehil (jan. 1998). « A numerical study of the formation of the Rhône River plume ». en. In : *Journal of Marine Systems* 14.1-2, p. 99-115. ISSN : 09247963. DOI : 10.1016/S0924-7963(97)00011-0.

Martin, S. et al. (sept. 2006). « Respiration, calcification, and excretion of the invasive slipper limpet, Crepidula fornicata L. : Implications for carbon, carbonate, and nitrogen fluxes in affected areas ». en. In : *Limnology and Oceanography* 51.5, p. 1996-2007. ISSN : 00243590. DOI : 10.4319/lo.2006.51.5.1996.

Medlin, L. K., K. Piwosz et K. Metfies (2017). « Uncovering hidden biodiversity in the Cryptophyta : Clone library studies at the Helgoland Time Series Site in the Southern German Bight indentifies the cryptophycean clade potentially responsible for the majority of its genetic diversity during the spring bloom. » In : *Vie et Millieu - Life and Environment* 67, p. 27-32.

Millot, C. (sept. 1990). « The Gulf of Lions' hydrodynamics ». en. In : *Continental Shelf Research* 10.9-11, p. 885-894. ISSN : 02784343. DOI : 10.1016/0278-4343(90)90065-T.

Miloslavich, P. et al. (juin 2018). « Essential ocean variables for global sustained observations of biodiversity and ecosystem changes ». en. In : *Global Change Biology* 24.6, p. 2416-2433. ISSN : 1365-2486. DOI : 10.1111/gcb.14108.

Minas, H. J. (1968). « Recherches sur la production organique primaire dans le bassin méditerranéen nord-occidental. Rapports avec les phénomènes hydrologiques. » Thèse de doct. Dr. Fac. Sci. Univ. Aix-Marscille.

Moestrup, O. et al. (2009). « IOC-UNESCO Taxonomic Reference List of Harmful Micro Algae ». In : doi : https://doi.org/10.14284/362.

Monaco, A., éd. (2009). *Le golfe du Lion : un observatoire de l'environnement en Méditerranée*. Update sciences & technologies. Versailles : Quae. isbn : 9782759203116.

Morel, A. et J.-M. André (1991). « Pigment distribution and primary production in the western Mediterranean as derived and modeled from coastal zone color scanner observations ». en. In : *Journal of Geophysical Research* 96.C7, p. 12685. issn : 0148-0227. doi : 10.1029/91JC00788.

Morin, P. et al. (2015). *The joint European Research Infrastructure Network for Coastal Observatories : Achievements and Strategy for the Future*. Report (Contract report).

Muniz Piniella, A. et al. (juil. 2018). *Strengthening Europe's Capability in Biological Ocean Observations*. isbn : 9789492043559.

Ng, A. Y., M. I. Jordan et Y. Weiss (2001). « On Spectral Clustering : Analysis and an Algorithm ». In : *Proceedings of the 14th International Conference on Neural Information Processing Systems : Natural and Synthetic*. NIPS'01. Cambridge, MA, USA : MIT Press, p. 849-856.

Officer, C. et J. Ryther (1980). « The Possible Importance of Silicon in Marine Eutrophication ». en. In : *Marine Ecology Progress Series* 3, p. 83-91. issn : 0171-8630, 1616-1599. doi : 10.3354/meps003083.

Omotosho, J. '. (mai 1992). « Long-range prediction of the onset and end of the rainy season in the West African Sahel ». en. In : *International Journal of Climatology* 12.4, p. 369-382. issn : 08998418, 10970088. doi : 10.1002/joc.3370120405.

OSPAR (1992). *Convention pour la protection du milieu marin de l'Atlantique du Nord-Est*. Rapp. tech., p. 176.

Ouillon, S. et A. Petrenko (2005). « Above-water measurements of reflectance and chlorophyll-a algorithms in the Gulf of Lions, NW Mediterranean Sea ». en. In : *Optics Express* 13.7, p. 2531. issn : 1094-4087. doi : 10.1364/OPEX.13.002531.

Pairaud, I. et al. (oct. 2011). « Hydrology and circulation in a coastal area off Marseille : Validation of a nested 3D model with observations ». en. In : *Journal of Marine Systems* 88.1, p. 20-33. issn : 09247963. doi : 10.1016/j.jmarsys.2011.02.010.

Petrenko, A. (sept. 2003). « Variability of circulation features in the Gulf of Lion NW Mediterranean Sea. Importance of inertial currents ». en. In : *Oceanologica Acta* 26.4, p. 323-338. issn : 03991784. doi : 10.1016/S0399-1784(03)00038-0.

Phan, T., E. Poisson-Caillault, A. Lefebvre et A. Bigand (2017). « Dynamic time warping-based imputation for univariate time series data ». In : *Pattern Recognition Letters*. issn : 0167-8655. doi : 10.1016/j.patrec.2017.08.019.

Philippart, C. J. M., G. C. Cadée, W. van Raaphorst et R. Riegman (jan. 2000). « Long-term phytoplankton-nutrient interactions in a shallow coastal sea : Algal community structure, nutrient budgets, and denitrification potential ». en. In : *Limnology and Oceanography* 45.1, p. 131-144. issn : 00243590. doi : 10.4319/lo.2000.45.1.0131.

Phlips, E. J., S. Badylak, E. Bledsoe et M. Cichra (2006). « Factors affecting the distribution of Pyrodiniumbahamense var.bahamensein coastal watersof Florida ». In : *MARINE ECOLOGY PROGRESS SERIES* 322, p. 99-115.

Planque, B., E. Bellier et C. Loots (juil. 2011). « Uncertainties in projecting spatial distributions of marine populations ». en. In : *ICES Journal of Marine Science* 68.6, p. 1045-1050. issn : 1095-9289, 1054-3139. doi : 10.1093/icesjms/fsr007.

Poisson-Caillault, E. (2020). « Contributions à la classification et segmentation de séries temporelles par apprentissage statistique non supervisé ou guidé. » Habilitation à diriger des recherches. Université Côte d'Opale.

Poisson-Caillault, E. et A. Lefebvre (juin 2017). « Towards Chl-a bloom understanding by EM-based unsupervised event detection ». In : p. 1-5. doi : 10.1109/OCEANSE.2017.8084597.

Pont, D., J.-P. Simonnet et A. Walter (jan. 2002). « Medium-term Changes in Suspended Sediment Delivery to the Ocean : Consequences of Catchment Heterogeneity and River Management (Rhône River, France) ». en. In : *Estuarine, Coastal and Shelf Science* 54.1, p. 1-18. issn : 02727714. doi : 10.1006/ecss.2001.0829.

Quéguiner, B. et P. Tréguer (1984). « Studies on the Phytoplankton in the Bay of Brest (Western Europe). Seasonal Variations in Composition, Biomass and Production in Relation to Hydrological and Chemical Features (1981 — 1982) ». In : *Botanica Marina* 27.10. issn : 0006-8055, 1437-4323. doi : 10.1515/botm.1984.27.10.449.

Racault, M.-F. et al. (mar. 2012). « Phytoplankton phenology in the global ocean ». en. In : *Ecological Indicators* 14.1, p. 152-163. issn : 1470160X. doi : 10.1016/j.ecolind.2011.07.010.

Ragueneau, O., L. Chauvaud, A. Leynaert et al. (nov. 2002). « Direct evidence of a biologically active coastal silicate pump : Ecological implications ». en. In : *Limnology and Oceanography* 47.6, p. 1849-1854. issn : 0024-3590, 1939-5590. doi : 10.4319/lo.2002.47.6.1849.

Ragueneau, O., L. Chauvaud, B. Moriceau et al. (août 2005). « Biodeposition by an Invasive Suspension Feeder Impacts the Biogeochemical Cycle of Si in a Coastal Ecosystem (Bay of Brest, France) ». en. In : *Biogeochemistry* 75.1, p. 19-41. issn : 0168-2563, 1573-515X. doi : 10.1007/s10533-004-5677-3.

Ragueneau, O., E. De Bias Varela et al. (1994). « Phytoplankton dynamics in relation to the biogeochemical cycle of silicon in a coastal ecosystem of western Europe ». en. In : *Marine Ecology Progress Series* 106, p. 157-172. issn : 0171-8630, 1616-1599. doi : 10.3354/meps106157.

Ragueneau, O., M. Raimonet et al. (avr. 2018). « The Impossible Sustainability of the Bay of Brest ? Fifty Years of Ecosystem Changes, Interdisciplinary Knowledge Construction and Key Questions at the Science-Policy-Community Interface ». In : *Frontiers in Marine Science* 5, p. 124. issn : 2296-7745. doi : 10.3389/fmars.2018.00124.

Reffray, G., P. Fraunié et P. Marsaleix (mai 2004). « Secondary flows induced by wind forcing in the Rhône region of freshwater influence ». In : *Ocean Dynamics* 54.2, p. 179-196. issn : 1616-7341, 1616-7228. doi : 10.1007/s10236-003-0079-y.

Reghunath, K. (fév. 2017). « Real Time Intrusion Detection System for Big Data ». In : *International Journal of Peer to Peer Networks* 08.1, p. 01-20. issn : 22295240, 22293930. doi : 10.5121/ijp2p.2017.8101.

REPHY-SEANOE (2019). *REPHY dataset - French Observation and Monitoring program for Phytoplankton and Hydrology in coastal waters. 1987-2018 Metropolitan data.* type : dataset. doi : 10.17882/47248.

Reynaud, J.-Y. et al. (2003). « The offshore Quaternary sediment bodies of the English Channel and its Western Approaches ». In : *Quaternary Science* 18, p. 361-371.

Reynolds, C. S. (1995). *Successional Change in the Planktonic Vegetation : Species, Structures, Scales | SpringerLink*. Springer-Verlag. Berlin : Joint I (ed) The molecular ecology of aquatic microbes.

Reynolds, C. S. (2006). *Ecology of phytoplankton*. English. OCLC : 76416312. Cambridge ; New York : Cambridge University Press.

Rodríguez, F. et al. (août 2000). « Temporal variability of viruses, bacteria, phytoplankton and zooplankton in the western English Channel off Plymouth ». en. In : *Journal of the Marine Biological Association of the United Kingdom* 80.4, p. 575-586. issn : 0025-3154, 1469-7769. doi : 10.1017/S0025315400002393.

Roger, D. (1981). « Etude de l'évolution saisonnière des sels nutritifs dans la Rade de Brest en fonction des apports fluviaux et des échanges avec l'Iroise. » Thesis. UBO.

Rousseeuw, K., É. Poisson-Caillault, A. Lefebvre et D. Hamad (jan. 2015a). « Hybrid Hidden Markov Model for Marine Environment Monitoring ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.1, p. 204-213. ISSN : 1939-1404. DOI : 10.1109/JSTARS.2014.2341219.

— (jan. 2015b). « Hybrid Hidden Markov Model for Marine Environment Monitoring ». In : *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8.1, p. 204-213. ISSN : 1939-1404. DOI : 10.1109/JSTARS.2014.2341219.

Ruser, A. et al. (1999). « Comparison of chlorophyll-fluorescence based measuring systems for the detection af algal groups and the determination of chlorophyll-a concentrations ». In : *Berichte Forsch.-u. technologiezentr. Westküste d. Univ. Kiel* 19, p. 27-38.

Ryckaert, M., P. Gros et E. Erard-Le Denn (1983). « Seasonal succession of coastal phytoplanctonic populations in the Channel ». In : *OCEANOLOGICA ACTA* SP, p. 171-175.

Sanchez-Garcia, J., M. Fennelly et S. N. et al. (sept. 2014). « Hierarchical Spectral Clustering of Power Grids ». In : *IEEE Transactions on Power Systems* 29.5, p. 2229-2237. ISSN : 0885-8950. DOI : 10.1109/TPWRS.2014.2306756.

Sanchez-Garcia, J., M. Fennelly, S. Norris et al. (2014). « Hierarchical Spectral Clustering of Power Grids ». In : *IEEE Transactions on Power Systems* 29.5, p. 2229-2237. ISSN : 0885-8950. DOI : 10.1109/TPWRS.2014.2306756.

Sanguinetti, G., J. Laidler et N. Lawrence (sept. 2005). « Automatic Determination of the Number of Clusters Using Spectral Algorithms ». In : *2005 IEEE Workshop on Machine Learning for Signal Processing*. ISSN : 1551-2541, 2378-928X, p. 55-60. DOI : 10.1109/MLSP.2005.1532874.

Sanogo, S. et al. (déc. 2015). « Spatio-temporal characteristics of the recent rainfall recovery in West Africa : RECENT RAINFALL RECOVERY IN WEST AFRICA ». en. In : *International Journal of Climatology* 35.15, p. 4589-4605. ISSN : 08998418. DOI : 10.1002/joc.4309.

Savoye, N. (2001). « Origine et transfert de la matière organique particulaire dans les écosystèmes littoraux macrotidaux ». 2001BRES2034. Thèse de doct., 324 p.

Sazhin, A. F., L. F. Artigas, J. C. Nejstgaard et M. E. Frischer (mai 2007). « The colonization of two Phaeocystis species (Prymnesiophyceae) by pennate diatoms and other protists : a significant contribution to colony biomass ». en. In : *Biogeochemistry* 83.1-3, p. 137-145. ISSN : 0168-2563, 1573-515X. DOI : 10.1007/s10533-007-9086-2.

Schapira, M., D. Vincent, V. Gentilhomme et L. Seuront (août 2008). « Temporal patterns of phytoplankton assemblages, size spectra and diversity during the wane of a *Phaeocystis globosa* spring bloom in hydrologically contrasted coastal waters ». en. In : *Journal of the Marine Biological Association of the United Kingdom* 88.4, p. 649-662. ISSN : 0025-3154, 1469-7769. DOI : 10.1017/S0025315408001306.

Schmitt, F. et A. Lefebvre (2016). *Mesures à haute résolution dans l'environnement marin côtier - CNRS Editions*. French. OCLC : 958293347. ISBN : 9782271085924.

Schoemann, V. et al. (jan. 2005). « Phaeocystis blooms in the global ocean and their controlling mechanisms : a review ». en. In : *Journal of Sea Research* 53.1-2, p. 43-66. ISSN : 13851101. DOI : 10.1016/j.seares.2004.01.008.

Schroeder, F., B. Mizerkowski et W. Petersen (jan. 2008). « The pocketFerryBox – A new portable device for water quality monitoring in oceans and rivers ». en. In : *Journal of Operational Oceanography* 1.2, p. 51-57. ISSN : 1755-876X, 1755-8778. DOI : 10.1080/1755876X.2008.11020103.

Sempéré, R., B. Charrière, F. Van Wambeke et G. Cauwet (juin 2000). « Carbon inputs of the Rhône River to the Mediterranean Sea : Biogeochemical implications ». en. In : *Global Biogeochemical Cycles* 14.2, p. 669-681. ISSN : 08866236. DOI : 10.1029/1999GB900069.

Shannon, C. E. (juil. 1948). « A mathematical theory of communication ». In : *The Bell System Technical Journal* 27.3, p. 379-423. ISSN : 0005-8580. DOI : 10.1002/j.1538-7305.1948.tb01338.x.

Shi, J. et J. Malik (août 2000). « Normalized cuts and image segmentation ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, p. 888-905. ISSN : 0162-8828. DOI : 10.1109/34.868688.

Shindler, M., A. Wong et A. Meyerson (2011). « Fast and Accurate K-means for Large Datasets ». In : *Proceedings of the 24th International Conference on Neural Information Processing Systems*. NIPS'11. USA : Curran Associates Inc., p. 2375-2383. ISBN : 9781618395993.

Simon, N., A.-L. Cras, E. Foulon et R. Lemée (fév. 2009). « Diversity and evolution of marine phytoplankton ». en. In : *Comptes Rendus Biologies* 332.2-3, p. 159-170. ISSN : 16310691. DOI : 10.1016/j.crvi.2008.09.009.

Smayada, T. J. (1990a). « Novel and nuisance phytoplankton blooms in the sea : evidence for a global epidemic. » In : *Proceedings of the 4th International Conference on Toxic Marine Phytoplankton.* Elsevier, p. 29-40.

— (1990b). « Novel and nuisance phytoplankton blooms in the sea : evidence for a global epidemic. » In : p. 29-40.

Smayda, T. J. et C. S. Reynolds (mai 2001). « Community Assembly in Marine Phytoplankton : Application of Recent Models to Harmful Dinoflagellate Blooms ». en. In : *Journal of Plankton Research* 23.5, p. 447-461. ISSN : 0142-7873. DOI : 10.1093/plankt/23.5.447.

Sournia, A., J. Brylinsk et S. Dallot (1990). « Fronts hydrologiques au large des côtes françaises : Les sites-ateliers de programme Frontal ». In : *Oceanologica acta* 13(4), p. 413-438.

Stern, R. D., M. D. Dennett et D. J. Garbutt (jan. 1981). « The start of the rains in West Africa ». en. In : *Journal of Climatology* 1.1, p. 59-68. ISSN : 01961748. DOI : 10.1002/joc.3370010107.

Stiger-Pouvreau, V. et G. Thouzeau (2015). « Marine Species Introduced on the French Channel-Atlantic Coasts : A Review of Main Biological Invasions and Impacts ». In : *Open Journal of Ecology* 05.05, p. 227-257. ISSN : 2162-1985, 2162-1993. DOI : 10.4236/oje.2015.55019.

Sverdrup, H. U. (jan. 1953). « On Conditions for the Vernal Blooming of Phytoplankton ». en. In : *ICES Journal of Marine Science* 18.3, p. 287-295. ISSN : 1054-3139, 1095-9289. DOI : 10.1093/icesjms/18.3.287.

Tett, P. et E. Barton (1995). « Why are there about 5000 species of phytoplankton in the sea ? » en. In : *Journal of Plankton Research* 17.8, p. 1693-1704. ISSN : 0142-7873, 1464-3774. DOI : 10.1093/plankt/17.8.1693.

Tew-Kai, E., V. Quilfen, M. Cachera et M. Boutet (août 2020). « Dynamic Coastal-Shelf Seascapes to Support Marine Policies Using Operational Coastal Oceanography : The French Example ». In : *JMSE* 8.8, p. 585. DOI : 10.3390/jmse8080585.

Toussaint, F. et al. (nov. 2014). « A new device to follow temporal variations of oxygen demand in deltaic sediments : the LSCE benthic station ». en. In : *Limnology and Oceanography : Methods* 12.11, p. 729-741. ISSN : 15415856. DOI : 10.4319/lom.2014.12.729.

Troadec, P. et R. Le Goff (1997). *Etat des lieux et des milieux de la rade de Brest et de son bassin versant.*
Phase préliminaire du Contrat de Baie de la rade de Brest. Brest : Communauté Urbaine de Brest.

Tsinaslanidis, P. et D. Kugiumtzis (2014). « A prediction scheme using perceptually important points and dynamic time warping ». In : *Expert Systems with Applications* 41.15, p. 6848-6860. issn : 0957-4174. doi : https://doi.org/10.1016/j.eswa.2014.04.028.

Utermöhl, H. (jan. 1958). « Methods of collecting plankton for various purposes are discussed. » In : *SIL Communications, 1953-1996* 9.1, p. 1-38. issn : 0538-4680. doi : 10.1080/05384680.1958.11904091.

Van Hoan, M., D. Huy et M. L.C. (2017). « Pattern Discovery in the Financial Time Series Based on Local Trend. » In : *Advances in Information and Communication Technology. ICTA 2016. Springer, Cham. doi 10.1007/978-3-319-49073-1_.48* 538.

Vantrepotte, V. et al. (mar. 2007). « Bio-optical properties of coastal waters in the Eastern English Channel ». en. In : *Estuarine, Coastal and Shelf Science* 72.1-2, p. 201-212. issn : 02727714. doi : 10.1016/j.ecss.2006.10.016.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc. isbn : 0-387-94559-8.

— (1998). *Statistical Learning Theory*. Wiley-Interscience.

— (juin 2013). *The Nature of Statistical Learning Theory*. en. Google-Books-ID : EqgACAAAQBAJ. Springer Science & Business Media. isbn : 9781475732641.

Vavrek, M. J. (2011). « fossil : palaeoecological and palaeogeographical analysis tools ». In : *Palaeontologia Electronica* 14.1. R package version 0.3.0, 1T.

Vaz, S., A. Carpentier et F. Coppin (mar. 2007). « Eastern English Channel fish assemblages : measuring the structuring effect of habitats on distinct sub-communities ». en. In : *ICES Journal of Marine Science* 64.2, p. 271-287. issn : 1054-3139. doi : 10.1093/icesjms/fsl031.

Veldhuis, M. J. W. et P. Wassmann (août 2005). « Bloom dynamics and biological control of a high biomass HAB species in European coastal waters : A Phaeocystis case study ». en. In : *Harmful Algae* 4.5, p. 805-809. issn : 15689883. doi : 10.1016/j.hal.2004.12.004.

Verity, P. et V. Smetacek (1996). « Organism life cycles, predation, and the structure of marine pelagic ecosystems ». en. In : *Marine Ecology Progress Series* 130, p. 277-293. issn : 0171-8630, 1616-1599. doi : 10.3354/meps130277.

Verma, D. et M. Meila (2003). *A Comparison of Spectral Clustering Algorithms*. Rapp. tech.

Wang, B. et LinHo (fév. 2002). « Rainy Season of the Asian–Pacific Summer Monsoon* ». In : *Journal of Climate* 15.4, p. 386-398. issn : 0894-8755. doi : 10.1175/1520-0442(2002)015<0386:RSOTAP>2.0.CO;2.

Ward, J. H. (mar. 1963). « Hierarchical Grouping to Optimize an Objective Function ». en. In : *Journal of the American Statistical Association* 58.301, p. 236-244. issn : 0162-1459, 1537-274X. doi : 10.1080/01621459.1963.10500845.

Wefer, G. (2015). « Ocean Margin Systems ». en. In : *Encyclopedia of Marine Geosciences*. Sous la dir. de J. Harff, M. Meschede, S. Petersen et J. Thiede. Dordrecht : Springer Netherlands, p. 1-9. isbn : 9789400766440. doi : 10.1007/978-94-007-6644-0_193-1.

Weisse, T. et al. (avr. 1994). « The trophic significance of Phaeocystis blooms ». en. In : *Journal of Marine Systems* 5.1, p. 67-79. issn : 09247963. doi : 10.1016/0924-7963(94)90017-5.

Wijnhoven, S. (2014). *Definition Strategy and Interfaces with the Monitoring of Biodiversity*. Report (Contract report).

Winder, M. et J. E. Cloern (oct. 2010). « The annual cycles of phytoplankton biomass ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 365.1555, p. 3215-3226. issn : 0962-8436, 1471-2970. doi : 10.1098/rstb.2010.0125.

WoRMS (2020). *WoRMS - World Register of Marine Species - IMIS*.

Xiang, T. et S. Gong (mar. 2008). « Spectral clustering with eigenvector selection ». en. In : *Pattern Recognition*. Part Special issue : Feature Generation and Machine Learning for Robust

Multimodal Biometrics 41.3, p. 1012-1029. ISSN : 0031-3203. DOI : 10.1016/j.patcog.2007.07.023.

Zelnik-Manor, L. et P. Perona (2004). « Self-Tuning Spectral Clustering ». In : *Advance in Neural Information Processing Systems 17* 2, p. 1601-1608.

Zhao, Q. (2012). *Cluster Validity in Clustering Methods. Publications of the University of Eastern Finland.*
Dissertations in Forestry and Natural Sciences., no 77. ISSN : 1798-5676.

Annexe A

Unités taxonomiques

TABLEAU A.1 – Correspondance entre les unités taxononimques et la liste des taxons identifiés en Manche lors des campagnes REPHY

Unités Taxonomique	Taxons
Achnanthes	*Achnanthes Actinocyclus*
Actinoptychus	*Actinoptychus*
Alexandrium	*Alexandrium, Alexandrium minutum*
Amphidinium	*Amphidinium, Amphidinium crassum*
Amphidomataceae	*Amphidomataceae*
Archaeperidinium minutum	*Archaeperidinium minutum*
Asterionellopsis	*Asterionella + Asterionellopsis + Asteroplanus, Asterionellopsis glacialis*
Azadinium	*Azadinium*
Bacillaria	*Bacillaria, Bacillaria paxillifera*
Bacillariaceae	*Bacillariaceae*
Bacillariophyceae	*Bacillariophyceae*
Bacteriastrum	*Bacteriastrum*
Bellerochea	*Bellerochea*
Brockmanniella	*Brockmanniella, Brockmanniella brockmannii*
Cerataulina	*Cerataulina, Cerataulina pelagica*
Ceratium	*Ceratium, Neoceratium furca, Trigonium alternans, Neoceratium furca,Tripos fusus Tripos lineatus*
Chaetoceros	*Chaetoceros, Attheya armata,Chaetoceros curvisetus, Chaetoceros curvisetus + debilis + pseudocurvisetus, Chaetoceros danicus, Chaetoceros decipiens, Chaetoceros diadema Chaetoceros densus + eibenii + borealis + castracanei Chaetoceros decipiens + lorenzianus, Chaetoceros densus, Chaetoceros, Chaetoceros didymus + protuberans didymus, Chaetoceros rostratus, Chaetoceros socialis, Chaetoceros socialis + socialis f. radians, Chaetoceros socialis f. radians*
Chlorophyceae	*Chlorophyceae*
Choanofila	*Choanofila*
Chrysochromulina	*Chrysochromulina*
Chrysophyceae	*Chrysophyceae*
Ciliophora	*Ciliophora*
Corethron	*Corethron*
Coscinodiscus	*Coscinodiscus, Coscinodiscus + Stellarima*
Cryptophyceae	*Cryptophyceae*
Cylindrotheca	*Cylindrotheca closterium, Nitzschia longissima*
Dactyliosolen fragilissimus	*Dactyliosolen fragilissimus*
Rhaphoneis - Delphineis	*Rhaphoneis, Delphineis,Rhaphoneis + Delphineis*
Dictyocha	*Dictyocha, Dictyocha fibula, Dictyocha speculum*
Dinoflagellata	*Dinoflagellata*
Dinophysis	*Dinophysis, Dinophysis acuminata, Dinophysis sacculus*
Diploneis	*Diploneis*
Diplopsalis	*Diplopsalis, Diplopsalis+Diplopelta+Diplopsalopsis+Preperidinium+Oblea, Diplopsalopsis*
Ditylum	*Ditylum, Ditylum brightwellii*
Eucampia - Climacodium	*Eucampia + Climacodium, Eucampia zodiacus*
Euglena	*Euglena, Euglenaceae*
Euglena	*Euglenaceae*
Eutreptiaceae	*Eutreptiaceae,Eutreptiella*
Fragilaria	*Fragilaria*
Gonyaulax	

Tableau A.2 – Continuation du tableau A.1

Unités Taxonomique	Taxons
Guinardia	*Guinardia,Guinardia delicatula, Guinardia flaccida,Guinardia striata*
Gymnodiniaceae	*Gymnodiniaceae*
Gymnodiniales	*Gymnodiniales, Gymnodinium*
Gyrodinium	*Gyrodinium, Gyrodinium spirale*
Helicotheca	
Heterocapsa	*Heterocapsa, Heterocapsa rotundata*
Heterosigma	*Heterosigma, Heterosigma akashiwo*
Hyalodiscaceae	*Hyalodiscaceae*
Karlodinium	*Karlodinium, Karlodinium veneficum, Katodinium*
Lauderia	*Lauderia + Detonula, Lauderia annulata*
Leptocylindrus	*Leptocylindrus, Leptocylindrus danicus, Leptocylindrus minimus, Leptocylindrus, complexe danicus groupe des larges (danicus + curvatus + mediterraneus + aporus + convexus + hargravesii + adriaticus)*
Licmophora	*Licmophora*
Melosira	*Melosira*
Mesodinium	*Mesodinium, Mesodinium rubrum*
Meuniera	*Meuniera, Meuniera membranacea*
Navicula	*Navicula, Navicula + Fallacia + Haslea + Lyrella + Petroneis, Navicula pelagica*
Nitzschia	*Nitzschia, Nitzschia + Hantzschia*
Noctiluca	*Noctiluca scintillans,Noctilucaceae, Noctilucales*
Odontella	*Odontella, Odontella aurita,Odontella granulata, Odontella sinensis Biddulphia, Biddulphia rhombus,Trieres mobiliensis,Trieres regia*
Paralia sulcata	*Paralia sulcata*
Pennees	*Pennées*
Peridiniales	*Peridiniales*
Phaeocystis	*Phaeocystis*
Phalacroma rotundatum	*Phalacroma rotundatum*
Plagiogramma	*Plagiogramma*
Plagiogrammopsis	*Plagiogrammopsis, Plagiogrammopsis vanheurckii*
Pleurosigma - Gyrosigma	*Pleurosigma, Gyrosigma Pleurosigma + Gyrosigma*
Podosira	*Podosira*
Polykrikos	*Polykrikos, Polykrikos schwarzii*
Porosira	*Porosira*
Prorocentrum	*Prorocentrum,Prorocentrum balticum + cordatum, Prorocentrum cordatum, Prorocentrum micans,Prorocentrum micans + arcuatum + gibbosum + scutellum, Prorocentrum triestinum, Protoceratium*
Protoctista	*Protoctista*
Protoperidinium	*Protoperidinium, Protoperidinium + Peridinium, Protoperidinium bipes, Protoperidinium depressum*
Pseudo-nitzsch autre	*Pseudo-nitzschia*
Pseudo-nitzsch autre	*Pseudo-nitzschia, complexe delicatissima, groupe des fines (calliantha + delicatissima + pseudodelicatissima + subcurvata), Pseudo-nitzschia, complexe seriata, groupe des larges (australis + fraudulenta + seriata + subpacifica), Pseudo-nitzschia americana, Pseudo-nitzschia delicatissima, Pseudo-nitzschia pungens, Pseudo-nitzschia seriata*
Pyramimonas	*Pyramimonas*
Pyrocystaceae	*Pyrocystaceae*
Pyrocystis	*Pyrocystis*
Pyrophacus	*Pyrophacus*
Rhizosolenia	*Rhizosolenia, Rhizosolenia hebetata, Rhizosolenia imbricata, Rhizosolenia imbricata + styliformis, Rhizosolenia robusta, Rhizosolenia setigera, Rhizosolenia setigera + setigera f. pungens, Rhizosolenia styliformis, Proboscia, Proboscia alata, Proboscia indica*
Scenedesmus	*Scenedesmus*
Scrippsiella	*Scrippsiella, Scrippsiella + Ensiculifera + Pentapharsodinium*

TABLEAU A.3 – Continuation du tableau A.1

Skeletonema costatum	*Skeletonema costatum*
Spatulodinium pseudonoctiluca	*Spatulodinium pseudonoctiluca*
Stauroneis	*Stauroneis*
Stephanopyxis	*Stephanopyxis*
Thalassionema nitzschioides	*Thalassionema nitzschioides*
Thalassiosira	*Thalassiosira, Thalassiosira + Porosira, Thalassiosira angulata, Thalassiosira antarctica, Thalassiosira gravida, Thalassiosira levanderi, Thalassiosira levanderi + minima, Thalassiosira nordenskioeldii*
Torodinium	*Torodinium*
Triceratium	*Triceratium*
Trigonium alternans	*Trigonium alternans*

Annexe B

Gammes "Capteur" et "Expert"

La gamme "capteur" est un intervalle de valeurs dites correctes et définies par le fabricant, la gamme "expert" est un intervalle de valeurs dites correctes et définies par un expert du domaine. Notons que la gamme "expert" étant définie comme plus précise que la gamme "capteur", celle-ci valide ou modifie par un intervalle plus strict le niveau de qualité de la gamme "capteur". Sur l'ensemble des paramètres, ne sont retenues que les valeurs comprises dans la gamme, alors que les valeurs hors gammes sont remplacées par des valeurs manquantes (*NA*)

Les gammes de MAREL-Carnot, MAREL-Iroise et MesuRho, représentées respectivement Tableau B.1, B.2 et B.3, ont été établies à partir partir des fiches techniques pour la gamme "capteur" et pour la gamme Expert à partir d'une étude statistique classique. Cette dernière a permis de dégager les propriétés de la série temporelle (minimum, maximum, moyenne, médiane, écart-type), les tendances et les valeurs particulières (extrême, ou erreur) des données brutes.

B.1 MAREL-Carnot

Tableau B.1 – Gamme "capteur" et "expert" pour le système de mesures. MAREL-Carnot

Noms			**Gamme Capteur**		**Gamme Expert**	
Paramètres	**Coriolis**	**Unités**	**min**	**max**	**min**	**max**
Température	TEMP.LEVEL1	°C	-5	35	0	30
Salinité	PSAL.LEVEL1	PSU	2	42	5	35
Oxygène dissous	//	$mg\,l^{-1}$	0	30	0	20
Oxygène dissous	DOX1.LEVEL1	$ml\,l^{-1}$	//	//	//	//
Saturation en Oxygène	OSAT.LEVEL1	%	0	150	0	150
Fluorescence	FLU3.LEVEL1	FFU	0	500	0,03	150
Turbidité	TUR4.LEVEL1	NTU	0	1500	0	150
PAR	LGH4.LEVEL1	$\mu mol\,m^{-2}\,s^{-1}$	//	//	0	2500
Direction du vent	WDIR.LEVEL0	°	0	360	0	360
Vitesse du vent	WSPD.LEVEL0	$m\,s^{-1}$	0	40	0	40
Nitrate	NTRZ.LEVEL1	$\mu mol\,l^{-1}$	0	100	0	100
Phosphate	PHOS.LEVEL1	$\mu mol\,l^{-1}$	0	100	0	10
Silicate	SLCA.LEVEL1	$\mu mol\,l^{-1}$	0	100	0	50
Niveau de la mer	SLEV.LEVEL1	m	//	//	0	20
Température de l'air	DRYT.LEVEL0	°C	-20	40	-20	40

B.2 MAREL-Iroise

Tableau B.2 – Gamme capteur et expert pour le système de mesures MAREL-Iroise

Noms			Gamme Capteur		Gamme Expert	
Paramètres	**Coriolis**	**Unités**	**min**	**max**	**min**	**max**
Température	TEMP.LEVEL1	°C	-5	35	0	30
Salinité	PSAL.LEVEL1	PSU	2	42	5	35
Oxygène Dissout	//	$mg\,l^{-1}$	0	30	0	20
Oxygène Dissout	DOX1.LEVEL1	$ml\,l^{-1}$	//	//	//	//
Saturation en Oxygène	OSAT.LEVEL1	%	0	120	0	150
Fluorescence	FLU3.LEVEL1	FFU	//	//	//	//
Fluorescence	FLU3_ADJUSTED	$\mu g\,l^{-1}$	0 $\mu g\,l^{-1}$	500 $\mu g\,l^{-1}$	0	50
Turbidité	TUR4.LEVEL1	NTU	0	1500	2	40
PAR	LGH4.LEVEL1	$\mu mol\,m^{-2}\,s^{-1}$	0	3000	0	2500
Température de l'air	DRYT.LEVEL0	°C	-20	40	-20	40

B.3 MesuRho

Tableau B.3 – Gamme capteur et expert pour le système de mesures MesuRho

Noms			Gamme Capteur		Gamme Expert	
Paramètres	**Coriolis**	**Unités**	**min**	**max**	**min**	**max**
Température	TEMP.LEVEL1	°C	-5	35	0	30
Salinité	PSAL.LEVEL1	PSU	2	42	5	35
Oxygène Dissout	//	$mg\,l^{-1}$	0	30	0	20
Oxygène Dissout	DOX2.LEVEL1	$ml\,l^{-1}$	//	//	//	//
Saturation en Oxygène	OSAT.LEVEL1	%	0	120	0	150
Fluorescence	FLU2.LEVEL1	$\mu g\,l^{-1}$	0	500	0	150
Turbidité	TUR4.LEVEL1	NTU	0	1500	0	150
PAR	LGH4.LEVEL1	V	0	5,104	0	1,5
PAR	//	$\mu mol\,m^{-2}\,s^{-1}$	//	//	0	3000
Direction du vent	WDIR.LEVEL0	°	0	360	0	360
Vitesse du vent	WSPD.LEVEL0	$m\,s^{-1}$	0	40	0	40
Niveau de la mer	SLEV.LEVEL1	m	//	//	//	//
Température de l'air	DRYT.LEVEL0	°C	-20	40	-20	40

Annexe C

Résultats de classification

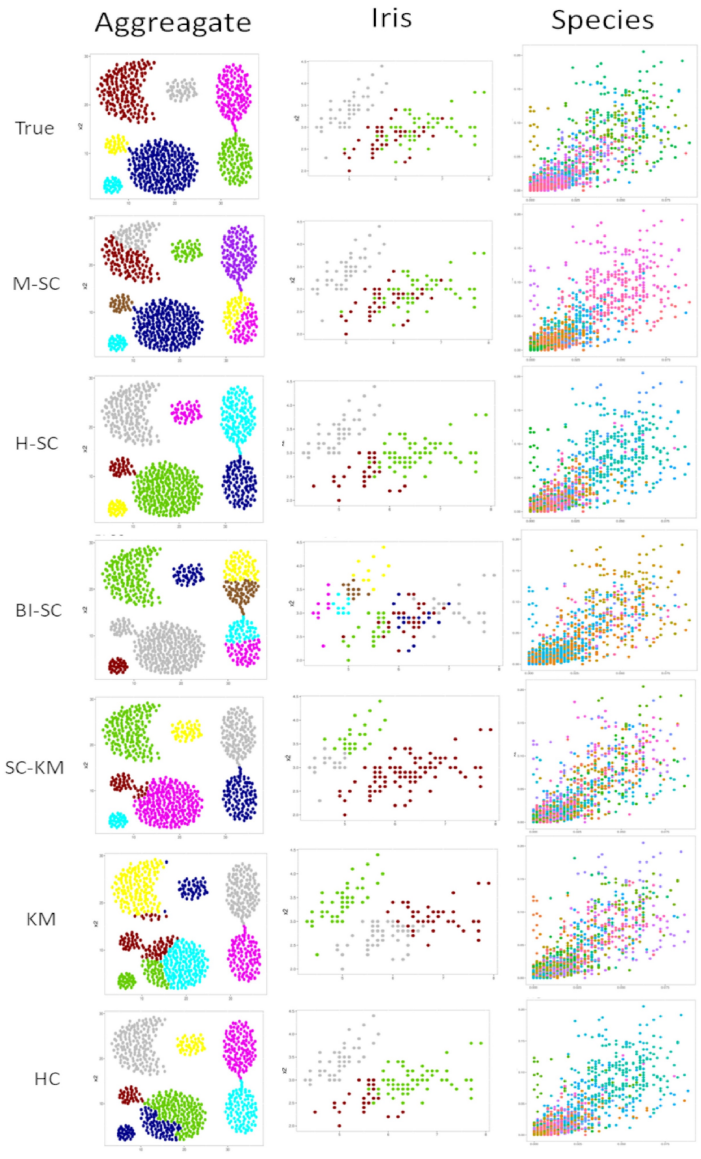


FIGURE C.1 – Résultats des classifications sur les jeux de données spatiales par les méthodes les plus efficaces. Les couleurs représentent les vrais labels C pour la ligne True et les classes K pour chaque méthode (M-SC, H-SC, Bi-SC, SC-KM, KM (*K-means*), HC).

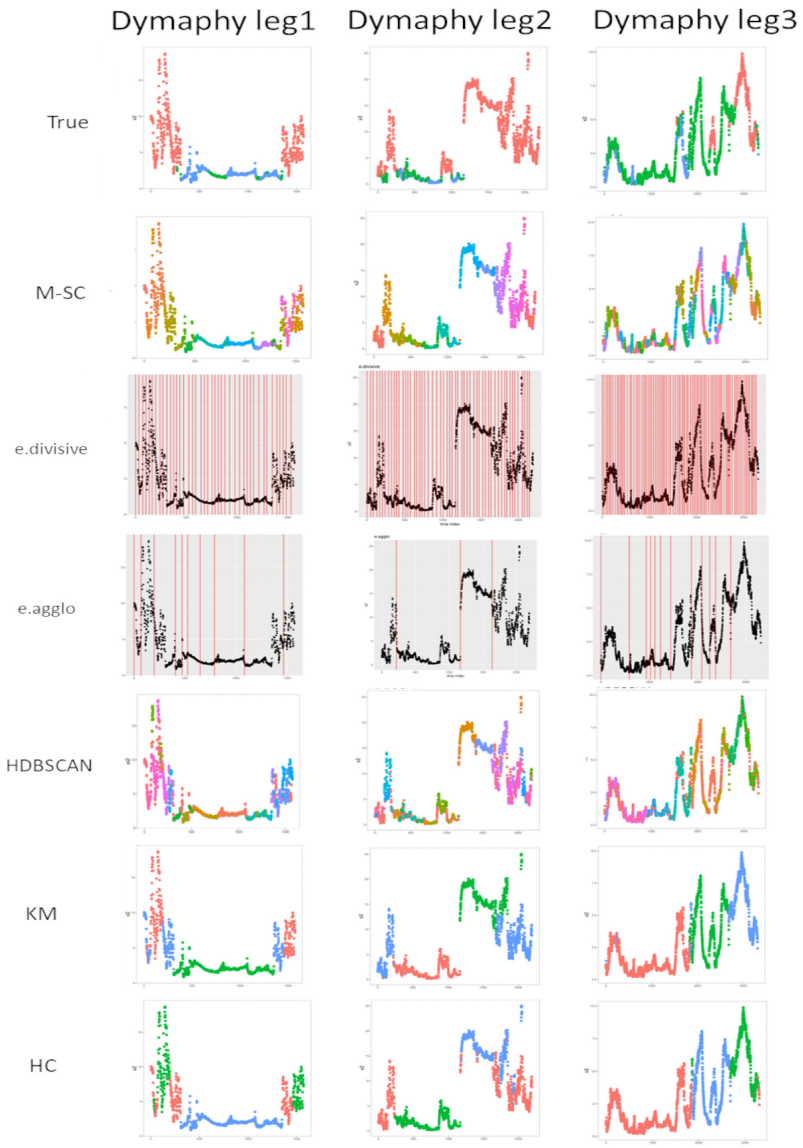


FIGURE C.2 – Résultats des classifications sur les séries temporelles par les méthodes les plus efficaces. Les couleurs représentent les vrais labels C pour la ligne True et les classes K pour chaque méthode (M-SC, e.divisive, e.agglo, HDBSCAN, KM (*K-means*), HC).

Annexe D

Étude de sensibilité : Fonction de complétion

D.1 Problématique

Lors du protocole de pré-traitement, les données sont complétées par *DTWBI* de Phan et al. 2017 modifiées pour être adaptées à nos problématiques(Section 3.2.2.3). Pour rappel, cette méthode recherche des correspondances dans le signal existant pour compléter les trous avec ce même signal. Cela permet de compléter une série de données dont le nombre de valeurs manquantes (*NA*) est important et d'obtenir une complétion qui préserve la dynamique de signaux complexes dans des séries avec une quantité conséquente de données manquantes consécutives.

La version initiale de T.T.H Phan (*the DTWBI* (R-CRAN:: DTWBI) est performante et permet de compléter 100 % du fichier MAREL-Carnot. Mais l'algorithme complète des séries de données manquantes de quelques jours à plus de 3 ans sans faire de distinctions. Or compte tenu des échelles de temps de la dynamique phytoplanctonique (Section 1.2.3, une complétion de plus de 3 ans de données est trop importante. C'est la raison pour laquelle, lors de cette thèse une version contrainte de la fonction *DTWBI* initiale est développée. Dans cette version deux paramètres (smallHole et acceptedHole) sont introduits. Ils permettent d'adapter la taille des fenêtres de complétion.

smallHole est le nombre de données manquantes (*NA*) consécutives que l'on considère comme petit, c'est-à-dire inférieur à la dynamique de changement du phénomène étudié (fenêtres à valeurs monotones).

acceptedHole est le nombre de données manquantes consécutives, que l'on considère comme trop grand en ce qui concerne le changement de la dynamique du processus, pour être complété.

Afin de définir au mieux ces deux paramètres une recherche d'optimale via une analyse de sensibilité est réalisée

D.2 Résultats

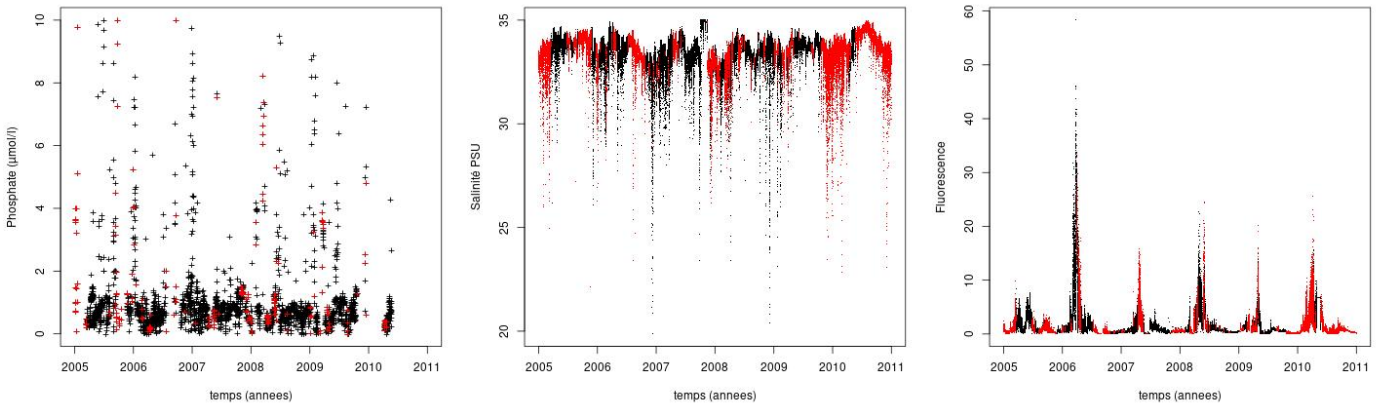
Six configurations de complétion (D.1) ont été testées sur un jeu de données MAREL-Carnot de 2005-2010 alignées et corrigées.

TABLEAU D.1 – Récapitulatif des tests de sensibilité pour les paramètres de l'algorithme de complétion .

tests	smallHole	acceptHole
1	1 jour	15 jours
2	1 jour	1 mois
3	1 jour	2 mois
4	1 jour	6 mois
5	1 semaine	1 mois
6	3 jours	1 mois

D.2.1 Test 1 : 15 jours

Avec seulement une fenêtre maximum de 15 jours complété nous avons 55.38 % de données utilisables. Et le nombre de NA par paramètre sont les suivants :



(a) Phosphate (b) Salinité (c) Fluorescences

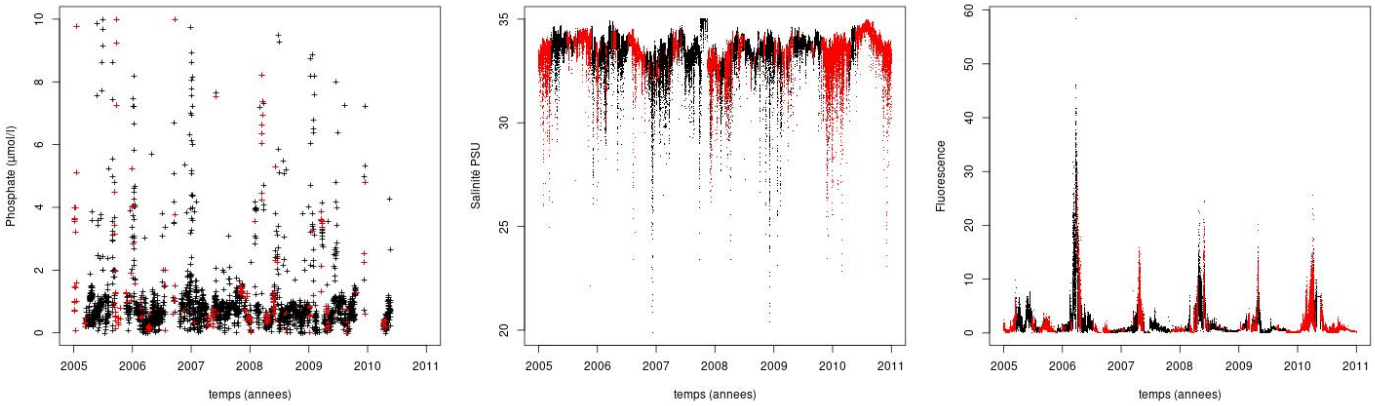
FIGURE D.1 – Série temporelle MAREL-Carnot. En rouge les données qui ne seront pas utilisées lors de la classification après la complétion paramétrée suivant le **test 1**.

TABLEAU D.2 – Statistiques du nombre de données manquantes pour chaque paramètre **pour le test 1**. Number.NA le nombre de données manquantes, Percentage.NA le pourcentage de données manquantes, Largest.Gap la taille de la plus longue séquence de données manquantes, Number.Gap nombre de séquences de données manquantes.

Paramètres	Number.NA	Percentage.NA	Largest.Gap	Number.Gap
Température (°C)	894	0,56	484	6
Salinité (PSU)	1043	0,66	484	5
Turbidité (NTU)	127	0,08	78	3
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	8510	5,39	7896	3
SeaLevel (m)	13628	8,64	8500	4
N ($\mu mol\,l^{-1}$)	54719	34,68	15837	62
P ($\mu mol\,l^{-1}$)	59277	37,57	15870	65
Si ($\mu mol\,l^{-1}$)	54605	34,61	15831	40
OxyDissolved ($mg\,l^{-1}$)	212	0,13	78	4

D.2.2 Test 2

Avec une séquence maximum de 1 mois nous augmentons les données utilisables de 55.38 % à 55.43 % (ce qui n'est vraiment pas énorme). Et le nombre de NA par paramètre sont les suivants :



(a) Phosphate (b) Salinité (c) Fluorescences

FIGURE D.2 – Série temporelle MAREL-Carnot. En rouge les données qui ne seront pas utilisées lors de la classification après la complétion paramétrée suivant le **test 2**

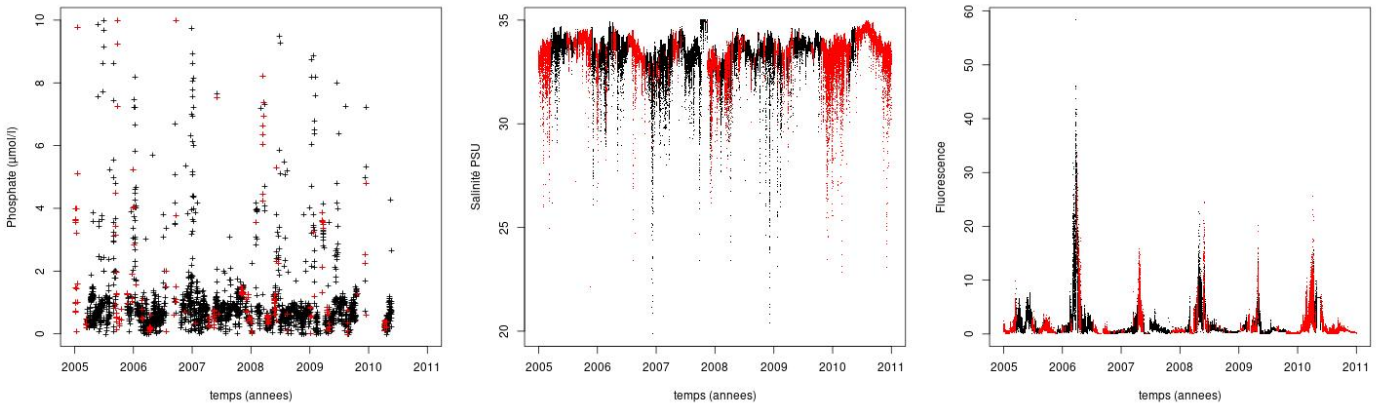
TABLEAU D.3 – Statistiques du nombre de données manquantes pour chaque paramètre **pour le test 2**. Number.NA le nombre de données manquantes, Percentage.NA le pourcentage de données manquantes, Largest.Gap la taille de la plus longue séquence de données manquantes, Number.Gap nombre de séquences de données manquantes.

Paramètres	Number.NA	Percentage.NA	Largest.Gap	Number.Gap
Température (°C)	894	0,57	484	6
Salinité (PSU)	1043	0,66	484	5
Turbidité (NTU)	127	0,08	78	3
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	8510	5,39	7896	3
SeaLevel (m) 1	2100	7,67	8500	3
N ($\mu mol\,l^{-1}$)	53021	33,61	15837	65
P ($\mu mol\,l^{-1}$)	56960	36,11	15870	70
Si ($\mu mol\,l^{-1}$)	51629	32,73	15831	47
OxyDissolved ($mg\,l^{-1}$)	212	0,13	78	4

D.2.3 Test 3 et 4

Cette fois, la taille de la séquence maximum est fixée respectivement à 2mois et 6mois, pour le test 3 et 4. Dans ce cas, le nombre de lignes utilisables et les statistiques de NA sont identique au Test2. DTWBI ne trouve plus de correspondance.

Pour le moment, l'étape 2 (complétion par moyenne mobile) est réalisée pour des séquences inférieures à 1 jour ($smallHole = 1$ jour). Dans son protocole initial de T.T.H Phan utilise une complétion par moyenne mobile sur une fenêtre de 7 jours. Nous avons donc fixé le smallHole à 7jours et le AcceptHole à 1 mois (Test 5).

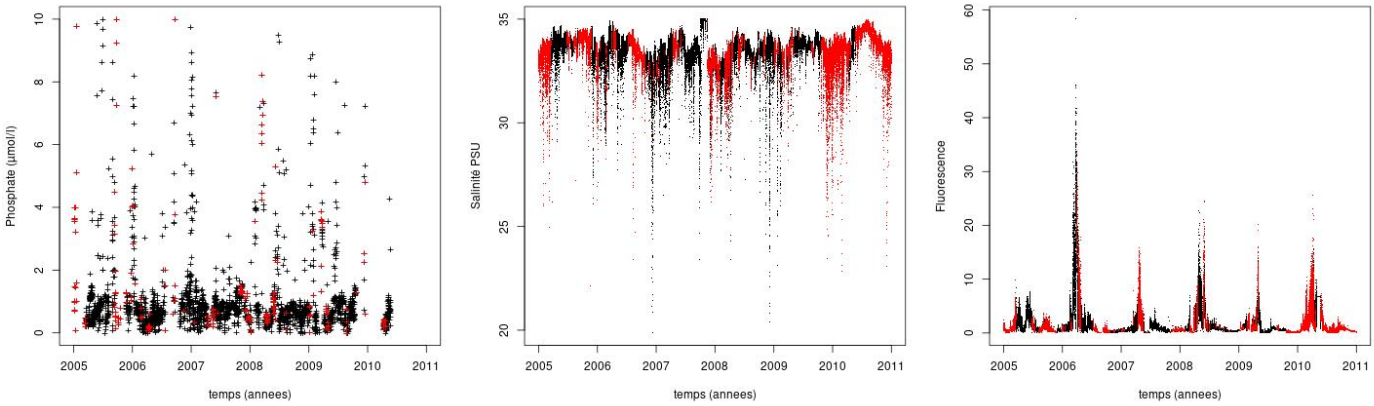


(a) Phosphate (b) Salinité (c) Fluorescences

FIGURE D.3 – Série temporelle MAREL-Carnot. En rouge les données qui ne seront pas utilisées lors de la classification après la complétion paramétrée suivant le **test 3**.

Tableau D.4 – Statistiques du nombre de données manquantes pour chaque paramètre **pour le test 3**. Number.NA le nombre de données manquantes, Percentage.NA le pourcentage de données manquantes, Largest.Gap la taille de la plus longue séquence de données manquantes, Number.Gap nombre de séquences de données manquantes.

Paramètres	Number.NA	Percentage.NA	Largest.Gap	Number.Gap
Température (°C)	894	0,57	484	6
Salinité (PSU)	1043	0,66	484	5
Turbidité (NTU)	127	0,08	78	3
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	8510	5,39	7896	3
SeaLevel (m)	12100	7,67	8500	3
N ($\mu mol\,l^{-1}$)	53021	33,61	15837	65
P ($\mu mol\,l^{-1}$)	56960	36,11	15870	70
Si ($\mu mol\,l^{-1}$)	51629	32,73	15831	47
OxyDissolved ($mg\,l^{-1}$)	212	0,13	78	4



(a) Phosphate (b) Salinité (c) Fluorescences

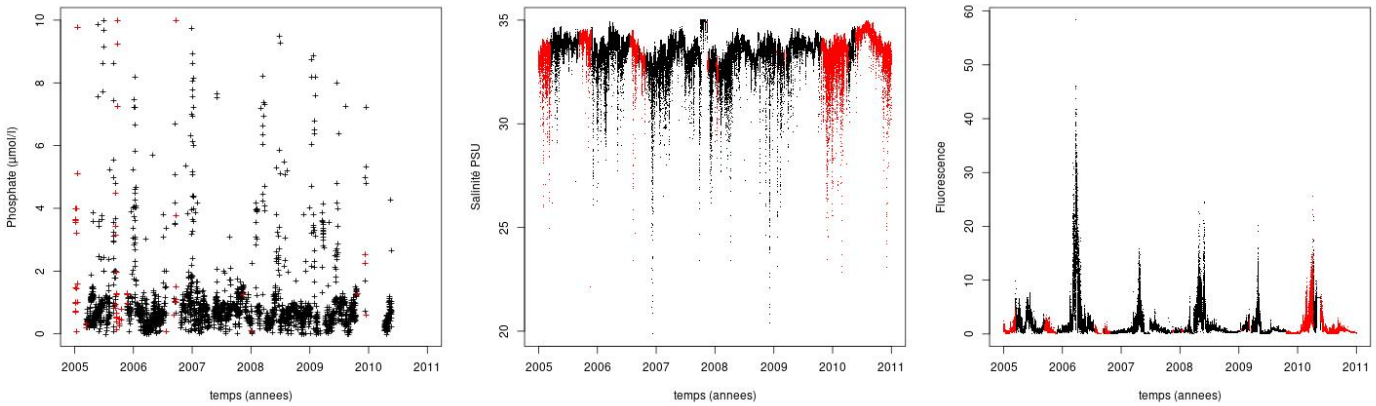
Figure D.4 – Série temporelle MAREL-Carnot. En rouge les données qui ne seront pas utilisées lors de la classification après la complétion paramétrée suivant le **test 4**.

TABLEAU D.5 – Statistiques du nombre de données manquantes pour chaque paramètre **pour le test 4**. Number.NA le nombre de données manquantes, Percentage.NA le pourcentage de données manquantes, Largest.Gap la taille de la plus longue séquence de données manquantes, Number.Gap nombre de séquences de données manquantes.

Paramètres	Number.NA	Percentage.NA	Largest.Gap	Number.Gap
Température (°C)	894	0,57	484	6
Salinité (PSU)	1043	0,66	484	5
Turbidité (NTU)	127	0,08	78	3
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	8510	5,39	7896	3
SeaLevel (m)	12100	7,67	8500	3
N ($\mu mol\,l^{-1}$)	53021	33,61	15837	65
P ($\mu mol\,l^{-1}$)	56960	36,11	15870	70
Si ($\mu mol\,l^{-1}$)	51629	32,73	15831	47
OxyDissolved ($mg\,l^{-1}$)	212	0,13	78	4

D.2.4 Test 5

Dans ce cas le nombre de lignes utilisables passe à 70,57 % et le nombre de NA par paramètre a largement diminué.



(a) Phosphate (b) Salinité (c) Fluorescences

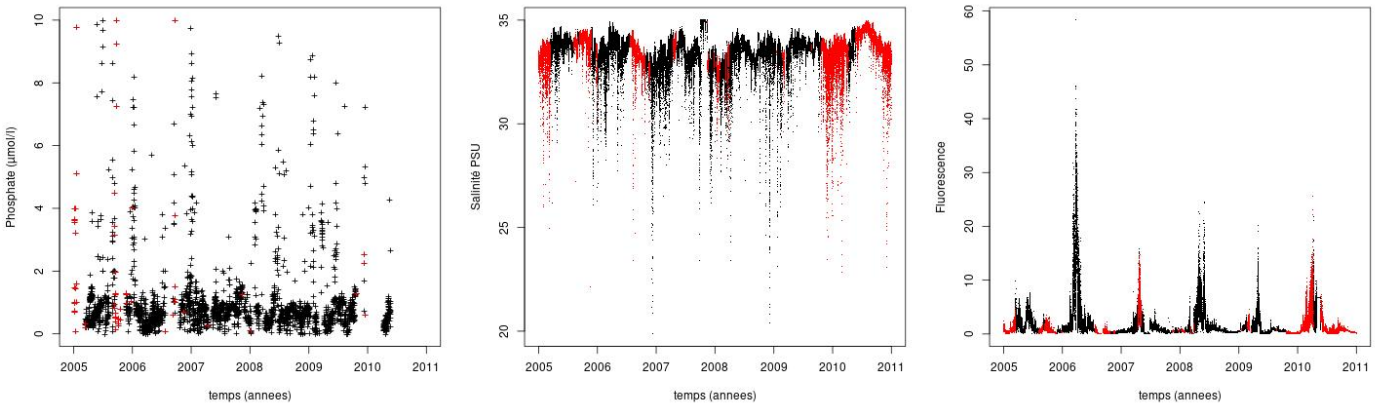
FIGURE D.5 – Série temporelle MAREL-Carnot. En rouge les données qui ne seront pas utilisées lors de la classification après la complétion paramétrée suivant le **test 5**.

Tableau D.6 – Statistiques du nombre de données manquantes pour chaque paramètre **pour le test 5**. Number.NA le nombre de données manquantes, Percentage.NA le pourcentage de données manquantes, Largest.Gap la taille de la plus longue séquence de données manquantes, Number.Gap nombre de séquences de données manquantes.

Paramètres	Number.NA	Percentage.NA	Largest.Gap	Number.Gap
Température (°C)	0	0,00	0	0
Salinité (PSU)	0	0,00	0	0
Turbidité (NTU)	0	0,00	0	0
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	8489	5,38	7896	2
SeaLevel (m)	11839	7,50	8500	2
N ($\mu mol\,l^{-1}$)	37567	23,81	15837	9
P ($\mu mol\,l^{-1}$)	40958	25,96	15870	8
Si ($\mu mol\,l^{-1}$)	42927	27,21	15831	8
OxyDissolved ($mg\,l^{-1}$)	0	0,00	0	0

D.2.5 Test 6

Dans ce cas le nombre de ligne utilisable passe a 66,7 %



(a) Phosphate (b) Salinité (c) Fluorescences

Figure D.6 – Série temporelle MAREL-Carnot. En rouge les données qui ne seront pas utilisées lors de la classification après la complétion paramétrée suivant le **test 6**.

TABLEAU D.7 – Statistiques du nombre de données manquantes pour chaque paramètre **pour le test 6**. Number.NA le nombre de données manquantes, Percentage.NA le pourcentage de données manquantes, Largest.Gap la taille de la plus longue séquence de données manquantes, Number.Gap nombre de séquences de données manquantes.

Paramètres	Number.NA	Percentage.NA	Largest.Gap	Number.Gap
Température (°C)	297	0,19	297	1
Salinité (PSU)	484	0,31	484	1
Turbidité (NTU)	39	0,02	39	1
PAR ($\mu mol\,m^{-2}\,s^{-1}$)	8489	5,38	7896	2
SeaLevel (m)	12100	7,67	8500	3
N ($\mu mol\,l^{-1}$)	39589	25,10	15837	13
P ($\mu mol\,l^{-1}$)	46297	29,35	15870	19
Si ($\mu mol\,l^{-1}$)	44176	28,00	15831	14
OxyDissolved ($mg\,l^{-1}$)	254	0,16	254	1

Annexe E

Résultats de l'apprentissage

E.1 Prédictions sur MAREL-Iroise de 2014 à 2016

E.1.1 Structuration des classes

Tableau E.1 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 1** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MAREL-Iroise 2014-2016 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Labels prédits	l1	l2	RI	ARI
cl1	3081	**6527**	0,60	0,20
cl2	**2748**	414		

Tableau E.2 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 2** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeux de données MAREL-Iroise 2014-2016 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Labels prédits	l1	l2	l3	l4	RI	ARI
cl1	523	1579	253	**4239**		
cl2	74	852	**1272**	815	0,60	0,20
cl3	344	**2360**	311	148		

TABLEAU E.3 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 3** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeux de données MAREL-Iroise 2014-2016 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Labels prédits	l1	l2	l3	l4	l5	l6	l7	l8	RI	ARI
cl1	2	10	435	1	94	39	2909	14		
cl2	233	269	1150	1	75	45	**1305**	11		
cl3	0	74	857	0	277	**987**	817	1	0,67	0.17
cl4	69	174	**1445**	8	211	22	75	0		
cl5	24	76	**904**	4	79	0	73	0		

E.1.2 Répartition temporelle des labels prédits

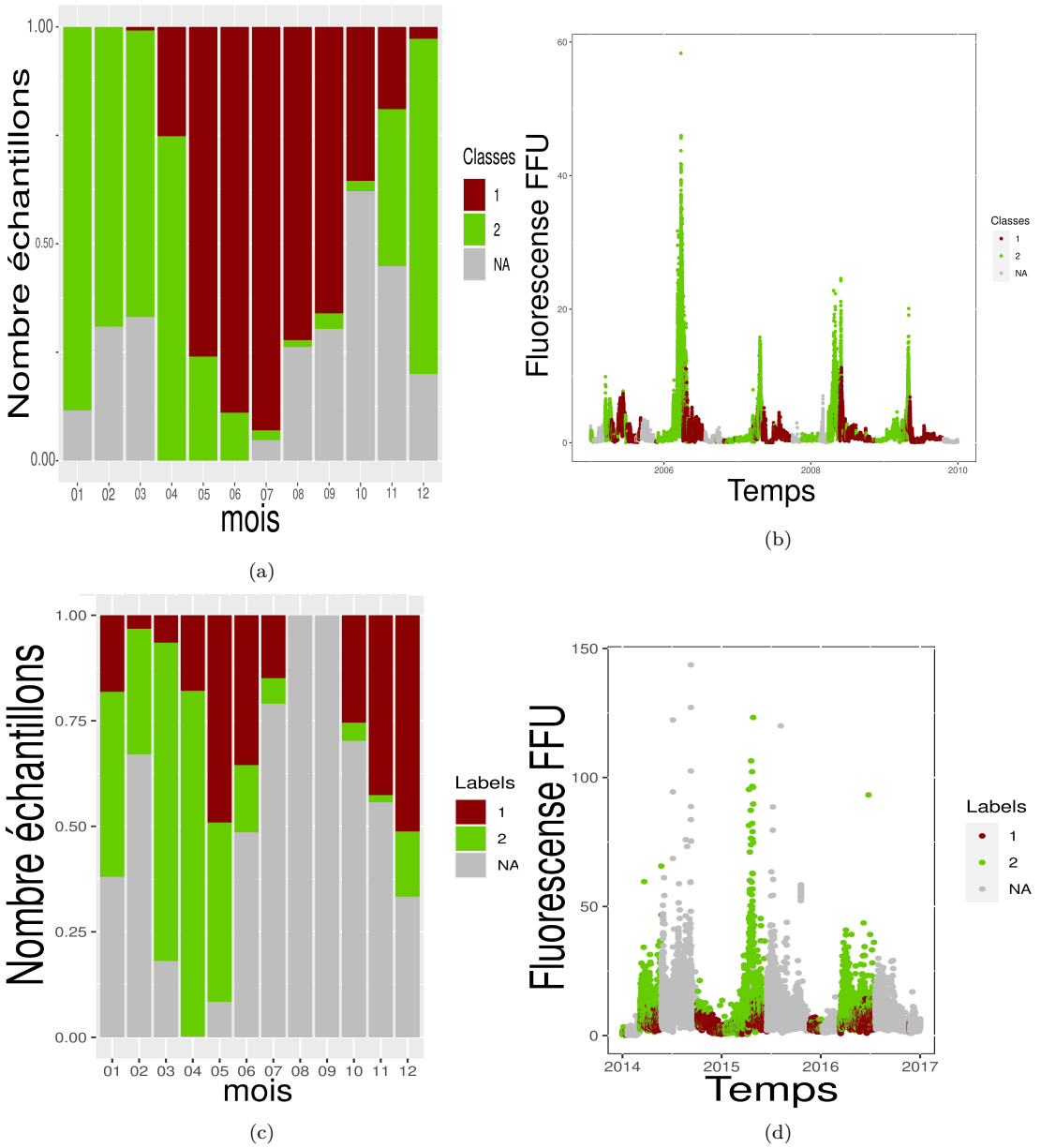
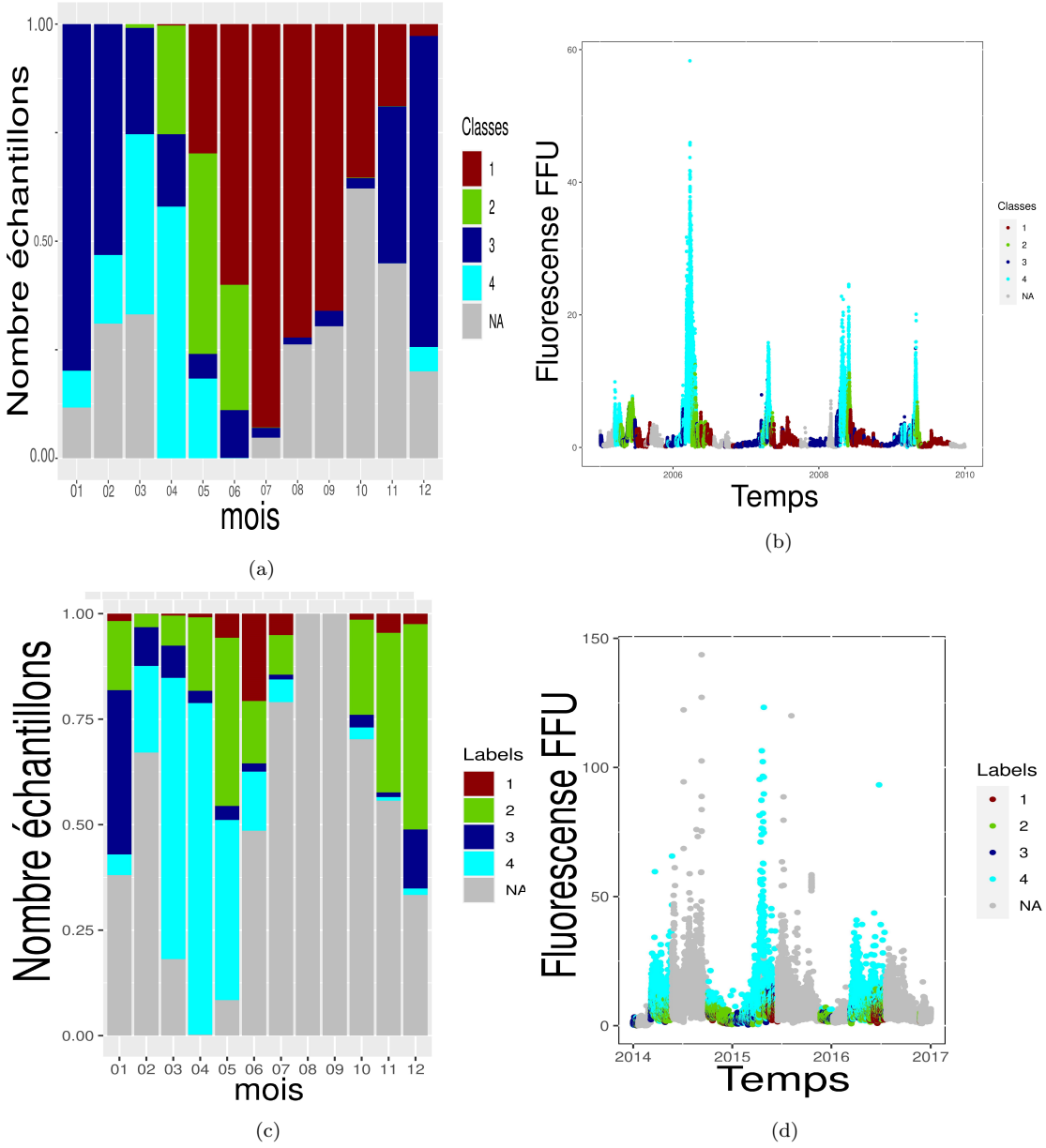


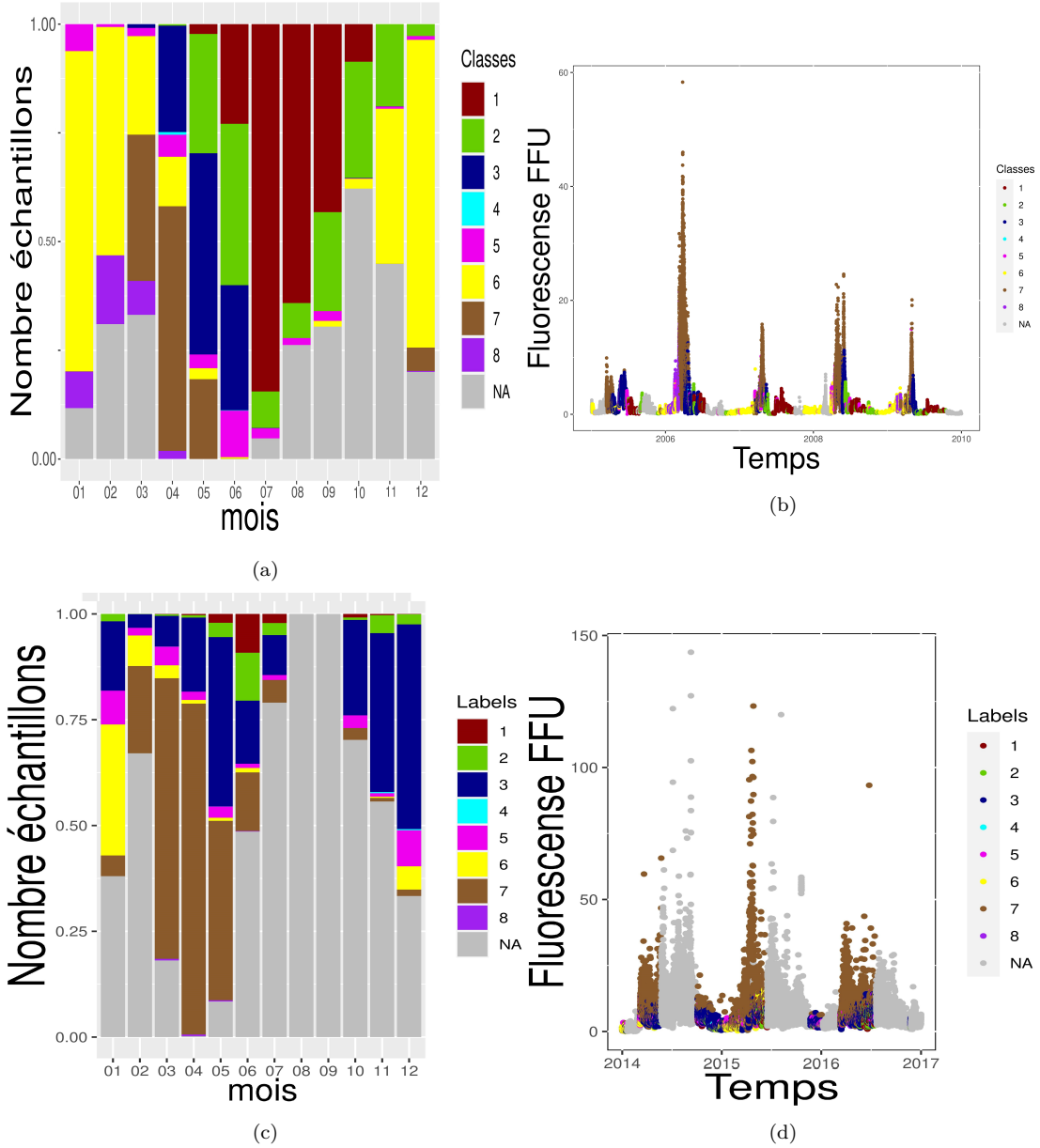
FIGURE E.1 – Comparaison au niveau 1 de la dynamique temporelle entre les classes labellisées par *M-SC* à MAREL-Carnot de 2005 à 2009 et les labels prédits pour MAREL-Iroise sur la période 2010-2014. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise et de la répartition sur le signale de fluorescence b) des classes issues de M-SC pour MAREL-Carnot sur 2005-2009, d) labels prédits pour MAREL-Iroise sur 2010-2014.



(a) (b)

(c) (d)

FIGURE E.2 – Comparaison au niveau 2 de la dynamique temporelle entre les classes labellisées par *M-SC* à MAREL-Carnot de 2005 à 2009 et les labels prédits pour MAREL-Iroise sur la période 2010-2014. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise et de la répartition sur le signale de fluorescence b) des classes issues de M-SC pour MAREL-Carnot sur 2005-2009, d) labels prédits pour MAREL-Iroise sur 2010-2014.



(a) (b)

(c) (d)

FIGURE E.3 – Comparaison au niveau 3 de la dynamique temporelle entre les classes labellisées par *M-SC* à MAREL-Carnot de 2005 à 2009 et les labels prédits pour MAREL-Iroise sur la période 2010-2014. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MAREL-Iroise et de la répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot sur 2005-2009, d) labels prédits pour MAREL-Iroise sur 2010-2014.

E.1.3 Prédictions sur MesuRho pour 2009

E.1.3.1 Structuration des classes

TABLEAU E.4 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 1** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MesuRho 2009 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	RI	ARI
cl1	**6371**	339	0,90	0,11
cl2	15	**27**		

TABLEAU E.5 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) **au niveau 2** (Classes non exp. cl_i) et les labels prédits (l_i) obtenus par Kppv à 1 voisin sur le jeu de données MesuRho 2009 et les indices RI et ARI. En gras, le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	l3	l4	RI	ARI
cl1	**4865**	8	213	7		
cl2	**751**	**747**	107	12	0,75	0,44
cl3	12	3	**27**	0		

TABLEAU E.6 – Tableau de contingence entre les classes non expertisées définies par classification spectrale (M-SC) au niveau 3 (Classes non exp.) et les labels prédits obtenus par Kppv à 1 voisin sur le jeu de données MesuRho 2009 et les indices RI et ARI. En gras le nombre d'occurrences majoritaires.

Classes non exp. vs Label prédits	l1	l2	l3	l4	l5	l6	l7	l8	RI	ARI
cl1	**3965**	899	8	0	91	123	7	0		
cl2	94	347	**430**	0	23	70	12	0	0,71	0,44
cl3	15	294	**317**	0	4	5	0	0		
cl4	11	3	4	0	0	**30**	0	0		

E.1.4 Répartition temporel des labels prédits

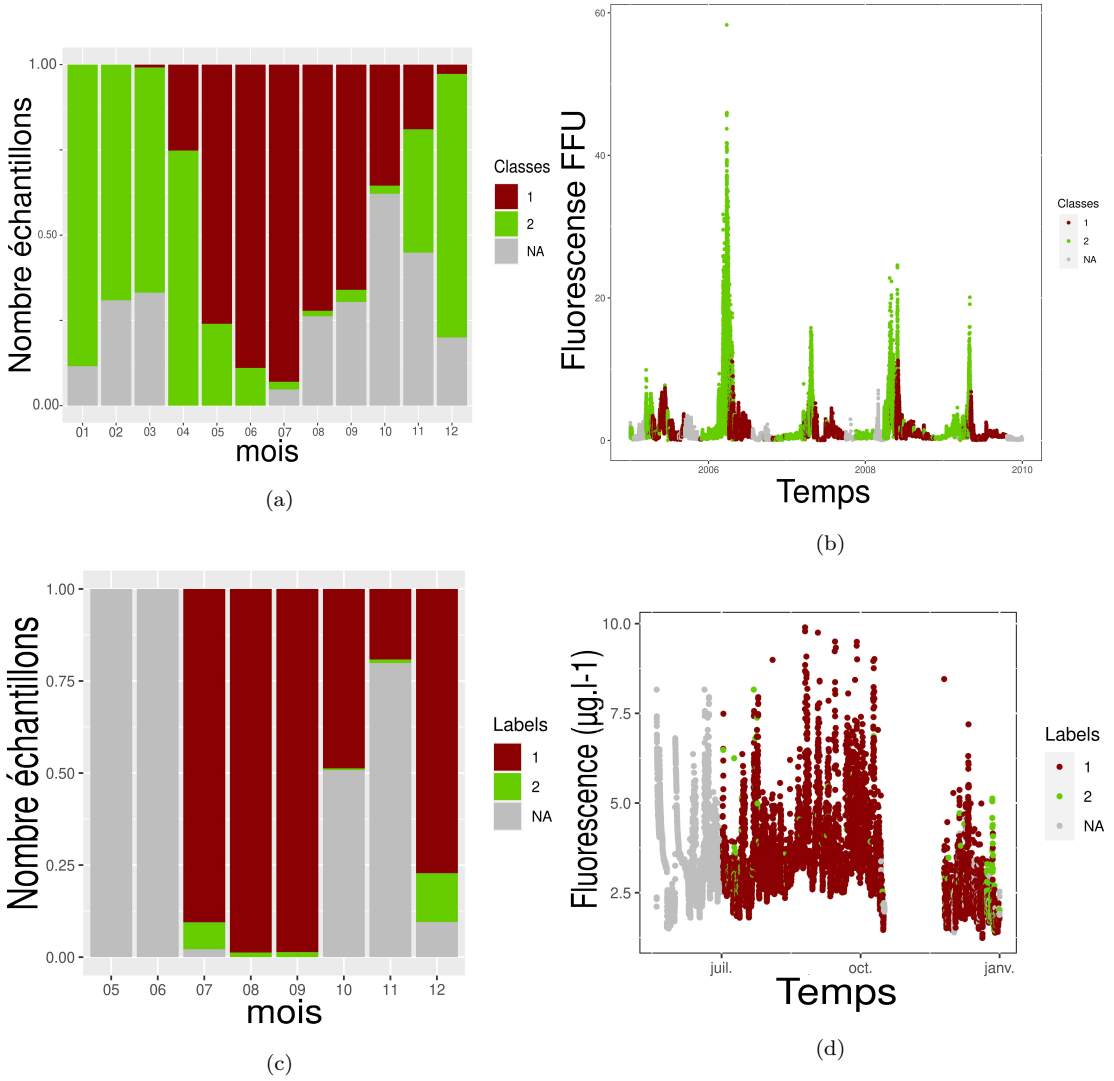
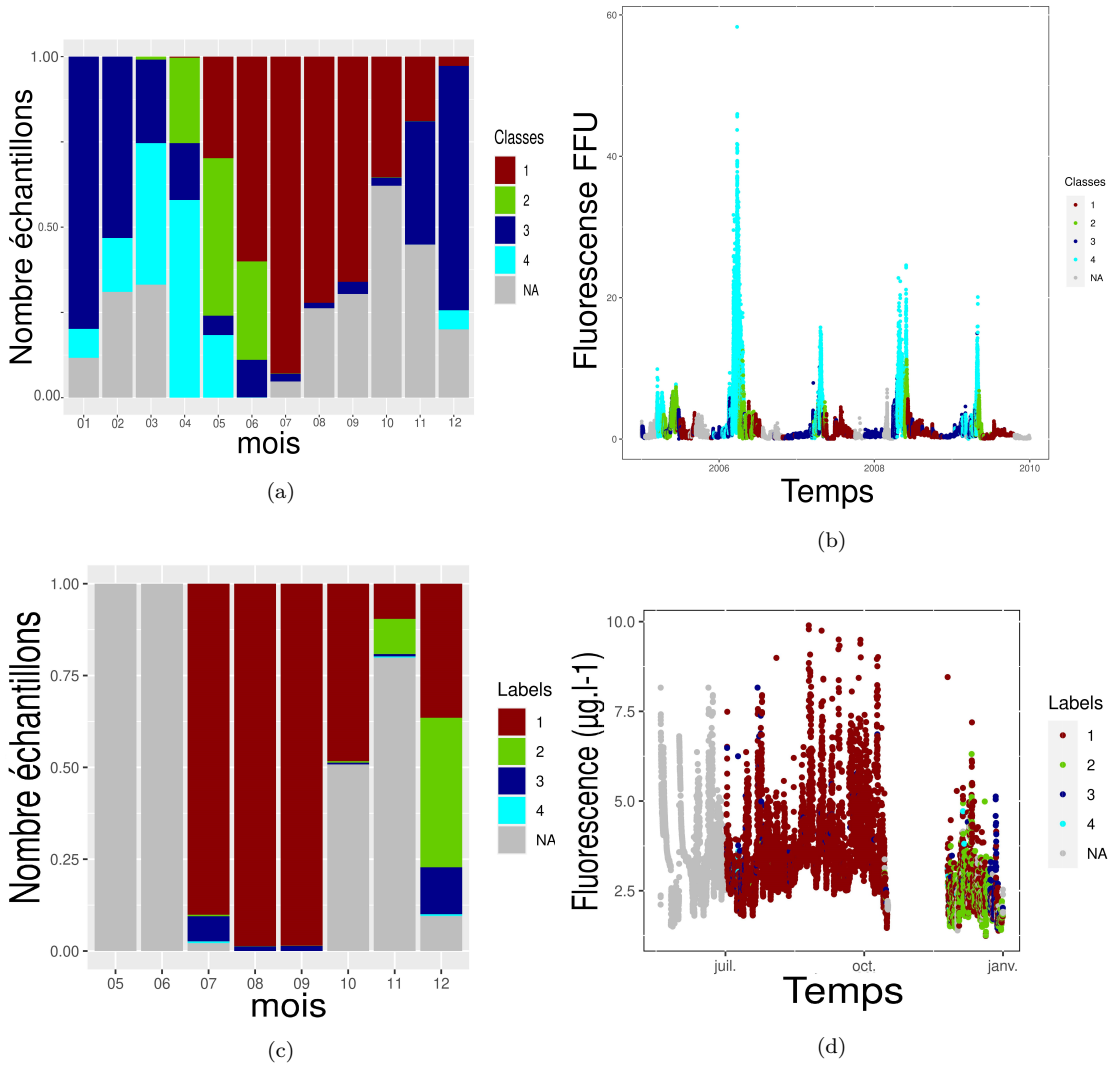
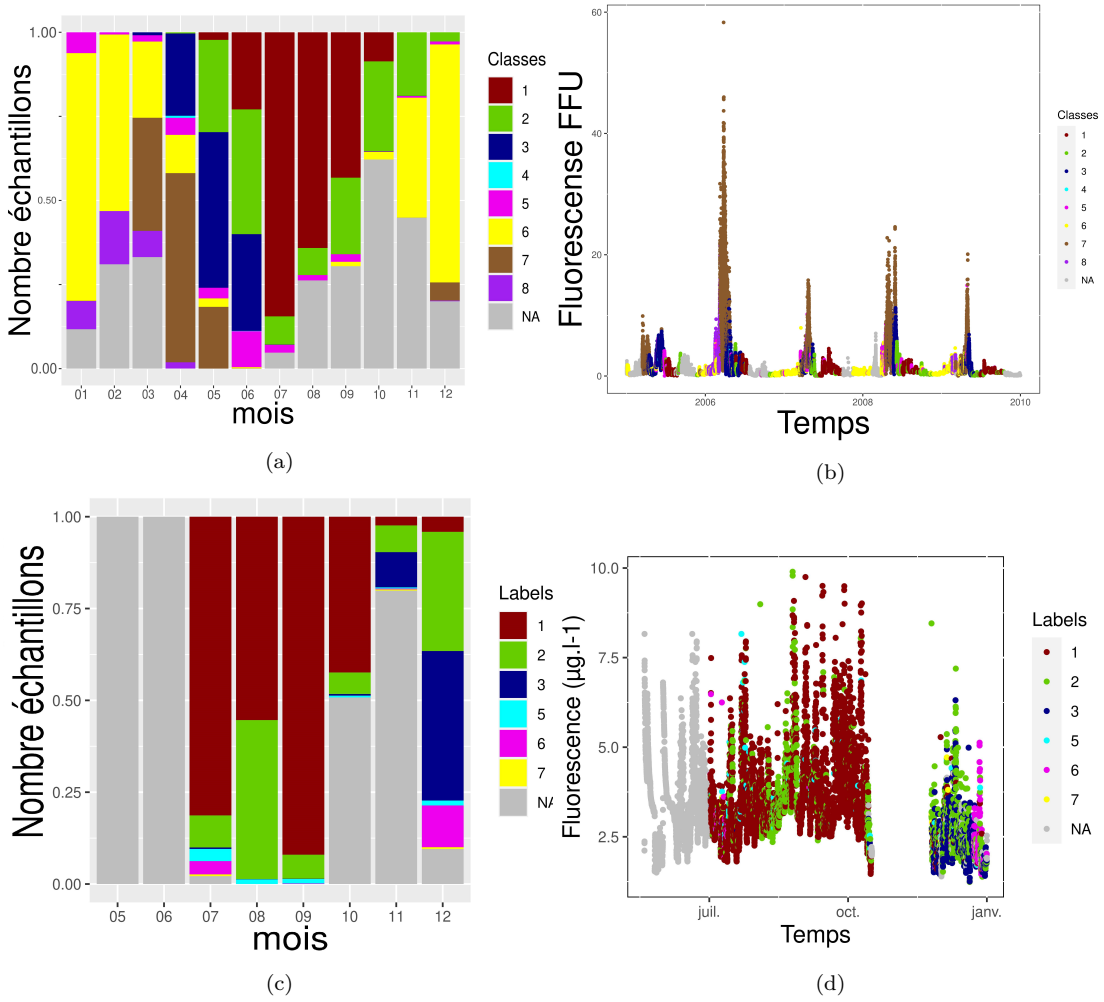


FIGURE E.4 – Comparaison au niveau 1 de la dynamique temporelle entre les classes labellisées par *M-SC* à MAREL-Carnot de 2005 à 2009 et les labels prédits pour MesuRho sur la période 2010-2014. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MesuRho et de la répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot, d) labels prédits pour MesuRho sur 2009.



(a) (b) (c) (d)

FIGURE E.5 – Comparaison au niveau 2 de la dynamique temporelle entre les classes labellisées par *M-SC* à MAREL-Carnot de 2005 à 2009 et les labels prédits pour MesuRho sur la période 2009. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MesuRho et de la répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot, d) labels prédits pour MesuRho sur 2009.



(a) (b) (c) (d)

FIGURE E.6 – Comparaison au niveau 3 de la dynamique temporelle entre les classes labellisées par *M-SC* à MAREL-Carnot de 2005 à 2009 et les labels prédits pour MesuRho sur la période 2009. Fréquence d'occurrence par mois a) de chaque classe labellisée par M-SC pour MAREL-Carnot, c) de chaque label prédit pour MesuRho et de la répartition sur le signal de fluorescence b) des classes issues de M-SC pour MAREL-Carnot, d) labels prédits pour MesuRho sur 2009.

Annexe F

Valorisations

Cette annexe présente l'intégralité des communications et publications faites lors de cette thèse. Elles sont triées par thématiques, catégories puis dates. Les deux articles publiés dans IEEE OCEAN19-Marseille et dans la spéciale issus *"Applications of Machine Learning in Marine Ecology Studies"* de *JMSE-Journal of Marine Science and Engineering* sont joints à cette annexe. Le bilan de formation et les activités supplémentaires sont aussi rappelés.

F.1 Développements méthodologique

— Publications [2] :

2019

- **Grassi K.**, Poisson-Caillault E., Lefebvre A. , 2019. Multilevel Spectral Clustering for extreme event characterization. OCEANS 2019 - Marseille, Marseille, France, 2019, pp. 1-7. doi : 10.1109/OCEANSE.2019.8867261

2020

- **Grassi K.**, Poisson-Caillault E., Bigand A., Lefebvre A., 2020. Comparative Study of Clustering approaches Applied to Spatial or Temporal Pattern Discovery. Journal of Marine Science and Engineering (ISSN 2077-1312). Special issues Applications of Machine Learning in Marine Ecology Studies. Accepted, sous presse Nov. 2020.

— Colloques internationaux [4] :

2018

- Poisson-Caillault E., **Grassi K.**, Phan T.T.H., Dezecache C., Prygiel J., and Lefebvre A., 2018. DTWBI and uHMM R-packages for multivariate time series preprocessing and interpretation. Colloque Earth science meeting, 22-26 octobre 2018, Lille.

- Poisson-Caillault E., Dezecache C., Phan T.T.H., **Grassi K.**, Prygiel J., Lefebvre A.,2018. Data completion, characterization of environmental states and dynamics using multiparameter time series : DTWBI, DTWUMI & uHMM R-packages. General Assembly H2020 JERICO-NEXT. 24-27 septembre, 2018, Galway, Ireland.

2019
- **Grassi K.**, Poisson-Caillaut E., Lefebvre A., 2019. Extrem Event detection from multivariate data time series. Application to marine observation. Journée IA atmosphère océan climat, 5-7 Février, 2019, Rennes, France.

- **Grassi K.**, Poisson-Caillault E., Bigand A., Lefebvre A. Multi-level Spectral Clustering for extreme event characterization. Juin 2019. Marseille, France.

— Colloques nationaux [2] :

2019
- Poisson-Caillault E., **Grassi K.**, Lefebvre A.,2019. Machine Learning et Dynamique Phytoplanctonique. Congrès ULCO, Avenir Littoral. 13 mars, 2019. Calais, France.

- Lefebvre A., **Grassi K.**, Devreker D., Poisson-caillault E., 2019. Phytoplancton, hydrologie et Haute Fréquence : développements numériques. Séminaire Ifremer / Fondation Tara Océan, 5 juin 2019, Maison de l'Océan, Paris.

— Communication interne [1] :
2018
- **Grassi K.**, Poisson-Caillault E., Bigand A., Lefebvre A., 2018. Multi-level spectral clustering for extreme event discovery. Journée Intelligence Artificielle du laboratoire Informatique et Signal de la Côte d'Opale, ULCO/LISIC, 18 juin 2018, Calais.

F.1.1 Observation et caractérisation des états environnementaux

— Publications [2] :
2019
- **Grassi K.**, Poisson-Caillault E., Lefebvre A. , 2019. Multilevel Spectral Clustering for extreme event characterization. OCEANS 2019 - Marseille, Marseille, France, 2019, pp. 1-7. doi : 10.1109/OCEANSE.2019.8867261

2020
- **Grassi K.**, Poisson-Caillault E., Bigand A., Lefebvre A., 2020. Comparative Study of Clustering approaches Applied to Spatial or Temporal Pattern Discovery. Journal of Marine Science and Engineering (ISSN 2077-1312). Special issues Applications of Machine Learning in Marine Ecology Studies. Accepted, sous presse Nov. 2020.

— Colloques internationaux [4] :
- Lefebvre A., Devreker D., **Grassi K**., Poisson-Caillault E., 2017. Analyse de tendance et classification spectrale couplée à un modèle de Markov caché. Colloque EVOLECO : EVOlution à Long terme des Ecosystèmes COtiers : Vers une mise en évidence des forçages et des processus associés, 5-7 décembre 2017, Bordeaux.

2018

- **Grassi K.**, Devreker D., Bigand A., Poisson-Caillault E., Lefebvre A., 2018. Trend analysis (TTA tools) and unsupervised clustering (uHMM tools) to characterize environmental

events from low and high Frequency data series. General Assembly H2020 JERICO-NEXT. 24-27 septembre 2018, Galway, Ireland.

2019

- Lefebvre A., **Grassi K.**, Poisson-caillault E., 2019. Identification of spatial Hydro-biological structures by spectral clustering. Towards implementation of machine learning for Ferry Box data processing. FerryBox Workshop, 24-26 mai 2019, Genoa, Italie.

- **Grassi K.**, Poisson-Caillaut E., Lefebvre A., 2019. Extrem Event detection from multivariate data time series. Application to marine observation. Final General Assembly H2020 JERICO-Next, 2-5 Juillet 2019, Brest, France.

— Colloques nationaux [3] :
2018
- **Grassi K.**, Phan T.T.H., Poisson-Caillault E., Bigand A., Devreker D., Lefebvre A., 2018. Results from measurements in the Eastern English Channel MAREL Carnot-station. Colloque H2020 JERICO-Next, 19-21 Mars 2018, MIO, Marseille.

- Lefebvre A., **Grassi K.**, Phan T.T.H., Devreker D., Bigand A., Poisson-Caillault E., 2018. Automated tools for analyzing outputs of automated sensors : High frequency Data. Third JERICO-NEXT Workshop on Phytoplankton Automated Observation. 19-21 Mars, 2018 - M.I.O., Marseille, France.

2019

- Devreker D., **Grassi K.**, Poisson-caillault E., Lefebvre A., Grassi K., 2019. Présentation des outils numériques développés pour l'exploitation des données haute-fréquence FerryBox. CGFS workshop, 25 avril 2019, Boulogne-sur-mer, France.

— Communication interne [2] :
2018

- **Grassi K.**, Poisson-Caillault E., Phan T.T.H., Dezecache C., Devreker D., Bigand A., Lefebvre A., 2018. Méthodes et outils numériques innovants pour l'exploitation des données d'observation (haute et basse fréquences) du milieu marin. Journée MMN, 20 Novembre 2018, Le Havre, France.

2019
- **Grassi K.**, Poisson-Caillault E., Bigand A., Lefebvre A., 2019. Identification d'éco-hydro-régions par classification spectrale multi-niveaux à partir de systèmes d'observations hautes résolutions, multi-paramètres. Journée MMN, 11 Octobre 2019, Lille, France.

F.2 Météorologie

— Colloque international [1] :
2020

- **Grassi K.**, Lefebvre A., and Poisson Caillault E., 2020. Short-term rainfall analysis by spectral clustering in Ivory Coast. Eumetnet Workshop on Artificial Intelligence for weather and climate. 26 février 2020, Brussels, Belgium.

— Colloque national [6] :
2018

- **Grassi K.**, Poisson-Caillault E., Lefebvre A.,2019. Machine Learning et observatoire marin. Séminaire Météo-France. 9 avril, 2019. Toulouse, France.

F.3 Rapports techniques

— Rapports techniques [4] :
2018
- Lefebvre A., Poisson-Caillault E., **Grassi K.**, 2018. MAREL Carnot : Rapport n° 12 : Bilan d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan 2017. Ifremer/RST.LER.BL/18.05, 24 pages.

2019
- Lefebvre A., **Grassi K.**, Poisson-Caillault E., 2018. MAREL Carnot : Rapport n° 13 : Bilan d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan 2018. Ifremer/RST.LER.BL/18.05, 25 pages.

- **Grassi K.**, Poisson-Caillault E., Lefebvre A., 2019. Contribution : Bilan d'activité CPER MARCO- Axe 1- Programme 3- Année 2019. Sept 2019, Rapport technique Projet CPER MARCO.

2020
- **Grassi K.**, Poisson-Caillault E., Lefebvre A., 2020. MAREL Carnot : Rapport n° 14 : Bilan d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan 2019. Rapport Ifremer/ODE/ LITTORAL/LER-BL/20.05, 25 pages

F.4 Valorisation Grand public

— Vulgarisation - Valorisation Grand public [2] :
2018
- **Grassi K.**, Dezecache C., Phan T.T.H., Poisson-Caillault E., Bigand A., Lefebvre A., 2018. High Frequency Observation and Characterization of the Marine Environment : Completion and Spectral Clustering of Multivariate Tiùe Series. Journée CPER MARCO-Fête de la science - 11 octobre 2018, Boulogne sur mer (Nausicaa)

- **Grassi K.**, Phan T.T.H., Poisson-Caillault E., Bigand A., Devreker D., Lefebvre A., 2018. Présentation des travaux en cours : Multi-level sepectral clustering : Validation sur un jeu de données artificiel et application sur le jeu de données MAREL-Carnot. Journée des étudiants, 8 Juin 2018, ifremer, Boulogne-sur-Mer.

F.5 Bilan productions scientifiques et techniques.

Tableau F.1 – Bilan productions scientifiques et techniques.

Indicateurs	Nombres
Publications dans des revues avec comité de lecture, de rang A	1
Publication parues dans d'autres revues et dans des ouvrages scientifiques et technologiques avec comité de lecture	1
IF moyen des publications	2,79
Rapports liés à : - réseaux de surveillance/observation - projets - communauté européenne- Working Group-FAO-	 3 1 0
Articles de vulgarisation, conférences et communications internes ou grand public	5
Ouvrages / chapitres d'ouvrages	0
Communications dans des colloques et congrès, posters	15
Brevets	0
Licences	0

F.6 Enseignements

- 20h00 équivalent TD maximum, au sein de la structure Services Centraux, pour la formation PRREL FSE, dans la discipline 27 - Informatique
- 5h00 équivalent TD maximum, au sein de la structure CGU Boulogne, pour la formation M1 Sciences de la Mer, dans la discipline 27 - Informatique
- Encadrelent d'un stage de M1-Traitement du Signal et des Images (TSI) au sein de l'Université Côte d'Opale ; sur la thématique : « Apprentissage et Prédiction d'épisodes extrêmes. Avec un réseau de capteurs non dédiés et souvent défaillants. Application à la détection des pluies et orages dans les pays en développement. »
- Encadrement d'un projet Industriel Collectif PIC- Génie Industriel au sein de EIL Côte d'Opale - École d'ingénieur du Littorale ; sur la thématique : « Analyse et Prévision par *Machine Learning* et Apprentissage Artificiel de Données de Pluie »

F.7 Formations

F.7.1 Formations doctorales et crédits acquis

Afin d'élargir mes connaissances tout au long de ma thèse, l'ED a défini un plan de formation, avec un large choix de possibilités de formation.

Le principe est le suivant, le doctorant doit obtenir un minimum de 60 crédits au cours des trois années de thèse avec un minimum de 10 crédits par domaine de formation :

- Scientifique (cours ED, cours Master, école d'été, ...)
- Ouverture (langues, mobilité internationale, modules outils)

— Professionnalisation (formations proposées par le Département Carrières – Emploi du Collège Doctoral, Doctoriales, . . .)

Les formations sont laissées au choix du doctorant, mais une formation à l'éthique est obligatoire. J'ai donc suivi les formations suivantes lors de cette thèse.

— Cours "Biostatistiques 2" du M1 FOGEM (Fonctionnement et Gestion des Écosystèmes Marins) de P-A Hébert et E.Poisson Caillault(25h) - Université Littoral Côte d'Opale **(Crédit : 10)**
— Cours "Apprentissage artificiel" du M1 informatique (Parcours "Ingénierie des Systèmes Informatiques Distribués" (ISIDIS)) de Fabien Teytaud (39h) - Université Littoral Côte d'Opale (ULCO) **(Crédit : 20)**
— ED Formation Parcours **(éthique)** : Développement et valorisation des compétences, communication : Améliorer la visibilité de sa production scientifique, 14 mai 2018. **(Crédit : 2)**
— ED Formation numérique : Composition efficace du mémoir de thèse (et autre documents) avec LaTex-niveau Avancer , 28-29 mai et 4-5 juin 2018. **(Crédit : 10)**
— Cours Anglais (9h/25h) d' avril à juin 2018 . **(Crédit : 5)**
— ED Formation Parcours : Définir et formuler son projet professionnel : 25-26- février 2019. **(Crédit : 7)**
— ED Participation à un événement de diffusion : La préparation et l'animation du stand Ifremer lors des Fêtes de la Mer, m'ont permis d'obtenir **(Crédit : 5)**.
— ED Participation à un événement de diffusion : la participation à plusieurs colloques internationaux **(Crédit : 5)**.

Au total, 64 crédits sur 60 crédits obligatoires ont été validés par l'école doctorale.

F.7.2 Autres Formations : Campagne CGFS

la campagne CGFS (Channel Ground Fish Survey) s'intègre dans le programme européen de suivi des ressources halieutiques, qui permet d'obtenir un ensemble de données relatives aux stocks exploités (maturité, structure en taille/âge, indices de recrutement). La série temporelle initiée en 1988 est utilisée chaque année par les groupes européens d'évaluation des stocks qui déduisent l'état de santé des principales espèces commerciales. Réalisée sur le N/O Thalassa, la campagne CGFS permet également un échantillonnage et une meilleure connaissance de l'ensemble de l'écosystème, répondant à la fois aux demandes de suivi des écosystèmes (DCSMM) et à la mise en place d'une approche écosystémique des pêches au niveau communautaire. Ainsi, les caractéristiques physico-chimiques de l'eau, les communautés de phytoplancton et zooplancton, l'abondance d'œufs de poissons, la composition spécifique des communautés nectoniques sont mesurées et analysées tout au long de la campagne Coppin et Travers -Trolet 2017.

Ma participation à la campagne CGFS 2019 m'a permis d'acquérir des compétences et des connaissances sur l'acquisition de données marines. J'ai pu suivre les méthodes d'échantillonnages et d'analyses en mer. Ayant été intégrée à l'équipe hydrologique, j'ai donc réalisé des mesures de sondes (CTD, Fluoroprobe), des prélèvements d'eau (bouteille niskin), des prélèvements de phytoplancton et zooplancton (filet WP2, babynet, manta), mesures HF (Ferry-Box). J'ai aussi

participé aux analyses : Filtrations, Cytométrie, Fluorométrie, identification d'espèces (Zooscam). Ces différentes manipulations m'ont offert une vue d'ensemble sur les différents moyens de prélèvements (BF et HF) et m'ont permis de faire le lien avec les différentes variables essentielles (EOVs) que j'utilise pour la classification des événements extrêmes.

F.8 Article 1 : OCEAN'19 (2019)

Grassi K., Poisson-Caillault E., Lefebvre A., 2019. Multilevel Spectral Clustering for extreme event characterization. OCEANS 2019 - Marseille, Marseille, France, 2019, pp. 1-7. doi : 10.1109/OCEANSE.2019.8867261.

Multilevel Spectral Clustering for extreme event characterization

Kelly *GRASSI*[1,2,3]
(1) WeatherForce
F-31000 Toulouse, France
kelly.grassi@ifremer.fr

Emilie POISSON CAILLAULT
(2) ULCO-LISIC
F-62228 Calais, France
emilie.poisson@univ-littoral.fr

Alain LEFEBVRE
(3) Ifremer, LER-BL
F-62321 boulogne-sur-mer, France
Alain.Lefebvre@ifremer.fr

Multilevel Spectral Clustering for extreme event characterization

Kelly $GRASSI^{1,2,3}$
(1) WeatherForce
F-31000 Toulouse, France
kelly.grassi@ifremer.fr

Emilie POISSON CAILLAULT
(2) ULCO-LISIC
F-62228 Calais, France
emilie.poisson@univ-littoral.fr

Alain LEFEBVRE
(3) Ifremer, LER-BL
F-62321 boulogne-sur-mer, France
Alain.Lefebvre@ifremer.fr

***Abstract*—Direct spectral clustering framework was first proposed to extract general pattern events within multivariate time series. This study investigated the way to identify extreme events, *i.e.* short duration and/or particular events, with no assumption about their emission date, duration and/or shape. A Multilevel Spectral Clustering (M-SC) architecture is proposed and compared with state-of-the-art clustering methods from a simulated manually labeled time series. Due to these promising empirical results, this new deep architecture is applied on marine field data.**

***Index Terms*—Spectral clustering, multilevel systems, extreme events, time series, environmental monitoring.**

I. Introduction

In marine ecology, pattern discovery within time series is crucial to help the understanding of ecosystem dynamics and to forecast and/or prevent harmful events. Therefore, many monitoring project initiatives promote the development and the implementation of Integrated Observing Systems that produce huge multivariate time series. The identification of environmental states in these nonlinear dynamic systems is a hard task. Indeed, patterns could have a large variety of distribution, shape and duration whether for recurrent, episodic or extreme events, *i.e.* short duration and/or particular events within general pattern. Moreover there is often no label or no oracle of these events. For instance considering the MAREL (Mesures Automatisées en Réseau pour l'Environnement Littoral) - Carnot dataset from the Eastern English Channel instrumented Station [1], no label of the Harmful Algae Bloom (HAB) or of environmental events is available.

So, we investigate the way to achieve a blind segmentation of these multivariate time series in events so as to facilitate the task of active learning by human expert, main step to build a specific environmental state and HAB forecasting system.

Segmentation methods could be divided according to the way they cut time series: by windows processing, by generative models, or from breaking points. In the first case, times series segmentation are based on window processing. Unsupervised approaches extract similar pieces of the series from fixed-length window [2] or sliding window by autocencoders [3]. In the second case, segmentation by generative models assume that time series is a mixture of motifs with an EM-step (Expectation Maximization): either [4] for uni-variate context, or Dirichlet Processes [5] for multivariate context and Hidden Markov Models [6]. In the last case, segmentation is processed according to breaking points (Extreme Values Theory or PIP: Perception Interest Points [7]). The Extreme values theory based on density function and frequency are judicious when time series follows some models/distributions. PIP-cut is based on a distance criteria combining time and level value on which ones, authors perform a clustering process to detect similar events. It is well adapted method to detect trend as financial prediction but is not be relevant to identify a specific events.

These methods cannot be directly applied in our nonlinear multivariate large data with such a high intra-variability. Another way is to apply clustering directly without any temporal cut/windows hypothesis and by considering collected multivariate points. In [8], a K-way Spectral Clustering (SC) approach has been proposed to discover recurrent environmental states from physico-chemical data. The authors used NJW spectral clustering algorithm (NJW-SC) [9] that consists in extracting spectral eigenvectors of a normalized Laplacian matrix issued from the data similarity matrix following by an unsupervised k-means clustering (K-means).

To detect both general environmental patterns and extreme events, we extend the direct clustering approach in [8] by a multilevel approach. A recursive/deep Spectral Clustering architecture named M-SC for Multilevel Spectral Clustering is proposed. From a labeled simulated multivariate time series, M-SC deep approach was first compared both with direct clustering approaches (K-means, NJW-SC) and hierarchical approaches: Hierarchical clustering (HC), Hierarchical spectral clustering (H-SC) [10], Bipartite spectral clustering (Bi-SC) [11]. Then, M-SC is applied to characterize different environmental states and their dynamics in the MAREL-Carnot data.

The rest of this article is structured as follows: Section II introduces our multilevel spectral clustering algorithm and its validation (data presentation, validation tools and criteria), Section III discusses the comparison results between M-SC and related approaches. Finally, Section IV concludes with a summary of the main findings.

II. Implicit Segmentation by Spectral Clustering

A. Data presentation

Labeled Simulated dataset $X(t) = \{x_i(t), i \in [1, 2, 3]\}$ is derived from three sine functions $x_i(t) = a_i \times \sin(2\pi f_i t + \phi_i) + b_i(t)$ of 1,000 time values. Each parameter $(a_i; f_i; \phi_i)$

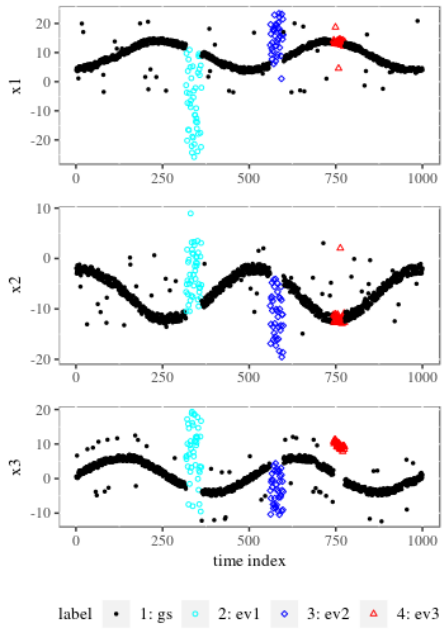


Fig. 1: Simulated dataset composed of three signals x_i with added specific events: two extreme events (ev1, ev2) and a sensor offset (ev3).

is different for each signal. Added noise $b_i(t)$ is generated from random uniform distributions with few offsets for some points. These global-shape signals (gs) are then disturbed to introduce three short events. Two strong variations similar to extreme events (ev1, ev2) were imputed on each signal $x_{i=1,2}$. The last event corresponds to an offset (ev3) of the x_3 values in order to simulate a sensor failure. Figure 1 illustrate this labeled time series $X(t)$.

The MAREL-Carnot system developed and implemented by Ifremer (French Research Institute for Exploitation of the sea) in 2004 is a moored buoy working in continuous and autonomous conditions. The instrumented buoy is located in coastal areas of Boulogne-sur-mer, France (eastern English Channel: 50°44'26.0"N , 1°34'03.7"E) with environmental conditions influenced both by marine coastal and fresh waters and the tidal conditions. Data are referenced in the SEANOE[1] system since 2015 [1]. Nineteen signals were measured over the period 2005-2008, including 2008. For these essential ocean variables (EOV) [12], [13], only nine uncorrelated variables (summarized in Table I) are selected to detect recurrent and rare events by clustering. The first six signals - sea level (Slev), sea temperature (T), PAR (Photosynthetically Active Radiation), practical salinity (Psal), concentration of dissolved oxygen (DOx), turbidity (Turb) - were collected every 20 minutes, while the last three signals - nitrate + nitrite (N), phosphate (P), silicate (S) - were collected every 12 hours. Other registered parameters are only used for the interpretation of the analytic results. In particular, fluorescence signal, proxy of the phytoplankton biomass, is not included in the clustering process. Moreover, the objective is also not to influence the definition of cluster by this strongly structuring variables when dealing with phytoplankton bloom.

[1]SEANOE: Sea scientific Open Data Edition

TABLE I: MAREL Carnot parameter characteristics involved in clustering after sensor range correction.

Parameters (units)	mean-median	min-max	Q1-Q3
Slev (m)	4.91 - 4.85	0.52 - 9.26	2.96 - 6.87
T level (°C)	12.59 - 12.40	3.60 - 21.40	8.50 - 17.10
Turb (NTU)	11.75 - 7.20	0.00 - 148.90	3.70 - 14.00
Psal (PSU)	33.36 - 33.52	20.41 - 34.92	33.05 - 33.88
PAR(µmol m s)	285.90 - 12.70	0.0 - 2487.8.	0.00 - 417.50
DOx ($mg\,l^{-1}$)	8.27 - 7.93	5.00 - 16.38	7.03 - 9.40
N ($\mu mol\,l^{-1}$)	20.32 - 17.70	0.01 - 99.54	7.34 - 28.47
P ($\mu mol\,l^{-1}$)	1.250 - 0.737	0.000 - 10.000	0.475 - 1.040
S ($\mu mol\,l^{-1}$)	6.699 - 5.350	0.010 - 48.840	2.365 - 9.707

B. Multilevel spectral clustering approach

A divisive hierarchical spectral clustering approach is proposed to build a multilevel implicit segmentation of a multivariate time series X (N signals of M time samples). Figure 2 illustrate the main scheme of the deep architecture. So, this architecture is obtained by applying recursively a variant of NJW-SC with an automatic determination of cluster number (K) at each level (l) and sub-states (s_l). It should be noted that time is not taken to account in the clustering process. At each level and state, a spectral clustering algorithm (Algorithm 1) is applied from the related observations, *i.e.* the reduced similarity matrix $W = W_{s_l}(X)$, $X \in s_l$. At the first level, input of M-SC is the set of M points $X = \{x_m \in \mathbb{R}^N, m \in [1, ., M]\}$. For each step, the eigenvectors v and eigenvalues e are extracted from the normalized Laplacian matrix L issued from similarity matrix W of X. Only dominant eigenvalues (e>0.9) are considered to ensure good separability of clusters (see [9]). K cluster number is computed from the maximal gap between two successive eigenvalues v of Laplacian matrix L of W. The K largest vectors v represent a new feature space (a spectral space) where a clustering step is performed. For clustering, initially proposed K-means are replaced by K-medoids algorithm, also called Partition Around Medoids (PAM). PAM is preferred for large databases. It avoid gravity center computation and save a medoid point corresponding to an existing observation for interpretation convenience.

M-SC output is a matrix S of labels where $s[m, l]$ corresponds to the cluster number of x_m at level l. The depth (level number) of M-SC depends on the intended use and the required standard of interpretation. It could be tuned by Dunn or silhouette criteria.

For large database as MAREL Carnot set, the maximal number of cluster accepted for each state have been set to Kmax = 20, it could be tune according to the dataset. The hierarchical process was voluntary stopped to three levels. Moreover, few pre-processing steps are necessary to facilitate inter-comparability and operability of the different variables: unit standardization, range correction, temporal alignment, data completion and a normalization step. Then for each

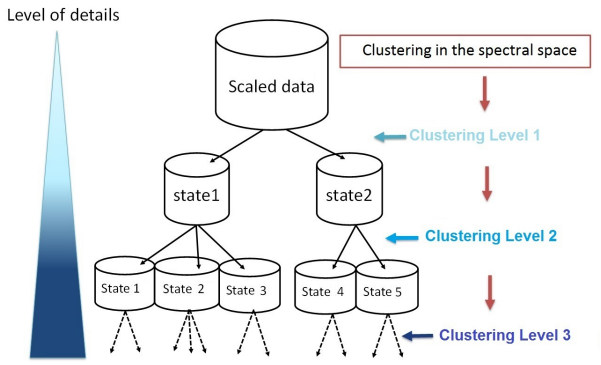


Fig. 2: Multilevel Spectral Clustering scheme.

Algorithm 1 Spectral-PAM

Require: $W(X)$ a $M \times M$ Gram matrix , Kmax maximum accepted clusters
Variables: W, D, L, e, v , m
 # Compute Laplacian
$\forall i, w_{ii} = 0$
Build $D = d_{ij}$ the Degree matrix $M \times M$
$d_{ii} = \sum_j w_{ij}$ et $d_{ij,i \neq j} = 0$
$L = D^{-1/2} W D^{-1/2}$ the Laplacian matrix
$\{e,v\}$=eigen(L,Kmax)
 # Compute number of clusters K
$K = \underset{i-1}{\operatorname{argmax}} (e_i - e_{i-1}),\ e_i > 0.9,\ i \geq 2$
 # Clustering in eigen-space
Select the K-largest eigenvectors v of L;
Form $V = [v_1 v_2 \dots v_K]$ matrix $M \times K$
Form Y matrix from the row-normalization of V.
Label = PAM (Y,K)
return Label

dataset the similarity matrix W is computed from a gaussian kernel. The simulated dataset is labeled manually in 4 clusters (gs, ev1, ev2, ev3). Therefore, hierarchical approaches (HC, H-SC, Bi-SC, M-SC) are cut in order to obtain 4 clusters then 8 clusters. Direct clustering approaches (K-means, NJW-SC) is implemented with $K = \{4, 8\}$.

C. Validation Tools

Performance of the MS-C approach was compared with published algorithms with direct clustering approaches (K-means, NJW-SC) and hierarchical clustering (HC, H-SC, Bi-SC).

Five performance indicators per algorithm are computed from the manual labeling and the obtained partitions of the simulated dataset [14]: the Adjusted Rand index (ARI), The Dunn index, the Silhouette score and the Accuracy (Table II) and Confusion/Agreement tables (Table III).

The Adjusted Rand index (ARI) is the corrected-for-chance version of the Rand Index (RI) available in fossil:adjustedRandIndex [15]. RI measures the number of agreements in two partitions (same grouping elements and same separating pairs) over the total number of pairs. Rand Index will never actually be zero. It can yield negative values if the index is less than the expected index.

TABLE II: Clustering performance indicators from ground-truth labels of a simulated dataset (best results in bold): Adjusted Rand Index (ARI), Dunn and Silhouette (Sil.) index, total accuracy (Acc.), detected events

	ARI	**Dunn**	**Sil.**	**Acc.**	**Detected events**		
ground-truth	1	0.008	0.15	-	ev1	ev2	ev3
KM8	0.37	0.010	**0.40**	0.92	35/45	16/38	none
HC8	0.41	**0.022**	0.34	0.90	21/45	none	none
H-SC8	0.42	0.007	0.21	0.89	31/45	none	none
NJW-SC	0.39	0.006	0.26	0.88	36/45	none	none
Bi-SC8	0.33	0.003	0.01	0.88	none	none	none
M-SC8	**0.43**	0.007	0.28	**0.94**	42/45	31/38	29/32

The Dunn index, clValid:dunn [16] is a ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn index has a value between zero and infinity, and should be maximized. It reflects the data/cluster separability. A Dunn index close to zero shows very connected data, while a large score indicates easily separable data.

The Silhouette score, cluster:silhouette [17] is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A large Silhouette (almost 1) is very well clustered, while a small Silhouette (around 0) means that the observations are connected between two clusters.

The Accuracy is the percentage of well recognized labels according to a given ground truth. Next majority vote principle is applied to assign each cluster to a label.

Confusion/Agreement tables between partitions and its labeling by majority vote rules are used to show if our expected events are detected.

III. Results and Discussion

A. Simulated time series

Note that simulated time series is composed of three events that we expect to detect and isolate from the global shape (gs). With the M-SC approach, two states were detected in level 1. Four clusters were obtained at level 2, and eight at level 3.

Figure 3 are the clusters of the time series X reported on one of its signals (x_3), as an illustration, by each different method obtained directly for $K = 4$ and $K = 8$ or by tree-cut. Each color corresponds to a cluster number. Globally, with a $K = 4$ setting for all algorithms the extreme states are never or poorly detected. That's why, only results of $K = 8$ are commented in the rest of this section.

Table II summarizes indicators for K=8 clusters obtained by a direct or multilevel way (K=8 required to isolate the three events for M-SC and $K >> 20$ for other methods). From ground-truth labels, low Dunn index (0.008) and the averaged silhouette score (0.157) show the complexity to isolate the three events named from the global-shape series (gs). In general for all algorithms, Dunn index are low due to high local connexity between the events and gs. Except for M-SC, spectral methods have lowest indexes and accuracies. High total accuracy is due to the large proportion of gs. HC8

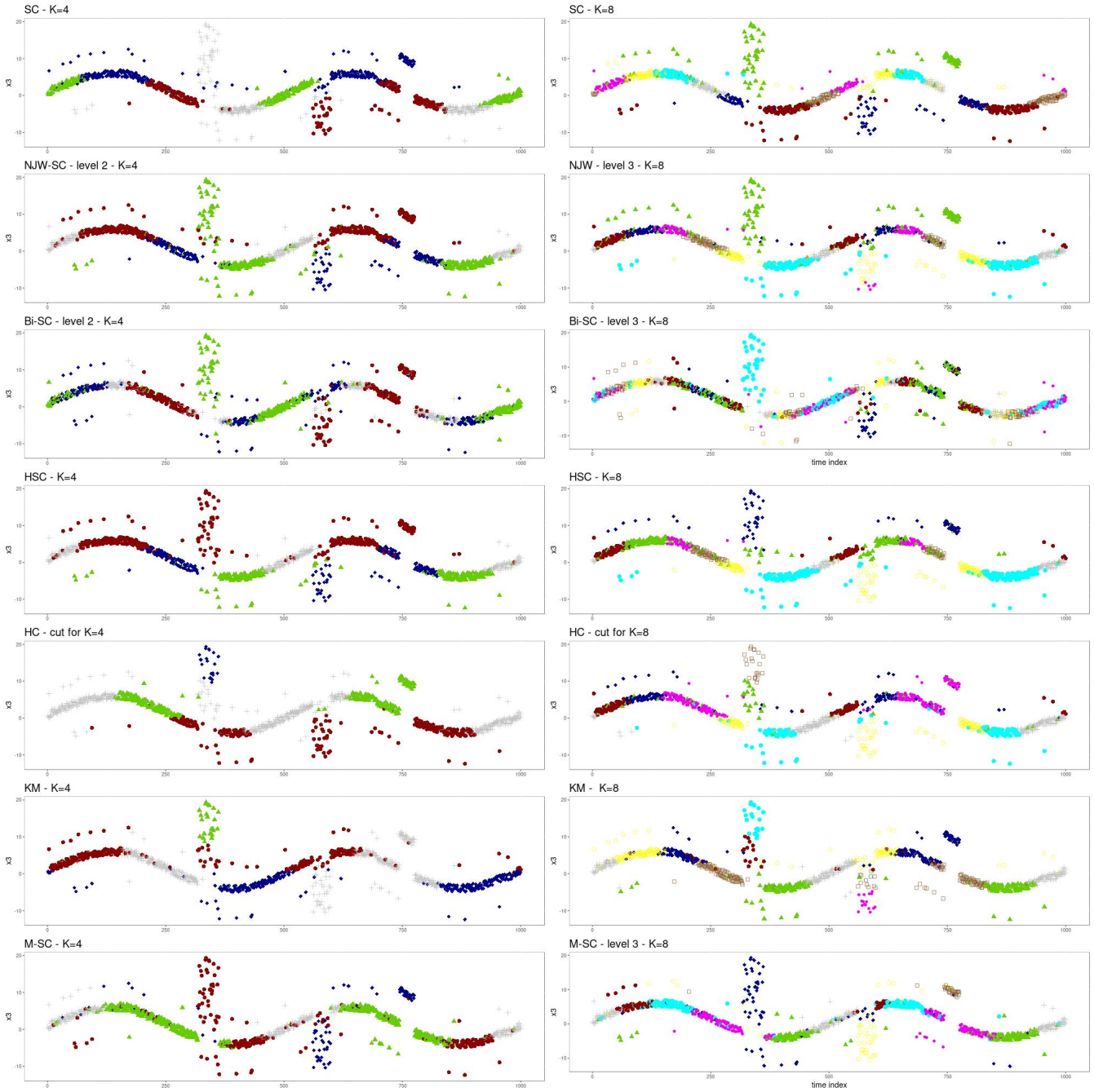


Fig. 3: Color-labeling of each K-clustering from X reported on the signal x_3. (HC4/HC8: Hierarchical clustering cut for $K = 4$ or $K = 8$, SC4/SC8: K-way Spectral clustering, M-SC4/M-SC8: our proposed approach.

TABLE III: Table of confusion between manual labeling and automatic clusterings. Bold values corresponds to label assignment by majority vote rules.

KM8	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	tot
sg	**224**	12	**176**	**146**	0	0	**157**	**170**	885
ev1	0	**14**	6	1	**21**	0	2	1	45
ev2	0	1	1	2	0	**16**	6	12	38
ev3	0	0	0	31	0	0	1	0	32
tot	224	27	183	180	21	16	166	183	
NJW-SC8									tot
sg	**98**	**187**	35	**110**	**101**	**105**	**128**	**121**	885
ev1	0	5	**36**	1	1	0	2	0	45
ev2	0	2	1	26	0	1	8	0	38
ev3	0	0	31	0	0	0	1	0	32
tot	98	194	103	137	102	106	139	121	
M-SC8									tot
sg	**209**	**76**	**163**	28	**163**	**228**	16	2	885
ev1	0	0	0	**42**	0	0	3	0	45
ev2	0	0	0	7	0	0	**31**	0	38
ev3	0	0	0	1	0	0	2	**29**	32
tot	209	76	163	78	163	228	52	31	

with the best Dunn index (0.022) only detects a part of ev1. K-means have the best silhouette but fails to isolate ev3. Bi-SC, closed to M-SC architecture detects no events. And, HSC (agglomerative approach in spectral space) and NJW-SC for K=8 are not able at this level to detect ev2 and ev3.

Thus, M-SC have greater indicators than spectral methods and better ARI and event detection ability than all others. Its false-positives are mostly confusions with closely connected points in the signal.

Table III showed the confusion matrices between manual labeling (ev1, ev2, ev3 and the signal gs) and clustering algorithm as KM8, NJW-SC8 and M-SC8. These algorithms have the best the three events recognition rate of all methods. According to majority vote, for KM8, ev1 is divided in two clusters S2 including almost same part of gs and S5. Cluster S3 of NJW-SC8 was assigned to ev1 (36 points), but including almost all ev3 (31 points) and 35 points of global-shape signal. Similarly for M-SC8, S4 was assigned to ev1 with close points of gs. Ev2 is detected only by KM8 and M-SC8. For NJW-SC8, ev2 was included in S4 and S6 identified as gs. Ev3 points are not connected with the global-shape signal for $x3$ signal. It could correspond to an offset of gs, so it is expected easier to detect. That is not the case for KM8 and NJW-SC8. M-SC8 successes to isolate it with only 2 false positives in the cluster S8.

Our experiments were conducted to have unbiased validation of all method and try to obtain the best results. K-setting conducted for all algorithms with 10 repetitions to assess output stability and a K-values have been increase until detection of the three events. In this case, K-means fails to isolate the three events before $K = 13$ with still 32% confusions for these clusters. HC dendrogram requires a K=12-cut to detect the three events with only 31/45 points detected in ev1 and 17/38 in ev2, *i.e.* 33% of event confusions. K-means and HC requires respectively $K > 241$ and $K > 206$ to obtain at least 85 percent of well recognized event rate. For spectral clustering methods, for two levels ($K = \{2, 4\}$, the specific events are never or poorly detected. The minimum number required to isolate each event is level 3 for M-SC with K=8 clusters.

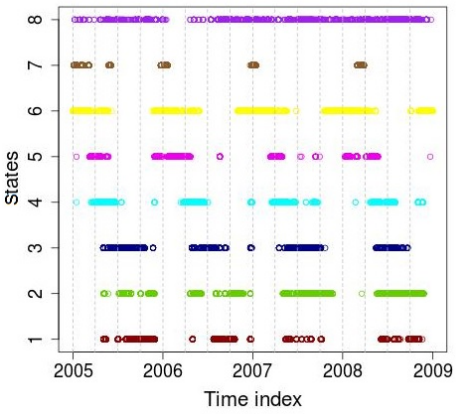
So, M-SC seems to have good ability for detecting the three simulated events with only a computation step of three levels. It isolated 93.3% of ev1, 91.4% of ev2 and 90.6% of ev3.

B. MAREL Carnot dataset

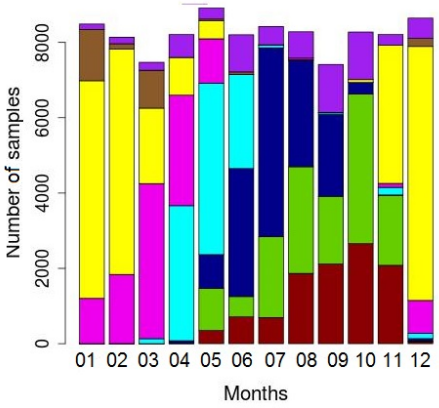
M-SC approach was also tested on high frequency data in coastal water monitoring context. Rousseeuw et al. [8] have proposed a direct K-way approach to detect environmental states in multivariate time series (available in R uHMM:FastSpectralNJW). M-SC are applied on the same dataset (MAREL Carnot) for comparison and we expect to detect also the global phytoplankton dynamics and to go further by detecting extreme events [18]. Thus, in this part, the M-SC clusters are labeled as specific environmental states.

At level 1, contrasted ecological states are identified: the productive and non productive period. At level 2, each of these two main periods are cut into sub-periods key environmental states: pre-bloom, bloom and post-bloom. level 3 allows going beyond the identification of general patterns strongly structured by environmental drivers in [8]. Intra-weekly or even hourly events could be identified from this level. For example, state s7 corresponds to winter events of a few hours/days where phosphate concentrations are particularly high (Figure 4). Then, in s7 value are high correlation between phosphate and turbidity variables. This matches sediment resuspension, which leads to a phosphate desorption. These kinds of events are linked to meteorological events (storms) or to anthropogenic activities (*e.g.*, dredging). This desorpted phosphate (in theory exhausted at this time) will become available in the water column and could be consumed by phytoplankton, which need phosphate for growth. This could explain secondary blooms superimposed onto the main general dynamics pattern. It appears that these secondary blooms are of particular interest when trying to further knowledge on HAB. Identification of such events may help to identify environmental drivers and therefore measures to implement in order to improve environmental quality (*i.e.*, reduction of nutrient inputs), maintain ecosystem structure and functions, and most importantly, to protect human health.

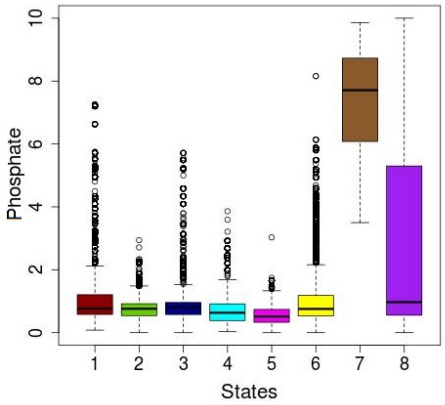
A second interest, M-SC highlight successive events by clustering process without time dimension. In fact, its possible to detect pressures and impacts states. For example, stats s4 (cyan), s5 (pink), s6 (yellow) show the biogeochimical response of nutrients input (Nitrate) and there are logical chronologies between this stats (Figure 4a,4d). As a first step, s6 is defined by high nitrate concentration and low phytoplankton and dissolved oxygen concentrations. As a second step (s5), they are an increase of phytoplankton and dissolved oxygen concentration and decrease of nitrate



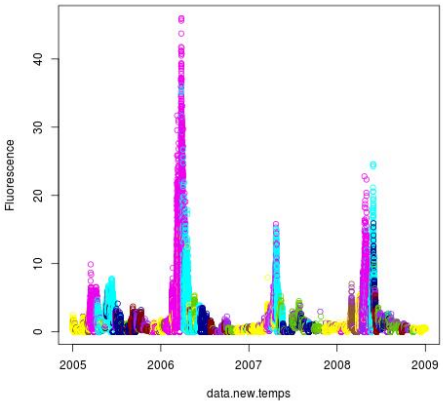
(a) State dynamics



(b) State distribution per month



(c) Phosphate box-and-whisker plot for each state



(d) State color in Fluorescence signal

Fig. 4: MAREL Carnot clustering over the period 2005-2008 (level 3).

concentration (Figure 5). For fine (s4), the variables are gradually re-balancing. These three steps are characteristic of phytoplankton bloom linked to nutrient uptakes. Thus, it is possible to detect states and impacts in response to natural and/or anthropogenic pressures (as described DPSIR framework [19]). Next, global dynamics and thereafter more specific events like secondary bloom or re-suspension events has been defined.

In [8], general patterns was identified and then labeled: non productive phytoplankton period, spring bloom, autumnal bloom, etc. M-SC approach was able to characterize also these different environmental states and propose more specific events thanks to its multilevel context. At level 3, M-SC isolates extreme events whose period can be infra-weekly or even hourly like phosphate desorption.

IV. Conclusions and perspectives

Extreme event detection in high-resolution and high-dimension time series processing are crucial for environment monitoring. The key for this successful application lies in applying the right combination of numerical methodology and representation to extract relevant information for stakeholder. In this case, we optimized processing without losing key-information for extreme events detection.

A Multilevel Spectral Clustering (M-SC) was proposed to segment multivariate time series from general patterns up to extreme events by unsupervised way. The test on simulated dataset with high local connexity between events and global-shape signals, has shown that deep M-SC architecture give added value to segment all this shapes in contrast with related published algorithms (direct clustering approach or hierarchical approach). For further details, in direct perspective of this work, a silhouette score threshold for each state could be used to set the number of levels in the M-SC approach.

Then, the results obtained by M-SC on real case MAREL Carnot time series are promising. Punctual events ($\leq$ 1-month) have been discovered by unsupervised clustering in a large, complex and high resolution multivariate time series (more than 5 years with a 20 minutes sampling frequency). Moreover M-SC detected temporally structured successive

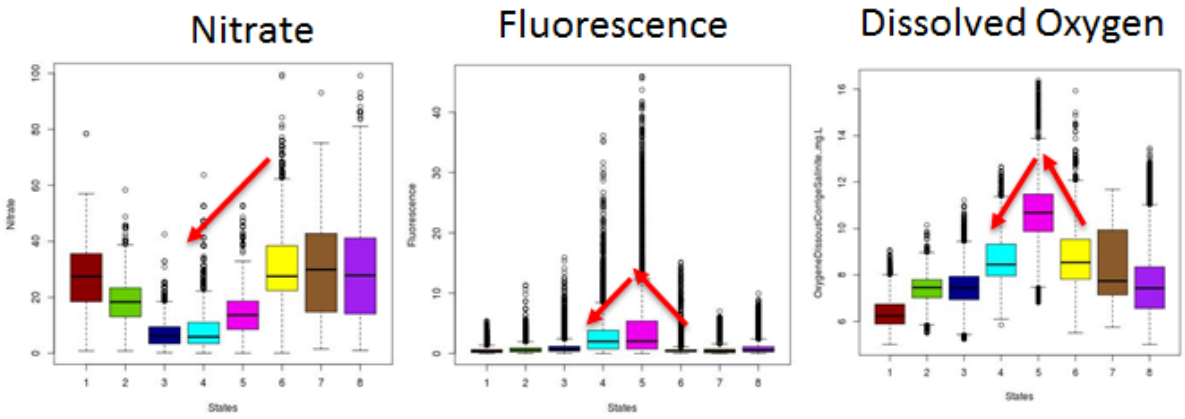


Fig. 5: box-and-whisker plot for each state : nitrate (µ), fluorescence (FFU) and dissolved oxygen () for each state over the period 2005-2008

events (pressure and impact events) while clustering processing was applied without any temporal information. Thus, we can extract knowledge on dynamics of events/environmental states. M-SC permits an unsupervised labeling of time series, basic element of machine learning, methods crucial to build an event forecasting system and to develop near real-time sampling strategies.

ACKNOWLEDGMENT

Kelly Grassi's PhD is funded by WeatherForce as part of its R & D program "Building an Initial State of the Atmosphere by Unconventional Data Aggregation".

MAREL-Carnot is part of the COAST-HF (Coastal ocean observing system - High frequency) network from the IR I-LICO (Infrastructure de Recherche LIttorale et COtière).

This project has received funding by JERICO-next from the European Union's Horizon 2020 research and innovation program under grant agreement No 654410.

This work has been also partly financially supported by the European Union (ERDF), the French State, the French Region Hauts-de-France and Ifremer, in the framework of the project CPER MARCO 2015-2020.

REFERENCES

[1] MAREL Carnot data and metadata from Coriolis Data Centre. SEANOE. http://doi.org/10.17882/39754, 2015.

[2] M. Van Hoan, D.T. Huy, and Mai L.C. Pattern discovery in the financial time series based on local trend. *ICTA 2016. Advances in Intelligent Systems and Computing. Springer*, 538, 2017.

[3] M. Längkvist, L. Karlsson, and A. Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11 – 24, 2014.

[4] É. Poisson Caillault and A. Lefebvre. Towards Chl-a bloom understanding by EM-based unsupervised event detection. In *OCEANS 2017 - Aberdeen*, pages 1–5, June 2017.

[5] R. Emonet, J. Varadarajan, and J-M. Odobez. Temporal analysis of motif mixtures using dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):140–156, 2014.

[6] J.G. Dias, J.K. Vermunt, and S. Ramos. Clustering financial time series: New insights from an extended hidden markov model. *European Journal of Operational Research*, 243(3):852 – 864, 2015.

[7] Prodromos E. Tsinaslanidis and Dimitris Kugiumtzis. A prediction scheme using perceptually important points and dynamic time warping. *Expert Systems with Applications*, 41(15):6848 – 6860, 2014.

[8] K. Rousseeuw et al. Hybrid hidden markov model for marine environment monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(1):204–213, Jan 2015.

[9] M. Ng, A.and Jordan and Y. Weiss. On Spectral Clustering: Analysis and an algorithm. pages 849–856. MIT Press, 2001.

[10] J. Sanchez-Garcia, M. Fennelly, and S. Norris et al. Hierarchical spectral clustering of power grids. *IEEE Transactions on Power Systems*, 29(5):2229–2237, Sept 2014.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[12] Patricia Miloslavich, Nicholas J. Bax, Samantha E. Simmons, Eduardo Klein, Ward Appeltans, Octavio AburtoOropeza, Melissa Andersen Garcia, Sonia D. Batten, Lisandro BenedettiCecchi, David M. Checkley, Sanae Chiba, J. Emmett Duffy, Daniel C. Dunn, Albert Fischer, John Gunn, Raphael Kudela, Francis Marsac, Frank E. MullerKarger, David Obura, and Yunne-Jai Shin. Essential ocean variables for global sustained observations of biodiversity and ecosystem changes. *Global Change Biology*, 24(6):2416–2433, June 2018.

[13] Magdalena. Strengthening Europe's Capability in Biological Ocean Observations, August 2018.

[14] Qinpei Zhao. Cluster validity in clustering methods. publications of the university of eastern finland. dissertations in forestry and natural sciences., no 77. issn: 1798-5676, 2012.

[15] Matthew J. Vavrek. fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14(1):1T, 2011. R package version 0.3.0.

[16] G. Brock, V. Pihur, Susmita Datta, and Somnath Datta. clvalid: An r package for cluster validation. *Journal of Statistical Software, Articles*, 25(4):1–22, 2008.

[17] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2018.

[18] 4 Decades of Belgian Marine monitoring: Uplifting historical data to todays needs RBINS Open Access Library.

[19] Rebecca L. Lewison, Murray A. Rudd, Wissam Al-Hayek, Claudia Baldwin, Maria Beger, Scott N. Lieske, Christian Jones, Suvaluck Satumanatpan, Chalatip Junchompoo, and Ellen Hines. How the dpsir framework can be used for structuring problems and facilitating empirical research in coastal systems. *Environmental Science Policy*, 56:110 – 119, 2016.

F.9 Article 2 : JMSE (2020)

Grassi K., Poisson-Caillault E., Bigand A., Lefebvre A., 2020. Comparative Study of Clustering approaches Applied to Spatial or Temporal Pattern Discovery. Journal of Marine Science and Engineering (ISSN 2077-1312). Special issues Applications of Machine Learning in Marine Ecology Studies. Accepted, sous presse Nov. 2020.

Journal of *Marine Science and Engineering*

MDPI

Article

Comparative Study of Clustering Approaches Applied to Spatial or Temporal Pattern Discovery

Kelly Grassi [1,2,3], Émilie Poisson-Caillault [2], André Bigand [2] and Alain Lefebvre [3,*]

1 WeatherForce, 31000 Toulouse, France; kelly.grassi@ifremer.fr
2 LISIC EA 4491 ULCO/University Littoral, 62228 Calais, France; emilie.poisson@univ-littoral.fr (É.P.-C.); andre.bigand@univ-littoral.fr (A.B.)
3 IFREMER Unité Littoral LER-BL, 62321 Boulogne-sur-Mer, France
* Correspondence: alain.lefebvre@ifremer.fr

Received: 7 August 2020; Accepted: 10 September 2020; Published: 15 September 2020

check for updates

Journal of
Marine Science and Engineering

MDPI

Article

Comparative Study of Clustering Approaches Applied to Spatial or Temporal Pattern Discovery

Kelly Grassi [1,2,3], Émilie Poisson-Caillault [2], André Bigand [2] and Alain Lefebvre [3,*]

1 WeatherForce, 31000 Toulouse, France; kelly.grassi@ifremer.fr
2 LISIC EA 4491 ULCO/University Littoral, 62228 Calais, France; emilie.poisson@univ-littoral.fr (É.P.-C.); andre.bigand@univ-littoral.fr (A.B.)
3 IFREMER Unité Littoral LER-BL, 62321 Boulogne-sur-Mer, France
* Correspondence: alain.lefebvre@ifremer.fr

Received: 7 August 2020; Accepted: 10 September 2020; Published: 15 September 2020

Abstract: In the framework of ecological or environmental assessments and management, detection, characterization and forecasting of the dynamics of environmental states are of paramount importance. These states should reflect general patterns of change, recurrent or occasional events, long-lasting or short or extreme events which contribute to explain the structure and the function of the ecosystem. To identify such states, many scientific consortiums promote the implementation of Integrated Observing Systems which generate increasing amount of complex multivariate/multisource/multiscale datasets. Extracting the most relevant ecological information from such complex datasets requires the implementation of Machine Learning-based processing tools. In this context, we proposed a divisive spectral clustering architecture—the Multi-level Spectral Clustering (M-SC) which is, in this paper, extended with a no-cut criteria. This method is developed to perform detection events for data with a complex shape and high local connexity. While the M-SC method was firstly developed and implemented for a given specific case study, we proposed here to compare our new M-SC method with several existing direct and hierarchical clustering approaches. The clustering performance is assessed from different datasets with hard shapes to segment. Spectral methods are most efficient discovering all spatial patterns. For the segmentation of time series, hierarchical methods better isolated event patterns. The new M-SC algorithm, which combines hierarchical and spectral approaches, give promise results in the segmentation of both spatial UCI databases and marine time series compared to other approaches. The ability of our M-SC method to deal with many kinds of datasets allows a large comparability of results if applies within a broad Integrated Observing Systems. Beyond scientific knowledge improvements, this comparability is crucial for decision-making about environmental management.

Keywords: clustering; pattern discovery; time series; Multi-Level Spectral Clustering; English Channel

1. Introduction

Detection of environmental states in spatial and temporal data is a fundamental task of marine ecology. It is crucial for many applications, especially to facilitate understanding of ecosystem dynamics (or part of ecosystem—i.e., phytoplancton—dynamics and more specifically Harmful Algal Blooms) and above all their vulnerability when considering anthropogenic impacts vs climatic changes at regional vs. global scales. It is also important for the evaluation of ecosystem health in order to put in place well-suited monitoring strategies for adaptive management. Although the global functioning scheme (from seasonal to monthly, from regional to spatial) are well known thanks to the implementation of conventional, historic methodologies, the small-scale variations are less studied because technologies and methodologies that are related to these issues are more recent or still in

development. Therefore, many monitoring project initiatives promote the development and the implementation of Integrated Observing Systems that produce complex databases. In the framework of water quality assessment and management, many marine instrumented stations, fixed buoys and Ferrybox were implemented with High Frequency multi-sensor systems to monitor environmental dynamics. The identification of environmental states in data sets from these instrumented systems is a hard task. The monitoring stations produce huge multivariate and complex time series with high local connexity between events, and the Ferrybox add a spatial dimension. In addition, environmental states could have a large variety of distributions, shapes and durations. Their dynamics can be an arrangement of general patterns and/or extreme events, that is, events that have a strong amplitude and/or short duration, and deviate from the general one.

1.1. General Issues

The key to obtaining correct detection, especially for extreme events, lies in applying the right numerical methodology. It is essential to optimise the processing step in order to extract relevant information. Indeed, current technologies generally allow for the introduction of extreme events, they are not commonly incorporated into new patterns of ecosystem functioning. Often, they are independently studied and not included in an integrated observing approach. Moreover, neither no open marine databases are labelled at a fine scale (time and/or space), nor available to build an efficient predictive model to forecast these events. So, we investigate a way to detect, segment and label spatial and/or temporal environmental states without any *a priori* knowledge about the number of states, their label, shape or distribution. This unsupervised labelling should provide an optimal set of clusters to facilitate identification by a human expert. We considered a cluster set optimal for interpretation of a phenomenon if it covers all the existing structures within the data from dense to sparse, frequent to rare. This labelled set is a crucial step to define an initial training database to build a prediction system by machine learning techniques.

1.2. The State-of-the-Art from an Algorithmic Point of View

Existing techniques to detect events without any *a priori* knowledge (unsupervised approach) can be divided according to the method they process cuts. Firstly, the segmentation is based on window processing (time window or spatial region). Unsupervised approaches extract similar pieces of the series from fixed-length windows [1] or sliding windows by autocencoders [2]. Secondly, segmentation by generative models such as Dirichlet Processes [3] or Hidden Markov Models [4], assumes that data are composed to mixture of models. These, two approaches need hypotheses about data distribution and pattern size which in our case it is unknown. Thirdly, the segmentation could be done by either temporal/geometric cuts such as breaking points—PIP-cut (Perception Interest Points) [5] or EVT (Extreme Values Theory). PIP approach is based on a distance criteria combining time and level values. Then, they perform a clustering process to group similar events. It is a well-adapted method to detect trend, such as financial prediction, but is not relevant without an aggregation of cut segments to identify similar events. The EVT approach based on density function and frequency is judicious when the time series follows some model/distribution, that is, rainfall estimation [6], where a clear threshold can be defined to discriminate events, that is, wet vs dry seasons or years. These methods cannot be directly used in our context: indeed, considering the high inter-annual variability of fluorescence (a proxy of phytoplankton biomass) measured in the English Channel, low values could still be representative of phytoplankton bloom of lower magnitude and duration compared in some years. The final method is to directly apply clustering without using any temporal cut/window hypotheses and in steal consider the collected multivariate points. Many clustering methods can be applied and they are often used for image segmentation problems [7]. The direct K-means (KM) and hierarchical clustering (HC) methods are the most common and are used for many applications. Density-based spatial clustering (DBSCAN) and its hierarchical version have a shape detection ability and are mainly applied for image segmentation problems. More recently, a Spectral Clustering (SC) approach has been proposed to

detect environmental states from physicochemical series. There are many variants of this algorithm. One variant to Ng et al. 2001 algorithm's (SC-Kmeans) [8] consists of extracting spectral eigenvectors of a normalised Laplacian matrix derived from the data similarity matrix followed by unsupervised k-means clustering (K-means). K-means step could be replaced by other partitioning algorithms. For instance, K-medoids algorithm, also called Partition Around Medoids (PAM) is preferred for large databases. Thus the spectral variant is named SC-PAM. Next, Shi and Malik [9] expressed the SC as a recursive graph bi-partitioning problem (algorithm Bi-SC). In Reference [10] a Hierarchical Spectral Clustering (H-SC) view is derived by replacing the initial k-means by a HC step for a specific case study.

1.3. Main Contributions

To address the issue of multi-scale and complex shape databases analyses, we proposed in Reference [11] an initial version of Multi-level Spectral Clustering (M-SC), combining spectral and hierarchical approaches with a multi-scale view. Contrary to usual spectral clustering, its deep architecture allows multi-scale and integrative approaches. It allows provides both the advantages of direct spectral clustering and also the possibility to use several levels of analysis from the general ones to more specific or deep ones. That is the reason why M-SC is said to have a deep architecture. Based on a specific temporal marine application in the English Eastern Chanel, the results of Reference [11], demonstrated the effectiveness of the M-SC method for the detection of environmental conditions and a good ability to detect extreme events. However, the method was not optimal and over-segmented the cluster with global distributions and leads a confusion with extreme events. In this article, we proposed to limit confusion and to improve the detection of extreme events with an extended version of Reference [11] including a supplementary «no-cut criterion». The «no-cut criterion» is based on density and connexity indexes and avoids cutting already well-structured clusters. Contrary to a bottom-up hierarchical approach, which segments the data into single isolated observations, this new version should enable stopping the segmentation for a given optimised level avoiding over-segmentation.

1.4. Main Objectives and Paper Organization

In this paper, we propose to confront our new adapted M-SC to several other clustering methods and also considering contrasted data sets (from the simplest one of the more complex one in terms of data geometry), to evaluate capacity to be used in various domains. The objective is to propose advice to the scientific community on how to choose the best suited unsupervised clustering method to detect global and extreme events, when processing time-series and spatial datasets with non-linear data with complex shapes and a high local connexity.

So, Section 2 introduces briefly several unsupervised clustering methods, and then focuses on our new adapted M-SC. Section 3 describes experiment protocol to compare them in the task of pattern discovery and time series segmentation. Results on artificial and in-situ data are discussed in Section 5 with a hydrographical and biological dataset during an eastern English Channel cruise. Their ability to facilitate labelling task data is afterwards discussed with a focus on some supervised approaches in Section 6.

2. Clustering Approaches for Pattern Discovery

Many clustering approaches succeed in pattern segmentation in numerous applications such as isolating objects in picture backgrounds [9] or spesific environmental events in marine multivariate time series [12]. In this tyoe of time series segmentation, the temporal information is not included in the clustering process. These approaches can be distinguished according to how they process and distribution cuts—direct or hierarchical approaches, raw space from the data, or kernel or spectral space. The space choice refers to data geometry. So, we propose viewpoint of direct and hierarchical methods and a new adapted M-SC.

2.1. Related Clustering Approaches

Direct clustering. K-meansor K-Partitioning Around Medoids (PAM) algorithms are well employed to partition convex clusters with no overlap or for vector quantisation (data reduction). They aim to partition N observations into a fixed number K of clusters, by minimising the variance within each cluster. Density-Based Spatial Clustering (DBSCAN) approaches [13] allow for relaxing the convexity constraint fot dense clusters: (1) two points are agglomerated in the same cluster if they respect a ϵ-distance and (2) the obtained clusters are saved if they have a minimal number of points (minPts). DBSCAN is useful for isolating noises; some observations will not be clustered, and it can be a default for sparse clusters. Spectral clustering (SC) techniques are used to separate clusters with low density. Require a point-to-point connexity within a cluster. SC is solved through a generalised problem of eigenvalues from a Laplacian matrix L. L is computed from a similarity matrix W obtained from the data and a cut criterion. The clustering step is done in this spectral space from the K-first eigenvectors. There are many variants like spectral k-means (SC-KM), which uses a standardised symmetric Laplacian matrix ($L_{NJW} = D^{-1/2}WD^{1/2}$; D the degree matrix of W) and a K-means algorithm for partitioning [8] or spectral PAM (SC-PAM) that uses K-medoids algorithms.

Hierachical clustering. Conventional hierarchical clustering (HC) techniques are based on the proximity between observations in the initial space. For the divisive ones, each observation is first assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters to from a single cluster. The partitioning trees differ by them proximity criterion; Ward.D2 is the most similar criterion to SC-KM [8]. An equivalent spectral approach was proposed by Reference [10], named Hierarchical-SC (H-SC), where the clustering step is based on HC with WARD.d2 criteria in L_{NJW} eigenspace. Bipartite-SC (Bi-SC) [9] leads to a binary tree: at each level, each node is subdivided in $K = 2$ clusters according to the sign of the second eigenvector from the Laplacian $L_{Shi} = I - D^{-1/2}WD^{1/2} = I - L_{NJW}$. This constraint of separation in 2 groups is well adapted when there is a dominant structure (like a background in an image). HDBSCAN is a hierarchical extension of the DBSCAN algorithm where a dissimilarity based on the ϵ-neighbourhood is used to aggregate observations.

For time-ordered observations, change-point analysis is also a possibility. We retain only these approaches with clustering—Divisive estimation (e.divisive) and agglomerative estimation (e.agglo), which are also hierarchical approaches based on (e=)energy distance [14]. e.divisive defines segments through a binary bisection method and a permutation test. e.agglo creates homogeneous clusters based on an initial clustering. If no initial clustering is defined as such, each observation is assigned to its own segment.

2.2. Proposed M-SC Variant

Multi-level spectral clustering. Our M-SC algorithm is a divisive spectral clustering approach use to build a multilevel implicit segmentation of a multivariate dataset [11]. The first level is a unique cluster with all data. At each level, observations from a related cluster are cut by SC-PAM with K computed from the maximal spectral eigengap. The spectral-PAM algorithm is detailed in Reference [11]. Here, we add a no-cut criterion ($sil.min$) for homogeneous clusters according to the silhouette index. (Algorithm 1, (Figure 1)).

The iterative segmentation of a cluster stops by not-cut criterion when it is well isolated from other clusters and has a good internal cohesion. Indeed, a cluster can be isolated at the first level. However, in the first version of the algorithm, it was systematically subdivided into deeper levels. The deeper levels thus allowed identifying clusters with extreme event characteristics but they over-segmented the more generic clusters.

This criterion is based on the Silhouette Index. Thus, the value of the silhouette of each cluster defines the conditions for stopping the segmentation. If this value is high, that is, higher than $sil.min$ it means that our cluster is sufficiently related (i.e., point are close enough to be part of the same set)

and thus sufficiently representative. In this case, the segmentation will be stopped for this cluster. Conversely, the segmentation will continue if the Silhouette of the cluster is less than $sil.min$.

$sil.min$ should be tuned for the application resolution needs, knowing that the closer the Silhouette Index is to 1, the more the cluster is well-formed and well isolated. For all experiments in this paper, the stop criterion $sil.min$ is set at 0.7; this value is a good compromise between having well-formed cluster and extreme pattern/events.

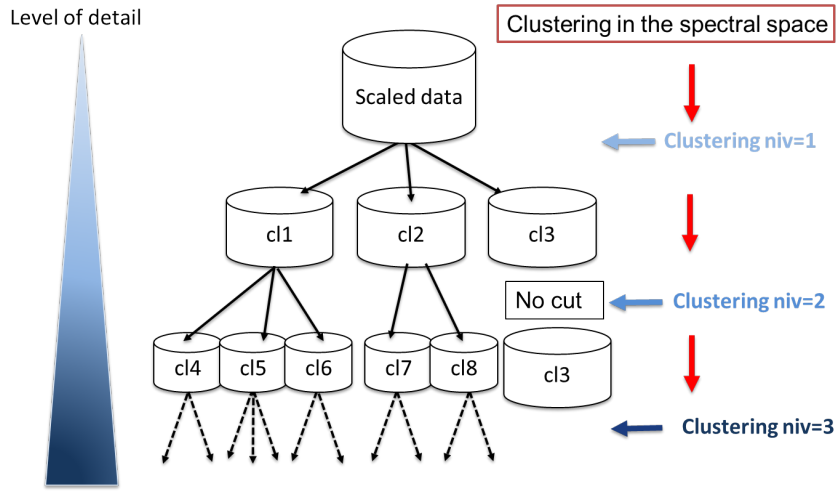


Figure 1. Multi-level Spectral Clustering (M-SC) extended with no-cut criteria scheme.

Algorithm 1 Multi-Level Spectral Clustering with no-cut stop criterion

Require: X, NivMax, Kmax, sil.min, W definition of a Gram matrix
Variables : W, cl, sil, clusterToCut = 1, niv = 1, stop = false, groups, k
 # Initialisation
cl[n, niv] = 1 matrix of $N \times NivMax$
 # Clusterings by level
while (stop! = false) **do**
for k in clusterToCut **do**
Compute similarity W of $X' = \{x_n \in X | \text{cl[n,niv]} = k\}$
groups = Spectral-PAM(W, Kmax = card(X'))
$\forall n \mid$ cl[n, niv] = k, cl[n, niv + 1] = groups+card(clusterToCut) + 1
end for
 # computation of silhouette of the new sub-clusters
clusterToCut = {} # empty vector
for k $\in$ unique(cl[n, niv + 1]) with n $\in$ [1,...,N] **do**
$X' = \{x_n | cl(n, niv + 1) = k\}$
sil = silhouette(X') # means of point silhouette
if sil<sil.min **then**
clusterToCut = {clusterToCut,k} # insertion
end if
end for
stop = ((niv + 1)$\geq$NivMax) | (card(clusterToCut) == 0)
niv = niv + 1
end while
return *cl* matrix of $N \times$ niv

3. Comparison Protocol

This work aims to comparing the ability of the above mentioned methods to propose an effective clustering as a first labelling. This section details labelled datasets, the tuning of each method should this step be required, and the list of performance metrics.

Dataset summary. For pattern discovery and time series segmentation, both selected artificial and experimental cases are briefly described in (Table 1). From UCI benchmark [15], two artificial datasets ("Aggregate" and "Compound") and two experimental ones ("Iris and Species") are chosen for their geometric characteristics. "Aggregate" has relative simple patches and "Compound" has nested patches, which are both clearly separated. They have respectively six and seven classes and both have two attributes. "Iris" and "Species" have more connected classes. Iris is a simple case because it only has three categories of plants with 50 observations per class, whereas Species has 100 classes with 16 observations per class.

For time series segmentation, the "Simulated" dataset was built and an experimental dataset from a cruise campaign was used. Simulated is composed of 3 signals based on 3 sinus global-shapes (gs) on which three short events have been inserted: two peaks and one offset (described in Reference [11]). For the experimental dataset provided by DYPHYMA program [16,17], we used the Pocket Ferry Box data (PFB), coupled with the four algae concentrations from a multiple-fixed-wavelength spectral fluorometer (Algae Online Analyser [AOA], bbe Moldaenke). The aim of this last dataset is to identify contrasting water masses based on their abiotic and biotic characteristics (details in Reference [17]). Each dataset is of dimension N observations $\times$ D features. The features are only the explicative variables like environmental parameters for "DYPHYMA" dataset. Time and spatial dimension are not included in clustering step.

Table 1. Dataset characteristics:name, area:exp. = experimental and art. = artificial, dimension (N observations $\times$ D features), number of classes C, distribution: the distribution percentage of the smallest class, (E) if equal. In bold: Time Series dataset.

	Dataset	Area	Dimension	C	Distribution
1	Aggregate [15]	art.	788×2	7	4.31
2	Compound [15]	art.	399×2	6	4
3	Iris [15]	exp.	150×4	3	33.33 (E)
4	Species [15]	exp.	1600×64	100	1.6 (E)
5	**Simulated** [11]	art.	1000×3	4	3.2
6	**DYMAPHY Leg1** [16]	exp.	2032×18	3	12.20
7	**DYMAPHY Leg2** [16]	exp.	3285×18	3	11.96
8	**DYMAPHY Leg3** [16]	exp.	5599×18	3	7.30

Data processing and parameter tuning. Firstly, all dataset X have normalised to avoid the impact of varying feature ranges in the clustering process. Zelnick and Perona locally adapted gaussian kernel with the 7th neighbours sigma distance in the similarity. For M-SC, min.points is defined at 7. Direct spectral methods and M-SC and H-SC are based on L_{NJW} Laplacian. So, all functions are computed with their default setting. But some parameters must be defined to choose the number of clusters (K). K is fixed to the ground-truth class number of direct approaches, so $K = C$. Tree cut in the hierarchical methods (HC, H-SC) and level of divisive methods (Bi-SC, M-SC) are defined to obtain at least C clusters and, sil.min is fixed at 0.7. For DBSCAN the determination of $K = C$ clusters requires ϵ-neighbourhood. It is automatically determined by Unit Invariant Knee (UIK) estimation of the average k-nearest neighbour distance. For species dataset, the 20 principal components are retained to obtain the cumulative sum of 95% explained variance.

Comparison metrics. To assess the performance of each algorithm, two indicators are computed: the total accuracy and #Iso. Then three conventional unsupervised scores are added for interpretation: Adjusted Rand index, Dunn index and Silhouette score [18].

Total accuracy is the percentage of well-recognised labels according to a given ground truth. It is a ratio of correctly classified observations (true positive) to the total observations. It is defined from the confusion table between the K clusters and C classes after a majority vote. The majority vote principle is applied to assign each cluster K to a class C according to the one that is the most represented.

#Iso is the number of well-isolated patterns, that is, represented by more than half of the true positive observations. This is the number of clusters with greater accuracy than 50 percents.

The Adjusted Rand index (ARI) is the corrected-for-chance version of the Rand Index (RI) available in fossil::adjustedRandIndex [19]. ARI measures the number of agreements in two partitions (same grouping elements and same separating pairs) over the total number of pairs. Rand Index will never actually be zero. It can yield negative values if the index is less than the expected index.

The Dunn index, clValid:dunn [20] is a ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn index has a value between zero and infinity, and should be maximised. It reflects the data/cluster separability. A Dunn index close to zero shows very connected data, while a large score indicates easily separable data.

The Silhouette score, cluster:silhouette [21] is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A large Silhouette (almost 1) is very well clustered, while a small Silhouette (around 0) means that the observations are connected between two clusters.

All indexes were computed in the raw space whatever the clustering methods used. Low Dunn index and averaged silhouette score from true label show the complexity to isolate each class, especially for DYOHYMA sets, Dunn index computed from the ground-truth labels are around 10^{-4}.

4. Clustering Results

Table 2 summarizes the clustering methods that succeed in isolating at least 50% of ground-truth patterns. They are ordered according to first #Iso, the total accucary and then minimum K to reduce the human labelling task.

Table 2. Clustering approaches applied to pattern discovery, ordered by well-isolated pattern numbers (#Iso) with performance indicators: Adjusted Rand Index (ARI), Dunn and Silhouette (Sil.) indexes, total accuracy (Tot.acc) and the number of clusters K. Bold: #Iso $= C$. 0.00: non zero number (value with more 3 decimal). n is the dataset number. Only methods that succeed in #Iso $> 50\%$ were shown.

n	Clustering	K	ARI	Dunn	Sil.	Tot.acc	#Iso
1	ground-truth	$C = 7$	1.00	0.04	0.49	1.00	7
	H-SC	7	0.99	0.04	0.49	**1.00**	**7**
	M-SC	9	0.89	0.03	0.42	**1.00**	**7**
	SC-PAM	7	0.97	0.03	0.50	0.98	**7**
	SC-KM	7	0.96	0.03	0.50	0.98	**7**
	Bi-SC	8	0.88	0.02	0.42	0.96	6
	HC Ward.d2	7	0.80	0.04	0.45	0.95	6
	KM	7	0.73	0.04	0.49	0.90	5
	DBSCAN	5	0.81	0.11	0.41	0.83	5
	HDBSCAN	5	0.81	0.11	0.41	0.83	5
2	ground-truth	$C = 6$	1.00	0.07	0.16	1.00	6
	M-SC	6	1.00	0.07	0.16	**1.00**	**6**
	KM	6	0.56	0.02	0.35	0.85	5
	Bi-SC	8	0.62	0.03	0.26	0.81	4
	SC-KM	6	0.45	0.03	0.29	0.74	4

Table 2. *Cont.*

n	Clustering	K	ARI	Dunn	Sil.	Tot.acc	#Iso
3	ground-truth	$C = 3$	1.00	0.06	0.50	1.00	3
	M-SC	3	1.00	0.06	0.50	**1.00**	**3**
	Bi-SC	8	0.72	0.06	0.27	**1.00**	**3**
	KM	3	0.62	0.04	0.51	0.83	**3**
	HC Ward.d2	3	0.61	0.07	0.50	0.83	**3**
	H-SC	3	0.45	0.05	0.53	0.67	2
	SC-KM	3	0.45	0.03	0.53	0.67	2
	SC-PAM	3	0.45	0.03	0.53	0.67	2
4	ground-truth	$C = 100$	1.00	0.11	0.10	1.00	100
	SC-KM	100	0.48	0.14	0.07	0.65	75
	SC-PAM	100	0.46	0.09	0.07	0.64	73
	KM	100	0.45	0.12	0.09	0.63	72
	H-SC	100	0.46	0.14	0.08	0.64	71
	HC Ward.d2	100	0.45	0.16	0.11	0.64	70
	M-SC	115	0.31	0.16	0.01	0.50	55
5	ground-truth	$C = 4$	1.00	0.01	0.16	1.00	4
	e.divisive	23	0.39	0.00	0.03	0.97	**4**
	M-SC	8	0.43	0.007	0.28	0.94	**4**
	HDBSCAN	5	0.62	0.01	0.13	0.94	3
	e.agglo	9	0.45	0.00	−0.03	0.94	3
6	ground-truth	$C = 3$	1.00	0.00	−0.02	1.00	3
	M-SC	32	0.66	0.00	−0.21	0.94	**3**
	HDBSCAN	42	0.68	0.00	−0.10	0.91	**3**
	e.divisive	42	0.49	0.00	−0.22	0.96	**3**
	e.agglo	10	0.48	0.00	−0.14	0.79	**3**
	KM	3	0.57	0.00	−0.04	0.84	2
	HC Ward.d2	3	0.57	0.00	−0.03	0.84	2
	SC-KM	3	0.53	0.00	−0.01	0.84	2
	SC-PAM	3	0.53	0.00	−0.01	0.84	2
	H-SC	3	0.53	0.00	−0.01	0.84	2
7	ground-truth	$C = 3$	1.00	0.00	−0.03	1.00	3
	M-SC	53	0.51	0.00	−0.18	0.94	**3**
	e.divisive	55	0.55	0.00	−0.17	0.92	**3**
	HDBSCAN	62	0.48	0.00	−0.25	0.89	**3**
	HC Ward.d2	3	0.21	0.00	0.26	0.72	2
	KM	3	0.21	0.00	0.25	0.72	2
	e.agglo	3	0.20	0.00	0.28	0.72	2
	SC-PAM	3	0.11	0.01	0.12	0.64	2
	SC-KM	3	0.06	0.01	0.20	0.64	2
8	ground-truth	$C = 3$	1.00	0.00	−0.02	1.00	3
	HC ward.d2	3	0.23	0.00	−0.00	0.80	2
	KM	3	0.21	0.00	0.01	0.79	2
	M-SC	4	0.41	0.00	−0.15	0.79	2

Spectral methods succeed in discovering all spatial patterns (#Iso = C) with a high score for hierarchical approaches: they could achieve 100% of accuracy, particularly M-SC except for "species". For example, within the "coumpound" dataset, M-SC is able to detect nested or overlaped clusters (clusters cyan and pink), while no other method makes the separation (Figure 2). However, within the "species" dataset, methods including M-SC have low scores and only part of the C classes are detected It could be explained by the low-class distribution and highly connected clusters (averaged silhouette = 0.1). For the time series segmentation task, hierarchical methods better isolated event patterns, particularly M-SC, e.divisive and HDBSCAN. For "DYPHYMA-leg3", none of the algorithms isolated 3 classes. M-SC succeeded in isolating them at level 3 with $K = 102$ and a total

accuracy of 93%. This number of clusters to label could be too many and unreasonable for the human expert, but we considered that who can do more can do less. Who can do more can do less.The objective is to offer the expert the possibility to define its own level of relevant expertise based on the best available number of clusters.

Multi-Level Spectral clustering seems to be effective to detect pattern structure in spatial data or time series (Figures 2 and 3). The obtained results reveal a good ability for generalisations. However, the no-cut criterion is a sensitive parameter and is not self-sufficient when the connection is too important, like the "species" or " DYPHYMA-leg3" datasets. Moreover, the deepest level permit to detect a large number of events or patterns but leads to over-segmentation. Human experts should tune M-SC silhouette parameters and levels of clustering according to a compromise between over-segmentation and cluster number for their labelling task. For large databases, the M-SC algorithm could be easily modified to obtain a fast computation process by using a reduced prototype set and an n-nearest neighbour algorithm.

The efficiency of a given method depends on the data distribution hypothesis and capacity to include it. For simple examples, that is, linear or with low connexity data, methods such as KM or HC are relevant. However in our application these direct methods (KM or HC) are less effective for our application because they require non-convexe shapes and linearly separable clusters to be optimal. Also, segmentation by breaking points (e.divisive and e.agglo) are performing when signals have a break between 2 stationary regimes. These breaks are not so obvious in marine data where extreme events are not an intensive variation with the overall signal in mean or variance. DBSCAN and HDBSCAN consider this extreme event often as outliers.

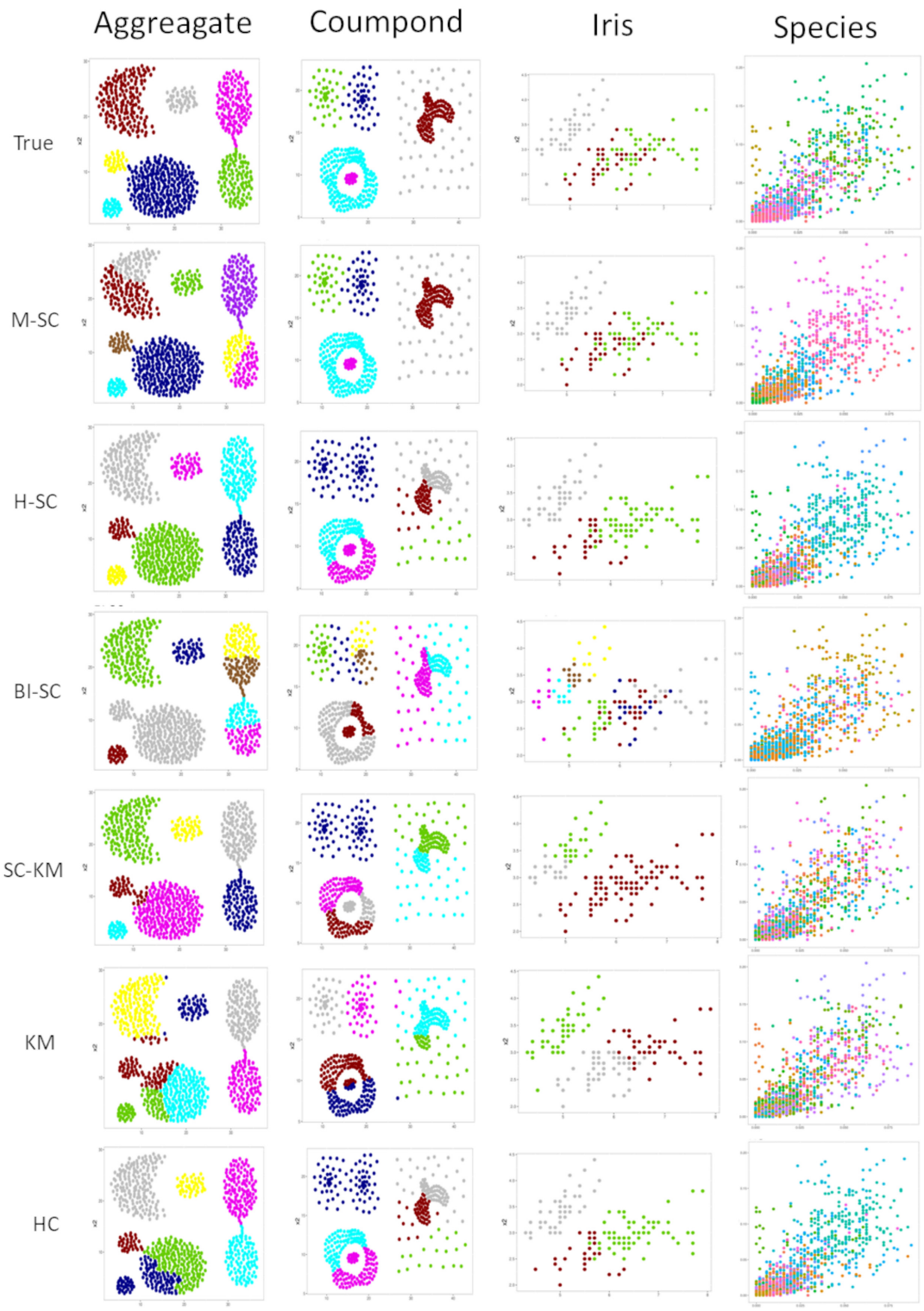


Figure 2. Color-labelling for the most efficient clustering methods on spatial dataset (True: ground-truth labels, M-SC Multi level Spectral Clustering, H-SC: Hierarchical clustering, Bi-SC; recursive bipartite SC, SC-KM: SC-Kmeans, KM: Kmeans, HC: Hierarchical clustering).

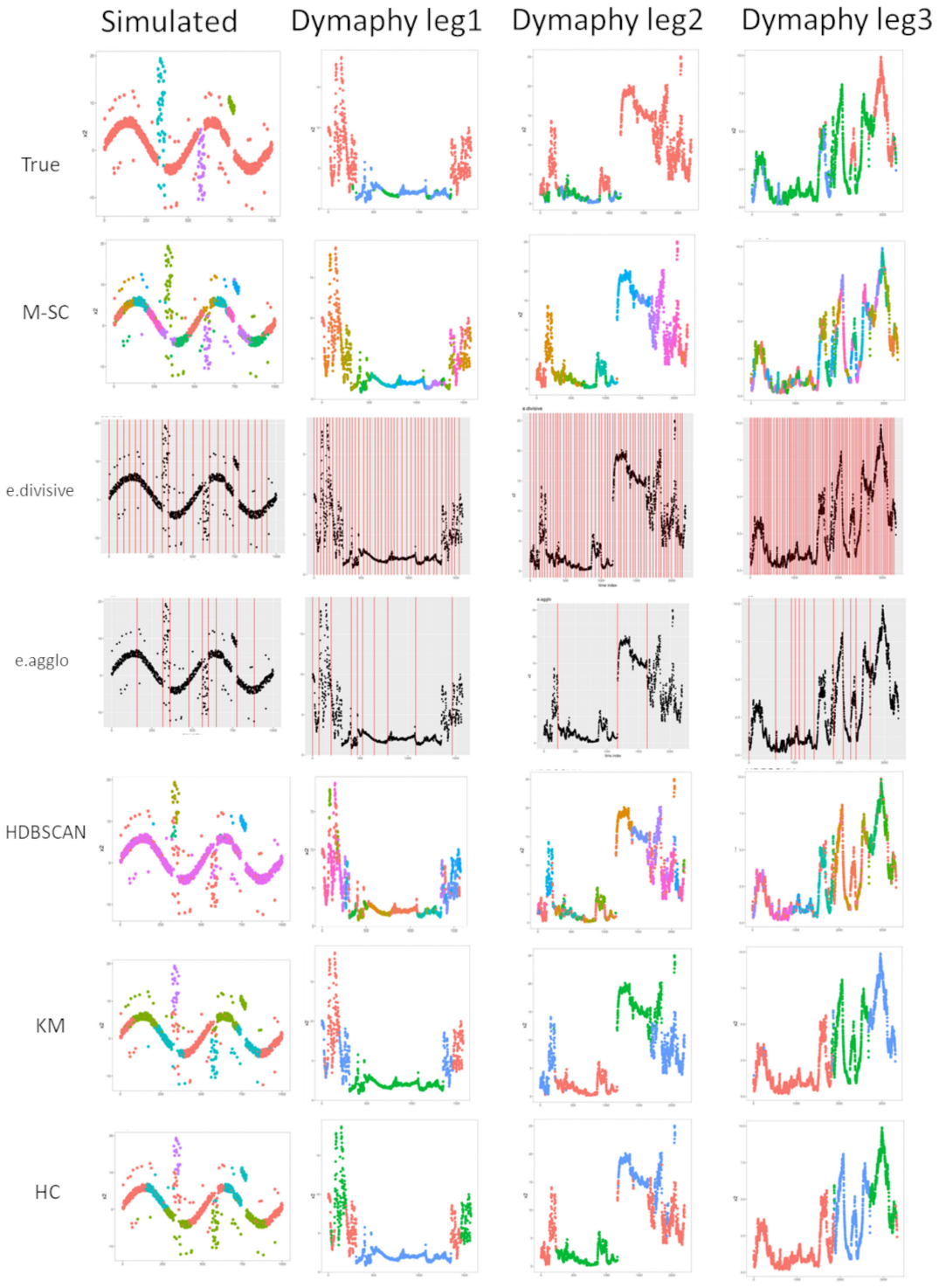


Figure 3. Color-labelling for the most efficient clustering methods on time series (True: ground-truth labels, M-SC Multivel Spectral Clustering, HDBSCAN: Hierarchical density-based spatial clustering, KM: Kmeans, HC: Hierarchical clustering).

5. Clustering for Labelling

Clustering methods rely solely on data geometry to provide data segmentation. The purpose of this last section is to test their ability to provide a first labelling. So we compare them now with supervised techniques in the task of pattern discovery in time series.

Three basic machine solutions were explored: k-Nearest Neighbours classification (k-nn), Breiman's Random Forest algorithm (RF) [22] and a Multi-Layer perceptron (MLP). For comparison, MLP was preferred to Time Delay Neural Networks to be fair with clustering approaches that do not take account time parameters also.

Two training databases per dataset (from 5 to 8 in Table 1) werre built. The first database represents 20% of the volume of each class in the Table (80% for the test database) and the second 50%. Whatever the training and test base is, it covers all $K = C$ temporal events. The attributes of an observation in the series are the classifier entries: for the 8th dataset, the input number I is equal to 18. And C is the output number of MLP and RF. For k-nn, two k values are chosen. k is set to 1 to assign the label of the closest observation and then $k = 7$ to obtain a a more unified segmentation (ie. with less than one class singleton among observations of other classes). MLP-1 here has one hidden layer whose neuron number is equal to $(I + C)/2$ and MLP-0 corresponds to linear perceptron (no hidden layer). Random Forest here consists of the vote of 500 trees with $\left\lfloor \sqrt{I} \right\rfloor$ variables randomly sampled as candidates at each split.

The same unsupervised and supervised scores as in Table 2 are computed on the test databases and reported in Table 3 in order to compare clustering and classification approaches.

Learning techniques do not over-segment the time series due to the fixed number of their classes.

RF and k-nn are able to isolate events with higher accuracy and so lead to low overlap between them for every dataset and every training cut. MLP-0 and MLP-1 are able to identify the 4 events of the simulated case but they are not efficient for in-situ cases (n = 6–8) due to a lack of observations and unequal classes.

Table 3. Classification approaches applied to pattern discovery, ordered by well-isolated pattern numbers (#Iso) with performance indicators for test database: Adjusted Rand Index (ARI), Dunn and Silhouette (Sil.) indexes, total accuracy (Tot.acc) and the number of clusters K. Bold: #Iso $= C$. 0.00: non zero number (value with more 3 decimal). n is the dataset number. RF = Random Forest, MLP-l = Multi-Layer Perceptron with l hidden layer, k-nn = k-nearest neighbors

n	20%-Training	K = C	ARI	Dunn	Sil.	Tot.acc	#Iso
5	ground-truth	4	1.00	0.03	0.16	1.00	4
	RF	4	1.00	0.03	0.17	**1.00**	**4**
	1-nn	4	0.85	0.02	0.17	0.97	**4**
	MLP−0	4	0.90	0.02	0.17	0.97	**4**
	MLP-1	4	0.91	0.05	0.18	0.97	**4**
	7-nn	4	0.65	0.05	0.21	0.94	2
6	ground-truth	3	1.00	0.00	−0.01	1.00	3
	RF	3	0.98	0.00	−0.01	0.99	**3**
	1-nn	3	0.77	0.00	0.002	0.91	**3**
	7-nn	3	0.56	0.001	0.02	0.82	**3**
	MLP−0	3	0.21	0.00	0.20	0.59	**3**
	MLP-1	3	-	-	-	0.50	1
7	ground-truth	3	1.00	0.00	−0.04	1.00	3
	RF	3	0.98	0.00	−0.03	0.99	**3**
	1-nn	3	0.80	0.00	−0.02	0.91	**3**
	7-nn	3	0.75	0.00	−0.01	0.88	**3**
	MLP−0	3	0.58	0.002	0.24	0.74	2
	MLP-1	3	-	-	-	0.63	1

Table 3. *Cont.*

n	20%-Training	K = C	ARI	Dunn	Sil.	Tot.acc	#Iso
8	ground-truth	3	1.00	0.00	−0.02	1.00	3
	RF	3	0.96	0.00	−0.01	0.98	**3**
	1-nn	3	0.70	0.00	−0.01	0.89	**3**
	7-nn	3	0.59	0.00	0.01	0.86	**3**
	MLP−0	3	0.28	0.001	−0.02	0.78	1
	MLP-1	3	-	-	-	0.70	1
n	50%-Training	K	ARI	Dunn	Sil.	Tot.acc	#Iso
5	ground-truth	4	1.00	0.03	−0.02	1.00	4
	RF	4	1.00	0.03	0.16	**1.00**	**4**
	1-nn	4	0.93	0.03	0.14	0.99	**4**
	7-nn	4	0.87	0.07	0.16	0.98	**4**
	MLP−0	4	0.99	0.03	0.15	0.99	**4**
	MLP-1	4	0.95	0.04	0.24	0.96	3
6	ground-truth	3	1.00	0.00	−0.05	1.00	3
	RF	3	1.00	0.00	−0.02	**1.00**	**3**
	1-nn	3	0.83	0.00	−0.02	0.93	**3**
	7-nn	3	0.73	0.00	−0.02	0.90	**3**
	MLP−0	3	0.65	0.007	0.03	0.83	2
	MLP-1	3	-	-	-	0.51	1
7	ground-truth	3	1.00	0.00	−0.015	1.00	3
	RF	3	0.97	0.00	−0.05	0.99	**3**
	1-nn	3	0.84	0.00	−0.05	0.92	**3**
	7-nn	3	0.80	0.00	−0.07	0.91	**3**
	MLP−0	3	0.75	0.001	0.14	0.79	2
	MLP-1	3	-	-	-	0.63	1
8	ground-truth	3	1.00	0.00	0.16	1.00	3
	RF	3	0.98	0.00	−0.01	0.99	**3**
	1-nn	3	0.81	0.00	−0.01	0.93	**3**
	7-nn	3	0.70	0.00	0.003	0.90	**3**
	MLP−0	3	0.35	0.00	0.04	0.78	2
	MLP-1	3	-	-	-	0.69	1

Divisive clustering techniques like M-SC do not suffer from unequal classes and can better to detect events in the series. M-SC reached the same objective of well-isolated pattern number as supervised techniques like RF or k-nn. ARI and connectedness scores are highly dependent on the number of classes. In the supervised case, with a fixed K-number and computed from the test database only, ARI scores are higher than those of clustering approaches. However, the connectedness indices are not better.

This study has shown that the Multi-level Spectral Clustering approach is a promising way to assist an expert in a labelling task for both spatial data and time series. M-SC also provides a deep hierarchy of labels depending on the desired depth of interpretation.

6. Conclusions

In marine ecology, understanding and forecasting events and environmental states is crucial for many applications, so artificial intelligence systems should especially to facilitate understanding of ecosystem processes and dynamics. It is also important for evaluation of ecosystem health in order to qualify environmental status and to put adaptative strategies to reduce the anthropogenic pressure on marine ecosystems. So, integrate and multi-scale optimal approaches are therefore needed to effectively monitor complex and dynamic ecosystems. So, the correct detection of environment state in no-linear multivariate dataset required the right numerical methodology. It is essential to optimise the processing in order to extract relevant information for stakeholders. We propose, a Multilevel Spectral

Clustering (M-SC) was proposed multivariate time series into general patterns up to extreme events by unsupervised way and demonstrate this algorithm outperforms existing algorithms for this task.

In this case, we improved the processing of the spectral method by adding hierarchical and density approaches. The deep architecture allows processing without losing key information for the detection of extreme events. In fact, this hierarchy allows eliminating the strong contributions of structuring variables linked to trends in early levels and general seasonal cycles and, to observe more specific environmental states at deeper levels while preserving all the explanatory variables. Then the no-cut criterion based on density and connexity indexes facilitates heterogeneous clustering. This allows the identification of classes of different sizes and duration. This is a significant advantage when studying processes whose phenology is highly variable, as in the case of harmful algae blooms. Thus, experts can use this criterion to find a compromise between over-segmentation and the number of classes necessary for their labelling task.

These different aspects are not present or only partially present in other clustering algorithms. The tests performed on artificial and experimental datasets with high local connexity between events and global-shape signals showed that the M-SC architecture can segment several kinds of shapes with which related algorithms struggle. M-SC often offers the most efficient segmentation. These results also reveal a good result for first labelling, it is close to the supervised machine learning techniques and includes a reasonable number of clusters with coherent structures.

Therefore, M-SC multilevel implicit segmentation will enable the implementation of nested approaches and to optimise extraction of knowledge when considering data covering different scales (temporal, frequency or spatial).

The extended M-SC approach seems well adapted for segmentation of time series or spatial datasets with coherent patterns that could appear several times or once. It combines the segmentation and clustering steps in the process and suggests different scales of interpretation, including a good detection of extreme events for an integrated observing approach.

However, M-SC has several limitations. The major drawback is that it requires a complete dataset: observations with no missing values (NA). In the case of NA values, data would not be assigned to cluster and could affect the clustering step. It is important to align the data. The choice of similarity and Laplacian operator was not studied here. The choice of Similarty matrix W or Laplacian definition L could affect the results and depend on the application. Furthermore, SC computation could be difficult for large datasets and in this case it will be replaced by Fast-NJW spectral algorithm based on a clustering after sampling (by a vector quantization). Another important point is sil.min parameter tuning. sil.min could be not been strong enough for applications with events or geographical pattern composed of very few observations.

We believe that the M-SC approach could be used for other marine applications (data from Ferry Box, gliders, etc.) and also for other applications when needing to segment data series and to identify general patterns and specific events without any *a priori* knowledge. For example, (I) The method could allow the characterisation of environmental states and associated phytoplankton assemblages. This would enable ecosystem dynamics studies and a better understanding at the environmental forcing (nutrient inputs, storms, ect) on multiple spatial and temporal scales. It should be of importance for eutrophication monitoring and Harmful Algal Blooms (HABs) forecasting. (II) Applied to Ferry Box data, it could permit the detection and characterisation of eco-regions and associated phytoplankton communities. This would make it possible to propose global or local assessments and could help in decision-making within the framework of the project to establish ecological status of marine waters(DCSMM, OSPAR conventions) or within the framework of the structuring of observation networks (H2020 JERICO-S3 project). From machine learning methods, it could enable the implementation of real-time sampling strategies during sea campaigns. (III) The classification could also be used to detect sensor failures and measurement anomalies on automated measuring stations. An alert system could be set up, which would speed up maintenance operations.

Author Contributions: Conceptualization, K.G., É.P.-C. and A.L.; methodology, K.G. and É.P.-C.; software, K.G. and É.P.-C.; validation, A.L., É.P.-C. and A.B.; formal analysis, K.G.; investigation, A.L., É.P.-C. and A.B.; resources, A.L.; data curation, K.G.; writing–original draft preparation, K.G., É.P.-C., A.B. and A.L.; writing–review and editing, K.G., É.P.-C., A.B. and A.L.; supervision, A.L. and É.P.-C.; project administration, A.L. and É.P.-C.; funding acquisition, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been financially supported (1) by the European Union (ERDF), the French State, the French Region Hauts-de-France and Ifremer, in the framework of the project CPER MARCO 2015–2020 (Grant agreement: 2016_05867 et 2016_05866), and (2) by the JERICO-S3 project which is funded by the European Commission's H2020 Framework Programme under grant agreement No. 871153. Project coordinator: Ifremer, France. Kelly Grassi's PhD is funded by WeatherForce as part of its R & D program "Building an Initial State of the Atmosphere by Unconventional Data Aggregation".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Hoan, M.; Huy, D.; Mai, L.C. Pattern Discovery in the Financial Time Series Based on Local Trend. In *Advances in Information and Communication Technology*; Springer: Cham, Switzerland, 2017; Volume 538. [CrossRef]
2. Längkvist, M.; Karlsson, L.; Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.* **2014**, *42*, 11–24. [CrossRef]
3. Emonet, R.; Varadarajan, J.; Odobez, J.M. Temporal Analysis of Motif Mixtures Using Dirichlet Processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 140–156. [CrossRef] [PubMed]
4. Dias, J.; Vermunt, J.; Ramos, S. Clustering financial time series: New insights from an extended hidden Markov model. *EJOR-Eur. J. Oper. Res.* **2015**, *243*, 852–864. [CrossRef]
5. Tsinaslanidis, P.; Kugiumtzis, D. A prediction scheme using perceptually important points and dynamic time warping. *Expert Syst. Appl.* **2014**, *41*, 6848–6860. [CrossRef]
6. Kuswanto, H.; Andari, S.; Oktania Permatasari, E. Identification of Extreme Events in Climate Data from Multiple Sites. *Procedia Eng.* **2015**, *125*, 304–310. [CrossRef]
7. Sharma, P.; Suji, J. A Review on Image Segmentation with its Clustering Techniques. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2016**, *9*, 209–218.
8. Ng, A.; Jordan, M.; Weiss, Y. *On Spectral Clustering: Analysis and an Algorithm*; MIT Press: Cambridge, MA, USA, 2001; pp. 849–856.
9. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905. [CrossRef]
10. Sanchez-Garcia, J.; Fennelly, M.; Norris, S.; Wright, N.; Niblo, G.; Brodzki, J.; Bialek, J.W. Hierarchical Spectral Clustering of Power Grids. *IEEE Trans. Power Syst.* **2014**, *29*, 2229–2237. [CrossRef]
11. Grassi, K.; Caillault, E.P.; Lefebvre, A. Multi-level Spectral Clustering for extreme event characterization. In Proceedings of the MTS IEEE OCEANS 2019, Marseille, France, 17–20 June 2019.
12. Rousseeuw, K.; Poisson Caillault, É.; Lefebvre, A.; Hamad, D. Hybrid Hidden Markov Model for Marine Environment Monitoring. *IEEE JSTARS-J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 204–213. [CrossRef]
13. Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD Proc.* **1996**, *96*, 226–231
14. James, N.; Matteson, D. ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data. *J. Stat. Softw.* **2015**, *62*, 1–25. [CrossRef]
15. Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, USA, 2017.
16. DYPHYMA Dataset. Continuous Phytoplankton Measurements during Spring 2012 in Eastern Channel. 2012. Available online: https://sextant.ifremer.fr/record/5dbafe69-81cf-4202-b541-9f8b564fa6f9/ (accessed on 12 September 2020).
17. Lefebvre, A.; Poisson-Caillault, E. High resolution overview of phytoplankton spectral groups and hydrological conditions in the eastern English Channel using unsupervised clustering. *Mar. Ecol. Prog. Ser.* **2019**, *608*, 73–92. [CrossRef]
18. Zhao, Q. *Cluster Validity in Clustering Methods*; University of Eastern Finland. Dissertations in Forestry and Natural Sciences: Joensuu, Finland, 2012; no 77, ISSN 1798-5676.

19. Vavrek, M.J. Fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontol. Electron.* **2011**, *14*, 16.
20. Brock, G.; Pihur, V.; Datta, S.; Datta, S. clValid: An R Package for Cluster Validation. *J. Stat. Softw. Artic.* **2008**, *25*, 1–22. [CrossRef]
21. Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions; R Package Version 2.1.0*; Brown Walker Press: Boca Raton, FL, USA, 2018.
22. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]