Modélisation de signaux temporels hautes fréquences multicapteurs à valeurs manquantes : Application à la prédiction des efflorescences phytoplanctoniques dans les rivières et les écosystèmes marins côtiers

Kévin Rousseeuw

Centre Ifremer Manche Mer du Nord
Laboratoire Environnement et Ressources

École Doctorale Sciences pour l'Ingénieur
CGR de Calais
Numéro d'ordre :

THÈSE

Présentée et soutenue publiquement le 11 décembre 2014
pour l'obtention du grade de

Docteur de l'Université du Littoral Côte d'Opale
Discipline : Traitement du Signal
par

Kévin ROUSSEEUW

Titre :

Modélisation de signaux temporels hautes fréquences, multicapteurs à valeurs manquantes. Application à la prédiction des efflorescences phytoplanctoniques dans les rivières et les écosystèmes marins côtiers.

Composition du jury

Rapporteurs :	***Philippe Grosjean*** *Professeur, Université de Mons, Belgique* ***Ali Mansour*** *Professeur, ENSTA Bretagne*
Membre :	***Cédric Bacher*** *CR-HDR, Ifremer Brest*
Invités:	***Jean Prygiel*** *HDR, Agence de l'Eau Artois Picardie* ***François Schmitt*** *DR-CNRS, Université de Lille 1*
Encadrants :	***Alain Lefebvre*** *CR, Ifremer, Boulogne-sur-Mer* ***Émilie Poisson-Caillault*** *MCF, Université du Littoral Côte d'Opale de Calais*
Directeur de Thèse :	***Denis Hamad*** *Professeur, Université du Littoral Côte d'Opale de Calais*

« Que la force soit avec toi »
Obiwan Kenobi, StarWars IV

Remerciements

Je remercie tout d'abord Denis Hamad pour avoir accepté de diriger ce travail de thèse, ainsi que mes deux encadrants, qui m'en ont fait baver et sans qui je ne serais pas là où j'en suis : Alain Lefebvre et Emilie Poisson Caillault pour tout ce que vous m'avez apporté, pour tous les échanges que nous avons eu, tous nos fous rires et le fait que vous avez toujours cru en moi : « Maintenant c'est fait ! ».

Merci à l'ensemble de mon jury pour avoir accepté d'évaluer ce travail de thèse : Cédric Bacher, Philippe Grosjean, Ali Mansour, François Schmitt et Jean Prygiel.

Je remercie de nouveau Jean Prygiel mais également l'ensemble de son équipe de l'Agence de l'Eau Artois Picardie pour tous les échanges que nous avons eu et la confiance que vous m'avez donné.

Merci à l'ensemble des personnes du projet DYMAPHY, ce fut un plaisir de travailler avec vous. Merci aux équipes du LISIC de Calais, du LER, de RH et les pôles administratifs de Boulogne-sur-Mer pour leur bonne humeur et pour tout le temps que l'on a passé ensemble. Restez comme vous êtes.

Arnaud et Manu, mes potes, merci pour vos conseils, nos parties de rigolades et nos délires permanents : « j'annonce : les mecs, vous allez être fier de moi ! ». A ma petite Marie, merci pour tous tes conseils et ta bonne humeur : « ça y est, je ne suis plus un padawan ! ». Guillaume (« mon p'tit chameau ») et David : merci pour votre soutien.

J'aimerais pour finir remercier toute ma famille : je vous aime très fort.

Table des matières

Glossaire

Allochtone (Matière organique) : Il s'agit de la matière organique issue de la dégradation des végétaux supérieurs qui est apportée aux environnements aquatiques par les eaux de nappe et les eaux de ruissellement. Ce compartiment comprend également les apports de matière organique anthropique provenant des déjections animales (fumiers, lisier), des résidus urbains et / ou industriels (boues de station d'épuration, hydrocarbures) ou encore de l'utilisation de produits phytosanitaires.

Anthropique : relatif à l'activité humaine. Qualifie tout élément provoqué directement ou indirectement par l'action de l'homme

Autochtone (Matière organique) : Il s'agit de la matière organique produite au sein des environnements aquatiques proprement dits, soit du fait du développement et de l'activité des organismes photoeutrophes (phytoplancton, macrophytes, …), soit du fait de la dégradation de ces mêmes organismes par les organismes brouteurs et les bactéries.

Benthique : relatif au fond des mers

Biofouling : accumulation de micro-organismes, plantes, algues ou animaux sur une matière solide immergée dans un milieu aquatique

Chlorophylle-*a* : principal pigment photosynthétique qui favorise à l'intérieur de la cellule l'absorption de l'énergie lumineuse chez les végétaux

Efflorescence (bloom) : prolifération et accumulation rapide et massive d'une espèce ou d'un groupe d'espèce de micro-algues à la surface d'un milieu aquatique

Epicontinentale (mer) : mer bordière occupant le domaine de la plate-forme continentale

Fluorescence : émission lumineuse provoquée par l'excitation des pigments algaux

Gamme capteur : intervalle de valeurs dites correctes définit par le fabricant du capteur

Gamme expert : intervalle de valeurs dites correctes définit par un expert du domaine

Gyre : gigantesque tourbillon d'eau formé d'un ensemble de courants marins

Haptonéme : filament s'étendant entre deux flagelles

Mégatidale (régime, mer) : environnement ou le marnage est supérieur à 8 mètres

Marnage : variation du niveau du plan d'eau d'une voie navigable

Photique (zone) : zone aquatique comprise entre la surface et la profondeur maximale d'un lac ou d'un océan exposée à une lumière suffisante pour que la photosynthèse se produise.

Pigment algal : molécule présente à l'intérieur de la cellule végétale et capable de transformer l'énergie lumineuse en énergie directement utilisable par la cellule

Phytoplancton : micro-algue en suspension dans l'eau qui dérive avec les courants

Production primaire : processus par lequel les algues micro-planctoniques transforment la matière minérale en matière organique nécessaire pour leur croissance.

Réseau trophique : définit comme l'ensemble des relations alimentaires entre espèces au sein d'une communauté et par lesquelles l'énergie et la matière circulent.

Taxon : entité d'êtres vivants regroupés parce qu'ils possèdent des caractères en communs du fait de leur parenté, et permet ainsi de classifier le vivant à travers la systématique.

Tempérée (zone) : se dit d'une zone où les températures ne sont pas extrêmes, ni torrides, ni glaciales.

Notations

$\boldsymbol{S}$	:	Etats du modèle
$\boldsymbol{V}$	:	Symboles du modèle
N	:	Nombre d'états
M	:	Nombre de symboles
π	:	Vecteur de probabilités initiales (de dimension $N \times 1$)
$\boldsymbol{A}$	:	Matrice de transition (de dimension $N \times N$)
$\boldsymbol{B}$	:	Matrice d'émission (de dimension $N \times M$)
λ	:	Modèle de Markov Caché : $\lambda(S, V, \pi, \boldsymbol{A}, \boldsymbol{B})$
$\boldsymbol{X}$	:	Base de données
N_p	:	Nombre de lignes de la matrice, nombre d'instants
D_p	:	Nombre de colonne / dimensions de la matrice : nombre de paramètres
vE	:	Variance Expliquée
τ	:	Chemin d'états
$card(i)$	:	Cardinal de i
p-value	:	Significativité : $*\leq 0{,}1$
		$**\leq 0{,}05$
		$***\leq 0{,}01$

Introduction générale

L'océan et sa frange côtière sont des milieux complexes, en mouvement permanent, gouvernés par des mécanismes physiques, chimiques et biologiques à haute fréquence et courtes longueurs d'onde, très incomplètement observés, décrits et compris, soumis à des aléas naturels (*i.e.* changement climatique) et à des pressions anthropiques (*i.e.* usages en amont des bassins versants par l'agriculture, l'industrie) de plus en plus fortes. En particulier, les domaines littoral et côtier concentrent des usages multiples en mer et sur le littoral, où les populations se sont notablement densifiées. Une des explications de la méconnaissance de ce milieu et de son état de santé tient à la difficulté d'y effectuer des observations régulières. Les programmes d'observation et de surveillance proposent le plus souvent des fréquences de suivi bimensuelles ou mensuelles, rarement hebdomadaires. Les approches à plus hautes fréquences (HF) se faisaient et se font encore sous forme d'opérations « coup de poing » lors de courte période dans le cadre de programme de Recherche. L'acquisition de mesures automatisées en mer de manière plus systématique n'est possible que depuis quelques décennies et seulement pour quelques paramètres et dans certaines zones. Ces mesures sont pourtant indispensables d'une part pour rendre compte de la dynamique multi-échelle de l'océan et des écosystèmes qu'il abrite, d'autre part pour surveiller leur état environnemental susceptible de se dégrader sous l'effet des pressions anthropiques. A ce jour, la valorisation des données HF reste faible car les méthodes numériques couramment employées pour traiter les données issues des programmes à plus basse résolution ne permettent pas d'optimiser l'extraction du maximum d'information portée par ces séries.

La prise de conscience générale des problèmes environnementaux, notamment au niveau du littoral, a donc conduit à renforcer la surveillance qui s'y exerce. Différentes directives, nationale, européenne ou des conventions à l'échelle de zones océaniques, telles que la Directive Cadre sur l'Eau (DCE – 2000/60/CE) ou la Directive Cadre Stratégie du Milieu Marin (DCSMM – 2008/56/CE) ou encore la convention d'Oslo et de Paris (OSPAR, 2010), ont ainsi émergées et imposent d'établir l'état des milieux aquatiques de manière à qualifier les masses d'eau, les classifier et pour prévoir le cas échéant des actions de restauration. Ces évaluations sont réalisées, entre autres critères, à partir de paramètres hydrologiques et biologiques.

Les paramètres physico-chimiques classiquement mesurés dans les systèmes aquatiques permettent de définir les conditions environnementales favorables au développement de la faune et de la flore, et reflètent également les effets directs et indirects de leur développement et de leurs interactions dans le milieu.

Parmi les paramètres biologiques, le phytoplancton, constitué d'organismes microscopiques, est un élément essentiel dans le cycle des matières ainsi que pour la productivité des zones

côtières et des océans (Cloern et Jassby, 2008). Ce compartiment est donc d'une importance capitale puisqu'il constitue la base des réseaux trophiques marins. Le maintien des biens et des services écologiques est donc en partie liés à la dynamique du phytoplancton. De par sa capacité à répondre rapidement aux modifications de la qualité de son environnement, le phytoplancton est régulièrement utilisé comme un indicateur de qualité pour les directives et conventions citées ci-dessus. La biomasse du phytoplancton, son abondance et sa composition sont régulièrement utilisées dans les métriques développées. La biomasse phytoplanctonique est historiquement évaluée par la concentration en chlorophylle-*a*, principal pigment des organismes végétaux (Lorenzen, 1966). Une technique plus rapide et plus simple consiste à estimer cette biomasse par fluorimétrie, en ayant cependant conscience de certaines limites liées à la variabilité de la relation fluorescence-chlorophylle-*a*.

Les utilisateurs des données HF se retrouvent face à des quantités de données pour lesquelles les méthodes statistiques classiques sont peu ou pas utilisables pour une optimisation de l'extraction du maximum d'information de ce type de signaux complexes. Les données sont principalement utilisées pour valider et / ou calibrer les modèles biogéochimiques ainsi que les algorithmes d'observation de la couleur de l'eau par satellite. Les données sont également utilisées sur des périodes de temps restreintes et sur une sélection de paramètres afin de répondre à des besoins d'amélioration de connaissance face à un épisode environnementale ponctuel.

L'application des méthodes d'analyse exploratoire et de fouille de données aux séries temporelles devient indispensable pour appréhender le volume de données des séries HF. Modéliser une série multivariée signifie d'une part de détecter des états ou segments sous-jacents et d'autre part apprendre la dynamique interne entre ces états. Une première classe de méthode consiste à explorer toutes les segmentations possibles par détection de ruptures séquentielles suivie d'une classification des segments générés. La seconde classe, abordée ici, consiste à détecter des changements structurels par classification suivie d'un alignement temporel. Qu'elle que soit la classe des méthodes choisies, elles ont recours aux algorithmes par programmation dynamique en amont ou aval de la partie classification pour diminuer drastiquement le nombre de segmentations possible. Selon l'information disponible, on distingue deux types d'approche d'apprentissage du système de classification : l'apprentissage supervisé et l'apprentissage non supervisé. Dans un cadre supervisé, le système est construit à partir d'un ensemble (données, label) afin d'affiner la discrimination (réseau de neurones, machines à vaste marge, arbre de décision, classification hiérarchique, …) ou la modélisation (modélisation markovienne, graphe, réseau bayésien). Dans un cadre non supervisé, la détermination du système est régi uniquement par les données, il s'agit alors de rechercher un partitionnement des données selon des critères de séparation optimale entre groupes obtenus. Face à la taille des bases de données actuelles de séries HF liés à l'émergence des capteurs, il existe un manque d'information important lié d'une part à la mesure elle-même, données

absentes ou aberrantes et d'autre part, au manque de connaissances actuelles à une si haute résolution. La labellisation de ces bases devient une étape trop lourde à valider par un opérateur humain. La modélisation markovienne a largement fait ses preuves en segmentation de formes ou séquences par apprentissage supervisé. Elle permet d'apprendre d'une part la distribution des données d'un état et d'autre part la dynamique. Pour ces raisons, nous avons choisi d'étendre cette approche dans un cadre non supervisé.

Dans cette thèse, une méthodologie sera présentée pour définir un système numérique automatisé robuste capable de traiter de tel volume de données complexes afin d'améliorer les connaissances quant à la qualité des systèmes aquatiques, avec une attention toute particulière portée à l'étude du déterminisme et de la dynamique des efflorescences du phytoplancton. Cette méthodologie est scindée en deux phases, la caractérisation et la complétion des données HF suivies de leur traitement :

- La première phase consiste à prétraiter une série temporelle multivariée et extraire l'information utile. Les capteurs, installés dans des milieux hostiles, sont sujets à des périodes de calibration, d'entretien voire des pannes et sont donc susceptibles de générer des données bruitées, manquantes voire aberrantes qu'il est nécessaire de filtrer et compléter avant toute exploitation ultérieure. L'apport des travaux dans cette phase est la proposition d'une méthode d'imputation des données manquantes par identification d'une séquence identique par appariement élastique.
- La deuxième étape concerne la modélisation de la série validée. Pour cela, une modélisation statistique de la dynamique des efflorescences est préconisée. Elle s'appuie sur un graphe probabiliste des états de l'environnement par modélisation markovienne. Les modèles de Markov cachés ont une très bonne aptitude à modéliser des processus stochastiques tel que ceux conduisant aux efflorescences du phytoplancton. L'apport de ces travaux se fera à la fois sur l'étude et la conception d'un Modèle de Markov Caché par apprentissage Non Supervisé (MMC-NS), notamment par l'hybridation du modèle de Markov par un algorithme de classification spectrale. Cet algorithme est utile pour son fort pouvoir discriminant à détecter des structures spécifiques dans un jeu de données non convexes.

Cette thèse s'inscrit dans le cadre d'une collaboration étroite entre l'Ifremer, le LISIC/ULCO et l'Agence de l'Eau Artois Picardie afin de développer des systèmes optimisés pour l'étude de l'effet des activités anthropiques sur le fonctionnement des écosystèmes aquatiques et du projet InterReg IVa « 2 mers » DYMAPHY (www.dymaphy.eu).

La méthodologie proposée sera appliquée sur trois bases de données HF de résolution et taille différentes : les données 2005-2009 issues de la station instrumentée MAREL-Carnot installée dans la rade de Boulogne-sur-Mer, les données d'une semaine acquise sur un système

embarqué sur un navire océanographique en Manche orientale et les données d'un mois d'une station déployée sur la rivière Deûle.

Le manuscrit se décompose en trois parties : une partie introductive présentant le contexte de la recherche liée aux compartiments hydrologie et phytoplancton, l'observation et la surveillance de la qualité des eaux (chapitre 1), une partie couvrant le volet numérique de ces travaux de recherche, l'une sur la complétion des données manquantes d'une série HF, l'autre sur l'apprentissage d'un Modèle de Markov Caché par classification spectrale (chapitres 2 et 3), puis une partie applicative consacrée à la caractérisation du fonctionnement d'écosystèmes particuliers par modélisation markovienne (chapitres 4 et 5).

Le chapitre 1 décrit la zone d'étude, la problématique de compréhension de l'environnement aquatique et plus particulièrement le besoin d'améliorer les connaissances sur les efflorescences phytoplanctoniques nuisibles. Les différents facteurs de contrôle et les réponses associées sont expliqués tant du point de vue hydrodynamique qu'hydrobiologique. Les données de l'application principale, issues de la station MAREL-Carnot sont alors présentées afin de mettre en évidence la problématique des données HF.

Le chapitre 2 a pour objectif de proposer un schéma directeur afin d'extraire une information utile et complète à partir de séries temporelles HF parfois bruitées et à valeurs manquantes. Une comparaison des techniques usuelles est apportée afin de mettre en évidence leur applicabilité et leur pertinence selon la distribution et la contiguïté des données.

Le chapitre 3 est consacré aux modalités de mise en œuvre d'un système d'estimation d'états, de leur caractérisation et de leur séquencement, le tout par apprentissage non supervisé. Une approche originale est de chercher à représenter la connaissance et la dynamique de ces états par une modélisation graphique et non d'arrêter le système à un arbre hiérarchique d'états ne permettant pas de visualiser le séquencement entre états. Une modélisation particulière est considérée : un modèle de Markov caché. Dans un premier temps, son architecture, ainsi que les méthodes d'apprentissage non supervisé de celle-ci rencontrées dans la littérature sont introduites. Dans un second temps, ce chapitre détaille le système implémenté et la construction non supervisée de l'architecture du modèle de Markov caché par classification spectrale.

Le chapitre 4 débute par le prétraitement réalisé sur les données de la station MAREL-Carnot. Afin de valider les méthodes de génération des paramètres du modèle de Markov caché non supervisé, un modèle à 2-états fixés a été construit à partir de ces données. Les sorties du système de classification issues de ce modèle appris ont été comparées avec les sorties d'une segmentation experte construite à partir de la stratégie de surveillance DCE. Une fois validé, le système, généralisation du modèle à N-états, est utilisé de manière totalement automatique

sur les données issues de la station MAREL-Carnot et les résultats seront de nouveau confrontés à une interprétation experte.

Le chapitre 5 correspond à l'application de l'approche Markovienne non supervisée via le développement de deux modèles spécifiques dédiés à des Etudes et Recherches ponctuelles mettant en œuvre des systèmes HF différents :

- Dans une première partie, il s'agira de s'intéresser à des données issues d'un système type FerryBox qui présente la particularité d'être couplé à un fluorimètre spectral (approche taxonomique préliminaire par définition de classes algales dominantes) afin de proposer une vision synoptique des états environnementaux rencontrés en Manche orientale lors de la période d'efflorescence du phytoplancton.
- La seconde partie de ce chapitre est consacrée aux données HF issues d'une station fixe de l'Agence de l'Eau Artois Picardie implémentée au printemps 2009 sur la rivière Deûle. L'objectif est de proposer une caractérisation du milieu pour contribuer à la définition d'un réseau de surveillance optimisé du phytoplancton en eau douce.

Chapitre 1 : Contexte environnemental et système d'observation MAREL

1.1. Introduction

L'objet de nos travaux est de construire un outil permettant de caractériser la dynamique phytoplanctonique ainsi que de détecter et de suivre des épisodes de fortes concentrations d'espèces nuisibles ou potentiellement toxiques pour en comprendre le déterminisme. Ce chapitre a pour but de définir les différents facteurs qui interagissent et jouent un rôle de contrôle quant à l'évolution de cette biomasse et de présenter le contexte ainsi que les contraintes opérationnelles.

Une première partie décrit la zone d'étude, soit la Manche orientale et plus précisément la zone côtière de Boulogne-sur-Mer avec sa géographie et son hydrodynamisme. L'accent est mis ensuite sur la description du développement phytoplanctonique avec un intérêt particulier pour la Prymnésiophycée *Phaeocystis globosa* et la Bacillariophycée *Pseudo-nitzschia sp.* dont la prolifération dans notre région (et ailleurs) peut avoir des conséquences négatives sur le fonctionnement de l'écosystème, comme sur la santé humaine. Afin d'étudier la dynamique des efflorescences à une échelle de temps adaptée, le recours à des systèmes de mesure dits à haute fréquence (par rapport aux approches conventionnelles mises en œuvre pour le milieu marin) est indispensable. Ainsi, la seconde partie est dédiée à ces systèmes actuellement en place. Puis, nous focaliserons sur le principal système de mesure étudié dans cette thèse, soit la station MAREL-Carnot installée dans la rade de Boulogne-sur-Mer.

1.2. Contexte environnemental

1.2.1. Géographie

La Manche, située dans une zone tempérée*, est une mer épicontinentale* délimitée au nord par le Royaume-Uni, au sud par la France, à l'ouest par l'océan Atlantique et à l'est par le détroit du Pas-de-Calais et la mer du Nord (figure 1.1).

* Les termes possédant un astérisque sont définis dans le glossaire page 1

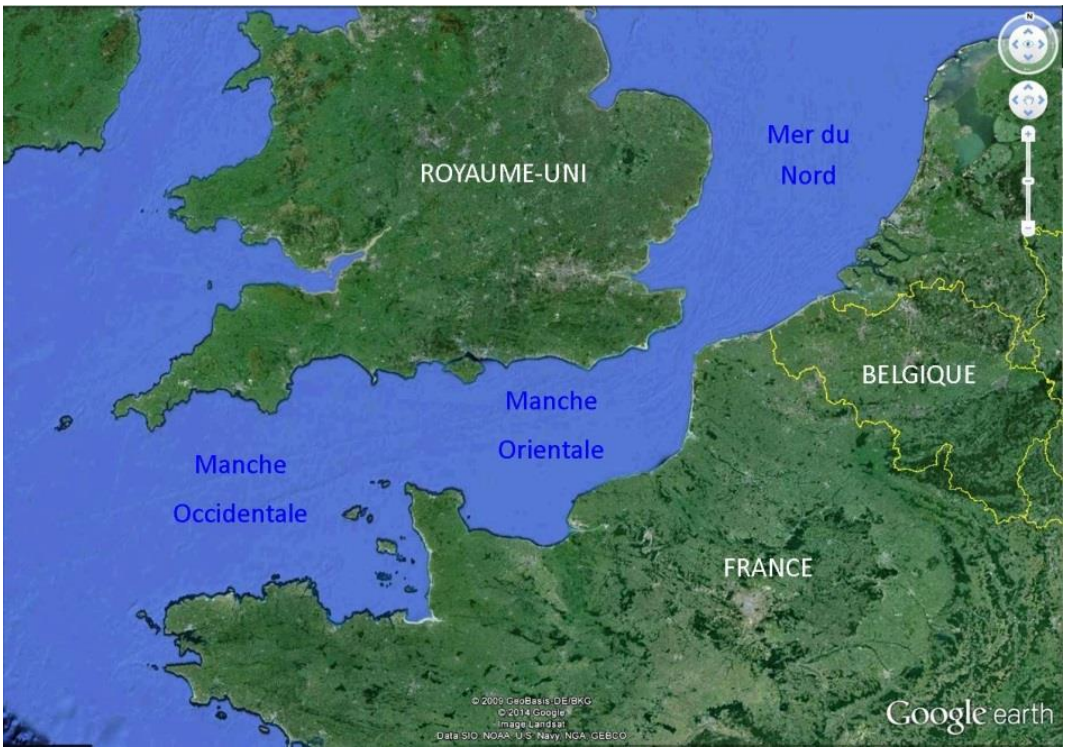


Figure 1.1. La région marine étudiée : La Manche et la baie sud de la mer du Nord.

La Manche occidentale a une profondeur supérieure à 50 mètres avec un maximum à 174 mètres dans la fosse centrale dans la partie nord-est du Cotentin. La Manche orientale s'étend sur 77 000 km² et a une profondeur inférieure à 50 mètres en sachant que la profondeur augmente de la côte vers le large. En forme d'entonnoir, elle est large de 35 km dans le détroit du Pas-de-Calais et de 80 km entre l'île de Wight et le Cotentin (Brylinski, 1993 ; Dauvin et Lozachmeur, 2006 ; Dauvin, 2008).

1.2.2. Hydrodynamisme

La Manche orientale est une mer mégatidale*, avec un marnage* qui varie entre 3,3 et 9,7 mètres. De ce fait, ces grandes marées provoquent des courants puissants, alternatifs et parallèles à la côte (figure 1.2, Sournia *et al.*, 1990) : courant de flot vers le nord-est (vive-eau), courant de jusant vers le sud-ouest (morte-eau) (Brylinski, 1993 ; Dauvin et Lozachmeur, 2006 ; Reynaud *et al.*, 2003 ; Sournia *et al.*, 1990).

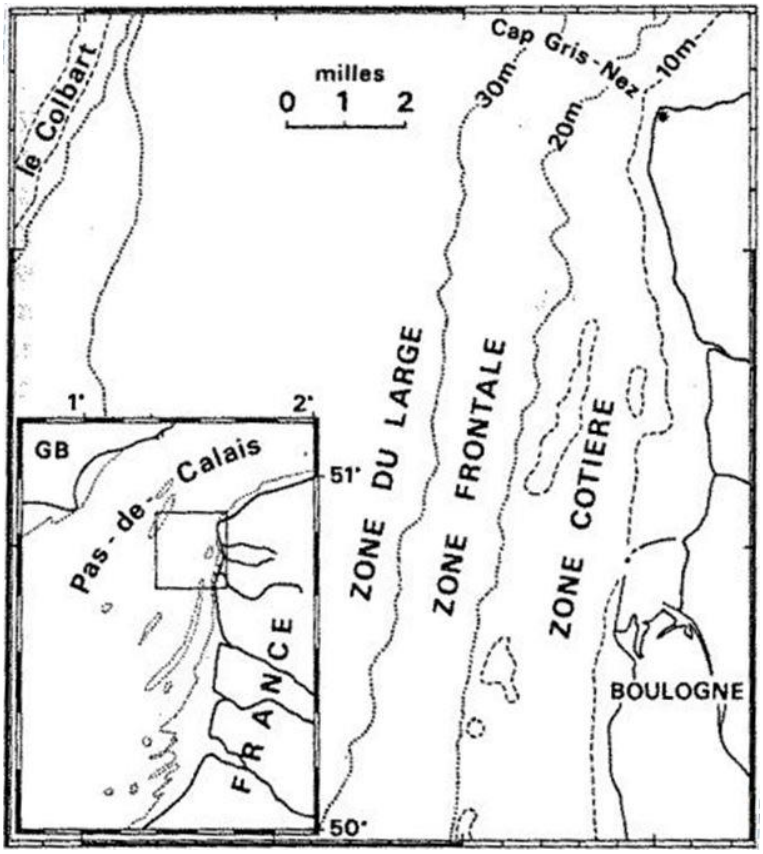


Figure 1.2. Structuration du fleuve côtier en Manche orientale en trois zones : zone du large, zone frontale et zone côtière (source : Sournia et al. 1990).

1.2.2.1. Le fleuve côtier

Ce contexte marégraphique favorise 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), ce qui provoque une chute de la salinité, un apport en nutriments, matières en suspension et matières organiques (Dauvin et Lozachmeur, 2006 ; Dauvin, 2008). Cette zone côtière a reçu le nom de « fleuve côtier » (figure 1.2). Le fort débit de la Seine en fait une source principale, cependant Brylinski *et al.* (1996) indiquent qu'elle n'est pas la seule initiatrice du fleuve côtier, on le doit également à la Somme dont ses apports se font ressentir jusqu'au détroit du Pas-de-Calais.

Le fleuve côtier est délimité sur sa « rive » droite par le littoral et sur sa « rive » gauche par une zone frontale (figure 1.2). Ce fleuve a une largeur oscillant entre 3 et 5 milles nautiques des côtes. Cette variation dépend du cycle de marée avec l'apparition ou la disparition d'une stratification dans la zone frontale (figure 1.3). Le fleuve côtier passe au-dessus de l'eau du large au moment de la morte eau : une stratification s'établit. Au contraire, lors de la vive eau, une barrière turbide verticale s'établit entre le fleuve côtier et l'eau du large : la stratification est détruite. Cet hydrodynamisme particulier est un lieu idéal pour la prolifération du phytoplancton : apport en nutriments constant, pas de profondeur critique, turbidité acceptable pour le développement (Brylinski et Lagadeuc, 1990 ; Brylinski, 1993). Ce phytoplancton peut ainsi être transporté sur de très longues distances (de la baie de Somme au détroit du Pas-

de-Calais) et traverse par conséquent les 3 écosystèmes caractéristiques de la Manche orientale le long des côtes françaises : un estuaire (la Somme), une zone côtière sous influence d'une structure frontale (Boulogne-sur-Mer), une zone côtière peu profonde faisant la transition vers la Mer du Nord (Dunkerque). Ces trois écosystèmes ont justifié la création du Suivi Régional des Nutriments (SRN) qui a permis d'approfondir les connaissances du fleuve côtier (section 1.3.1). Ce suivi a été renforcé dans la zone du fleuve côtier en implantant la station MAREL-Carnot (section 1.3.3), devenant la station la plus côtière de la radiale SRN de Boulogne-sur-Mer, ce qui amène le choix de la rade comme point 0.

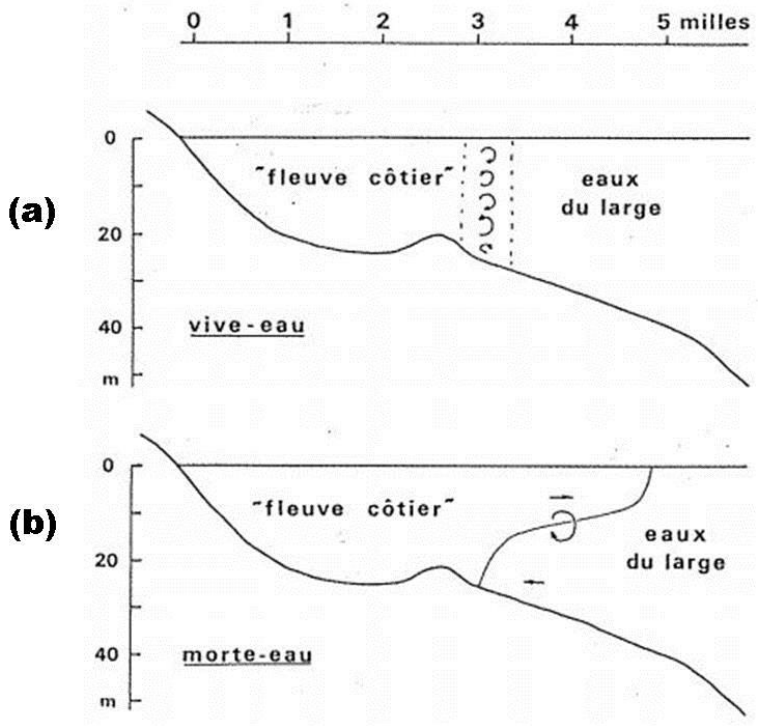


Figure 1.3. Structuration du fleuve côtier en fonction des conditions marégraphiques : (a) Vive-eau, (b) Morte-eau (source : Brylinski, 1993).

1.2.2.2. La rade de Boulogne-sur-Mer

La rade de Boulogne-sur-Mer est entourée du littoral, de la digue Carnot et de la digue Nord. Cette rade est également située à la sortie de la Liane qui y déverse ses eaux à chaque ouverture du barrage Marguet (figure 1.4).

Figure 1.4. La rade de Boulogne-sur-Mer avec l'emplacement du barrage Marguet, de la digue Nord, du caisson, de la digue Carnot et de la station MAREL-Carnot (source : Google Earth).

Du fait de cet agencement géographique, les circulations des masses d'eau entre l'extérieur et l'intérieur de la rade sont déphasées de 2 à 3 heures (Hebert et Lefebvre, 2004) (illustration en 2005 sur la figure 1.5).

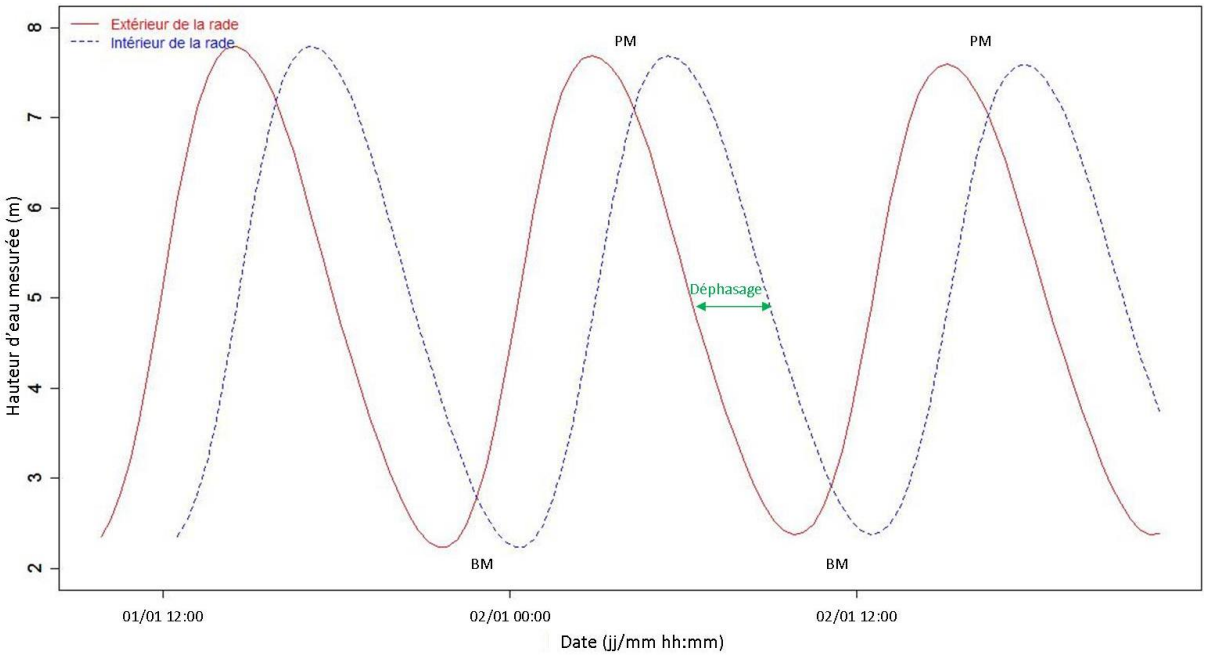


Figure 1.5 – Déphasage de 2 à 3 heures des marées entre l'intérieur et l'extérieur de la rade de Boulogne-sur-Mer.

Lors de l'étude de la circulation des masses d'eau à l'intérieur de la rade (Dolle *et al.*, 2001 ; Hebert et Lefebvre, 2004), il a été montré que le flot se fait approximativement en 5h10' alors

que le jusant se fait en 7h20' (figure 1.6) et que cette circulation pouvait être découpée en 5 phases différentes (figure 1.7) :

- Phase 1 (BM à BM+2h) : La rade se remplit suite à l'étale de la Basse Mer. Le courant au niveau de la digue Carnot est orienté vers le sud-ouest.
- Phase 2 (BM+2 à BM+2h45) : Le remplissage de la rade se poursuit alors qu'à l'extérieur il y a une renverse de courant. Le courant au niveau de la digue Carnot est orienté nord-est.
- Phase 3 (BM+2h45 à PM-1h15) : La renverse de courant à l'extérieur de la rade est terminée, le courant est maintenant orienté nord-est. La rade quant à elle, continue son remplissage avec un courant au niveau de la digue Carnot orienté sud-ouest.
- Phase 4a (PM-1h15 à PM) : Les courants à l'extérieur de la rade sont très soutenus en opposition à ceux se trouvant à l'intérieur qui sont très faibles. Cette différence engendre la formation d'un gyre* à l'intérieur de la rade. Le courant au niveau de la digue Carnot est orienté nord-est.
- Phase 4b (PM à PM+3h30) : La rade se vidange. L'orientation du courant au niveau de la digue Carnot reste inchangée.
- Phase 5 : (PM+3h30 à BM) : La vidange de la rade se poursuit alors qu'il y a une renverse de courant à l'extérieur de celle-ci. Aucun changement de direction du courant au niveau de la digue Carnot n'est signalé.

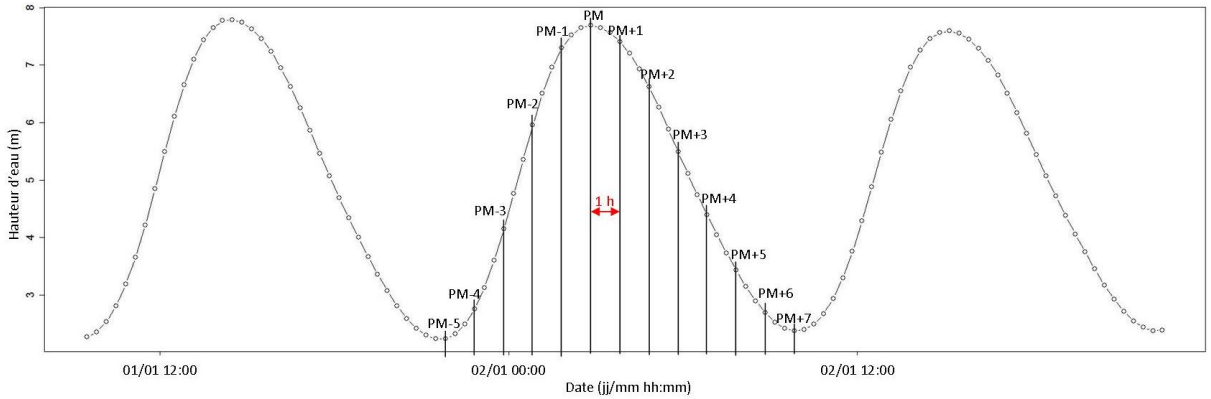


Figure 1.6. Hauteur d'eau mesurée par la station MAREL-Carnot du 01/01/05 au 03/01/05 inclus avec la projection du niveau de la marée.

Cette courantologie permettra de connaître l'origine de la masse d'eau étudiée : comme il a été détaillé dans l'analyse des phases ci-dessus, un effet marin est principalement ressenti lors du remplissage de la rade, contrairement au moment de la vidange lors duquel l'impact anthropique* est dominant et est fonction de l'intensité des vidanges via le barrage Marguet.
Le déplacement des masses d'eau a une incidence sur le développement des particules inertes ou vivantes (phytoplancton).

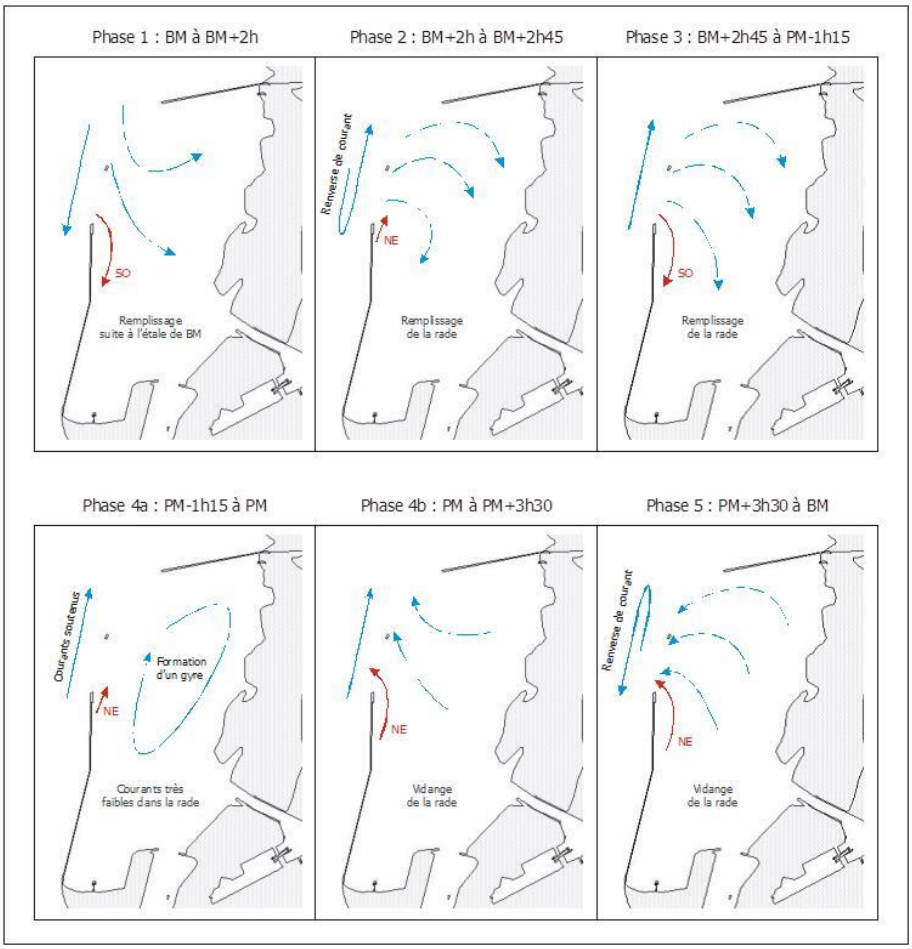


Figure 1.7. Circulation des masses d'eaux dans la rade de Boulogne-sur-Mer (en rouge: courants mesurés au niveau du site MAREL-Carnot) (Hebert et Lefebvre, 2004).

1.2.3. Hydrobiologie

Le phytoplancton est le premier maillon de la chaîne alimentaire puisqu'il se situe à la base d'un ensemble d'interactions qui structurent un réseau trophique*. Selon la concentration en nutriments disponible ainsi que des facteurs physiques (température, luminosité, turbulence, turbidité des masses d'eau), il y a, ou non, présence de phytoplancton (figure 1.8). La croissance optimale du phytoplancton est atteinte lors des conditions normales d'abondance en éléments nutritifs : allongement des journées et donc de l'éclairement ainsi que du réchauffement des masses d'eau. Durant la période post-automnale et hivernale, les conditions n'étant plus favorables au développement du phytoplancton, une reconstitution des stocks de nutriments est réalisée grâce notamment aux différents fleuves se jetant dans la mer. L'amélioration des connaissances de la dynamique du phytoplancton telle que historiquement décrite par Margalef (1978) ci-dessous justifie le recours à une étude multi-paramètres et à une fréquence temporelle adaptée.

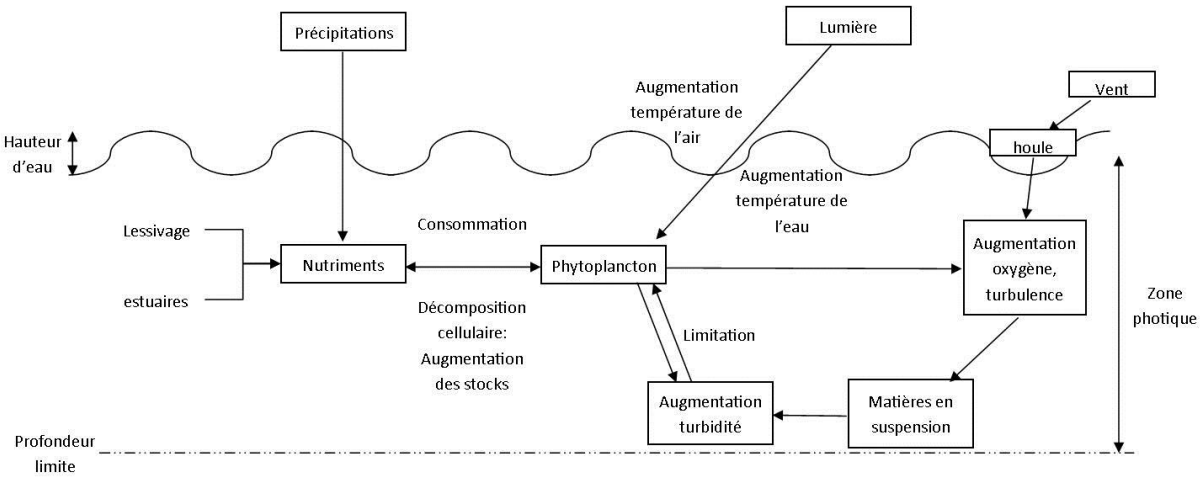


Figure 1.8. Représentation schématique des évènements physiques entrant dans le processus du développement phytoplanctonique.

Margalef (1978) expose que peu importe l'environnement aquatique où l'on se trouve, les nutriments sont le principal facteur limitant au développement du phytoplancton, comme l'explique Redfield (1958) en définissant chaque ratio Phosphate : Nitrate : Carbone disponible dans l'environnement par rapport au besoin du phytoplancton pour son développement. Ce rapport a été repris par Brzezinski (1985) qui mesura les ratios Silicate : Carbone : Nitrate et Phosphate : Nitrate selon le cycle de lumière imposé sur des diatomées. La température étant basse en hiver, le développement du phytoplancton est par conséquent limité et le stock en nutriments est renouvelé. Les sources de ces nutriments sont variées, elles peuvent provenir des différents fleuves se jetant en mer et donc des estuaires (création du fleuve côtier), des précipitations ainsi que du lessivage. A la fin de l'efflorescence phytoplanctonique, les concentrations en nutriments sont très faibles. La décomposition cellulaire du phytoplancton en matière organique permet une remobilisation du stock de nutriments. Margalef (1978) explique que l'efflorescence printanière est caractérisée par une température modérée et une intensité lumineuse élevée. Il précise que la majorité du phytoplancton est constituée de petites cellules. Lorsque l'on se trouve dans une eau riche en nutriments, les petites cellules, ayant un rapport surface / volume plus important, se développent plus rapidement que les cellules volumineuses. De plus, la turbulence générée par le vent participe au développement phytoplanctonique. En effet, selon la forme des espèces, la turbulence permet au phytoplancton de voyager verticalement et horizontalement dans la colonne d'eau afin d'accéder aux nutriments qui y sont dispersés. Cependant, une forte turbulence peut également entraîner une éjection des cellules phytoplanctoniques hors de la zone photique* pouvant provoquer leur mort si elles n'y reviennent pas. La probabilité de cette éjection de la zone photique additionnée à leur possible consommation par la faune s'oppose au taux d'augmentation des cellules phytoplanctoniques dont la division cellulaire s'effectue la nuit. La profondeur maximale de la zone photique est déterminée à partir de la turbidité du milieu. L'augmentation de la turbulence par l'action du vent augmente la quantité de matière remise en suspension et donc de la turbidité. Il faut ajouter à cela l'effet du

développement phytoplanctonique qui, en grande quantité, réduit la pénétration de la lumière dans l'eau (création de mousse lors des efflorescences de *Phaeocystis*). Les différentes familles phytoplanctoniques ne réagissent pas de la même manière aux facteurs physico-chimiques entrants dans leur développement. Par exemple, les diatomées prolifèrent dans des eaux plus turbides que les Prymnésiophycées.

En plus de cette description, Reynolds *et al.* (2002) proposent un modèle permettant de regrouper les espèces de phytoplancton par groupes fonctionnels, c'est-à-dire en des ensembles d'espèces caractérisés par la morphologie (taille, dimensions linéaires maximales, rapport surface / volume) et la physiologie (taux de croissance, efficacité photosynthétique, taux d'absorption des éléments nutritifs) de celle-ci. Wyatt (2014) reprend les recherches des cinquante dernières années sur la dynamique du phytoplancton qui a été étudiée selon les facteurs de contrôle (lumière, température, nutriments et turbulence analysés deux à deux). Il présente ses travaux où les facteurs de contrôle utilisés sont la turbulence et les concentrations en nutriments, comme l'avait fait Margalef (1978). L'espace construit avec ces facteurs est découpé en quatre zones :

- Fortes concentrations en nutriments et forte turbulence : Domaine où l'efflorescence printanière est la plus élevée (typiquement dans les mers et lacs tempérés et boréaux) et les blooms automnales ;
- Fortes concentrations en nutriments et faible turbulence : Domaine des marées rouges (qui ont été observées par Margalef) ;
- Faibles concentrations en nutriments et faible turbulence : Typique des eaux du large où la stratification est possible. Margalef caractérisait ce domaine avec les dinoflagellés ;
- Faibles concentrations en nutriments et forte turbulence : C'est un domaine dit « vide ». Les conditions sont celles que l'on retrouve en période hivernale dans les eaux côtières.

La prolifération de certaines espèces phytoplanctoniques entraine un risque environnemental qui est à l'origine d'une surveillance des masses d'eau côtières. A l'échelle planétaire, environ 70 des 4 000 espèces phytoplanctoniques sont toxiques pour l'homme, les coquillages et les poissons. A l'échelle locale de la Manche orientale, la diatomée du genre *Pseudo-nitzschia* est l'une de ces espèces toxiques (production de toxines amnésiantes), d'autres cependant sont nuisibles comme la Prymnésiophycée *Phaeocystis globosa* (forte biomasse) (Ifremer environnement, 2014a). En Manche orientale, la prolifération des diatomées et de *Phaeocystis* n'est plus à démontrer (Davies *et al.*, 1992 ; Lamy *et al.*, 2006 ; Lancelot *et al.*, 1998), de même que son cycle de vie complexe (Rousseau *et al.*, 1994).

La Prymnésiophycée *Phaeocystis globosa* est une algue nuisible qui se développe majoritairement de mars à juin. Les périodes où il y a très peu de vent (donc peu de turbulence) et une forte lumière sont des moments favorables au développement de

Phaeocystis (Seuront et Souissi, 2002). Elle est capable de se développer sur des sources organiques du phosphore quand le phosphate est épuisé, mais son principal facteur limitant reste le nitrate. De ce fait, lorsque la concentration en nitrate diminue, il y a une augmentation du nombre de cellules de *Phaeocystis* au détriment des diatomées car elle est plus compétitive que ces dernières dans un milieu turbulent. De même lors de la compétition pour la lumière, *Phaeocystis* possède des propriétés de flottabilité lui permettant de bénéficier de conditions favorables pour la photosynthèse, contrairement aux diatomées qui sont plus lourdes (Breton *et al.*, 2006). Davies *et al.* (1992) montrent dans leurs travaux que *Phaeocystis globosa* n'est pas une nourriture appréciée par le zooplancton, sa consommation par les prédateurs est donc fortement restreinte, ceux-ci se rabattant sur les diatomées.

Rousseau *et al.* (1994) présentent dans leur article plusieurs types de cellules de *Phaeocystis* : les non-mobiles, les flagellées et les microzoospores. Les cellules non-mobiles sont similaires aux cellules coloniales avec une taille de 4,5 à 8 µm. Capables d'une division végétative, elles ont une forte habilité à générer de nouvelles colonies. Ces cellules ont des propriétés pour adhérer à des surfaces solides ce qui explique leur habitat benthique*. Les flagellés, identifiés après la désorganisation des colonies, possèdent deux flagelles et un haptonéme*. Leurs tailles varient entre 4,5 et 8 µm et ont une durée de vie très courte (24 – 48 h). Les microzoospores sont des cellules de type flagellé de petites tailles (3 - 5 µm). Elles se développent à l'intérieur des colonies qui n'augmentent pas de taille par division végétative. Elles seraient une anomalie de développement. Les colonies, quant à elles, sont constituées de cellules qui migrent vers le bord et restent situées sur une surface sphérique à 15 - 20 µm de distance. Le diamètre des colonies varie de 10 µm à 8 mm voire 20 mm pour *Phaeocystis globosa*. Quand la colonie grandit, sa forme sphérique disparaît pour être allongée.

Les diatomées étant plus aptes à se développer dans les masses d'eau à forte turbulence et sur une plus grande gamme de température, elles se développent sur une plus longue période que *Phaeocystis*. Certaines espèces de diatomées sont toxiques. La diatomée *Pseudo-nitzschia* est l'un de ces genres. Ses espèces toxiques sont capables de sécréter de l'acide domoïque (substance amnésiante) qui engendre des problèmes de santé chez l'Homme et chez certains organismes marins. Les efflorescences côtières de la diatomée *Pseudo-nitzschia*, de faible densité, ne sont pas visibles à l'œil nu comparées à la mousse de *Phaeocystis*. Les espèces présentent sur notre site d'étude sont *Pseudo-nitzschia fraudulenta*, *P. pseudodelicatissima* (toxique), *P. pungens*, *P. seriata* et *P. multiseries* (toxique) (Trainer *et al.*, 2012). Il est noté dans Ifremer environnement (2014a) que des proliférations de *Pseudo-nitzschia* ont lieu régulièrement au printemps sachant que *P. pseudodelicatissima* peut proliférer à des concentrations importantes, alors que *P. multiseries* n'est jamais abondante. De forme allongée, elles sont souvent assemblées en chaînes. Leur taille peut varier entre 50 et 180 µm et leur largeur de 1,5 à 3,4 µm. Lorsque cette taille diminue, la cellule se reproduit sexuellement pour ne pas mourir.

Nous avons vu que la dynamique et la composition du phytoplancton varient rapidement selon l'environnement marin dans lequel il se trouve. De ce fait, pour mesurer la qualité de l'eau et se rapprocher de son objectif de « bon état » des eaux, la Directive Cadre sur l'Eau 2000/60/CE a mis en œuvre un programme de surveillance consistant à suivre un certain nombre d'éléments de qualité et de leur attribuer un classement : mauvais, médiocre, moyen, bon ou très bon. Parmi les éléments de qualité biologiques, on y trouve le phytoplancton dont l'indicateur utilisé résulte d'une combinaison des indices de biomasse, d'abondance et de composition. Pour les éléments de qualité physico-chimiques soutenant la biologie, il y a les nutriments. Pour ceux-ci, la correspondance entre la concentration d'azote inorganique dissous normalisée à une salinité de 33 PSU et les valeurs de RQE (Ecological quality ratio) de l'indice de chlorophylle-*a* est calculée (Ifremer environnement, 2014b).

Le bilan de santé 2010 issu de la commission Oslo-Paris (OSPAR) indique que notre zone d'étude (région II) possède de nombreux problèmes d'eutrophisation (OSPAR, 2010). La quantité de nitrate dans le fleuve côtier étant importante, de même que les efflorescences de *Phaeocystis*, l'étude de la qualité de l'eau à travers des réseaux basse et haute fréquence permet d'approfondir les connaissances sur la dynamique de l'environnement dans notre contexte particulier (efflorescence de *Phaeocystis* et problèmes d'eutrophisation).

L'amélioration des connaissances de la qualité de l'environnement marin via des paramètres physico-chimiques et biologiques permet donc de définir des indicateurs qui servent à leur tour de juger de l'efficacité des mesures de gestion qui sont prises pour atteindre le « bon état » écologique et limiter l'eutrophisation. Un des objectifs de ce travail de thèse est de développer un outil de diagnostic exploitant les données d'un système instrumenté haute fréquence permettant aux gestionnaires de l'environnement de mettre en œuvre des programmes de mesures adaptés.

1.3. MAREL

1.3.1. De la basse fréquence…

Un certain nombre de réseaux basse fréquence (échantillonnage selon les approches conventionnelles, c'est-à-dire avec une fréquence de prélèvement mensuel à hebdomadaire au mieux) ont été mis en place par l'Ifremer afin d'observer et de surveiller la qualité de l'environnement marin. On peut citer le réseau de surveillance du phytoplancton et des phycotoxines (REPHY) créé en 1984. Il a pour objectifs :

- Observer l'ensemble des espèces phytoplanctoniques des eaux côtières.
- Recenser les évènements tels que les eaux colorées, les efflorescences exceptionnelles et les proliférations d'espèces toxiques ou nuisibles pour la faune marine.
- Surveiller plus particulièrement les espèces produisant des toxines dangereuses pour les consommateurs de coquillages.

Les prélèvements pour ce réseau sont mensuels sauf en période printanière où ils deviennent bimensuels dans les eaux côtières (Ifremer environnement, 2014a).

Depuis 1992, une extension vers le large de ce réseau REPHY est réalisée régulièrement via le réseau Suivi Régional des Nutriments (SRN). Des prélèvements sont effectués dans 3 écosystèmes contrastés : la Baie de Somme, Boulogne-sur-Mer et Dunkerque, sur une radiale qui a été définie suivant un gradient côte-large. La radiale de Boulogne-sur-Mer comprend 3 points de prélèvement : l'un dans la zone côtière (fleuve côtier), l'un dans la zone du large et le dernier entre les deux dans la zone frontale qui est selon la période : soit sous l'influence d'une stratification fleuve côtier / zone du large, soit constitué uniquement de l'eau du large. Les paramètres prélevés en chaque point sont les suivants : la salinité, la température, la turbidité, l'ammonium, le nitrate, le nitrite, le phosphate, le silicate, les matières en suspension, la matière organique particulaire, la chlorophylle-*a* et phéopigments, la composition phytoplanctonique (Nzigou et Lefebvre, 2013). Les prélèvements se font mensuellement. La fréquence d'échantillonnage devient bimensuelle entre les mois de mars et juin, période de prédilection de *Phaeocystis globosa*. A ces trois points s'ajoute le point 0 réalisé par la station MAREL-Carnot (section 1.2.2.1).

Dickey (2003) 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 1.9). La turbulence dont l'échelle temporelle est inférieure à l'heure, joue un rôle dans les efflorescences phytoplanctoniques qui peuvent s'étendre de quelques mètres à une dizaine de kilomètres sur une période inférieure à la semaine et pouvant s'étendre à plus d'un mois. Dans son papier, Dickey (2003) pointe du doigt le fait que l'océanographie dépend d'un grand nombre de disciplines (biologie, chimie, physique des océans) et qu'il faut des systèmes capables de mesurer dans l'ensemble de ces domaines à une haute fréquence pour englober toute la variabilité de l'écosystème étudié. Ces systèmes de mesures automatisées permettent l'acquisition de données à des fréquences journalières, horaires et infra-horaires et seront définis par convention comme des systèmes à haute fréquence ou à haute résolution. Par conséquent, afin d'étudier au mieux le développement phytoplanctonique dans notre milieu structuré par les marées avec des fréquences de l'ordre de 14 jours (Morte Eau, Vive Eau), 12 heures et 6 heures (Pleine Mer, Basse Mer) et influencé par l'ouverture du barrage et les périodes de dragages, les systèmes d'acquisition à haute fréquence s'imposent. Alors que l'approche conventionnelle du type REPHY / SRN permettra de définir la dynamique des efflorescences d'une manière générale en tenant compte des variabilités saisonnières et interannuelles et de conclure quant à des tendances à long terme, la haute fréquence permettra d'appréhender des phénomènes et des processus clefs potentiellement déterminants pour expliquer l'initiation, le maintien et le déclin d'un bloom.

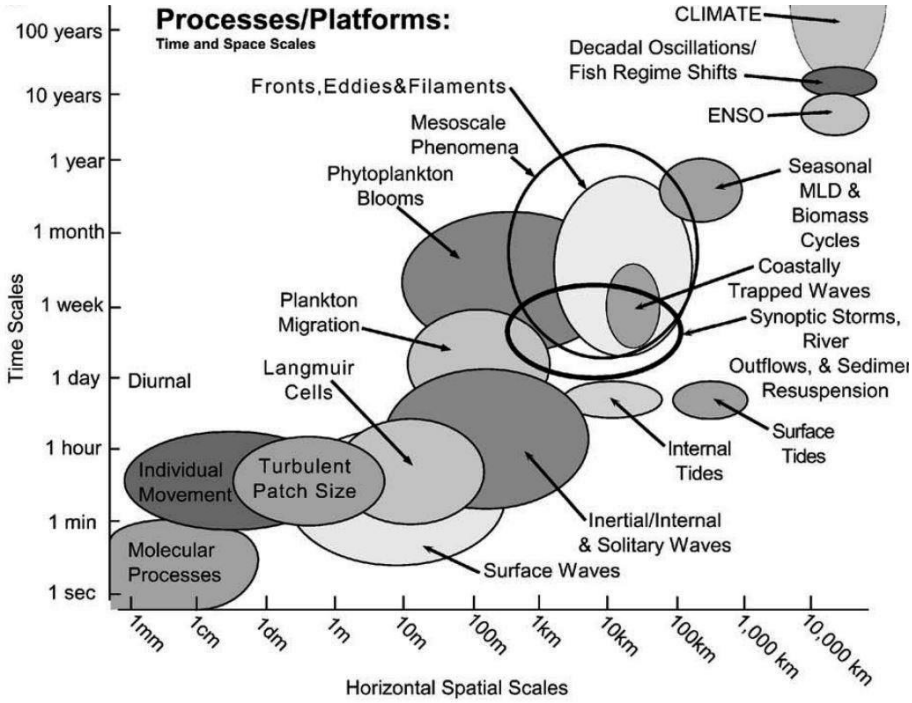


Figure 1.9. Les échelles de temps et d'espace de plusieurs processus physiques et biologiques illustrés par des ovales (Dickey, 2003).

1.3.2. … à la haute fréquence

La prise de conscience générale des problèmes d'environnement, notamment au niveau du littoral, conduit à renforcer la surveillance qui s'y exerce tout en étant le plus proche possible des échelles de temps qui caractérisent les phénomènes étudiés.

Par l'expérience acquise depuis de nombreuses années dans l'exploitation des réseaux de surveillance de l'environnement, l'Ifremer a mis en évidence le besoin de développer des systèmes de surveillance automatisée de l'environnement et des effets directs et indirects des activités humaines sur le milieu marin. Les développements technologiques concernant les capteurs physico-chimiques permettent la réalisation de réseaux de stations instrumentées autonomes, effectuant des mesures à fréquence élevée et rapidement disponibles pour les utilisateurs via notamment des interfaces web.

C'est dans ce contexte qu'est né le concept des stations MAREL (Mesures Automatisées en Réseau pour l'Environnement et le Littoral). Ainsi, depuis le début des années 90, le concept des stations MAREL a été validé puis décliné selon différentes familles de produits adaptés aux contraintes environnementales ainsi qu'aux demandes des utilisateurs et par conséquent en fonction des paramètres qui doivent être mesurés. On peut ainsi identifier, par exemple, les systèmes :

- MAREL SMATCH (nke instrumentation ©) : Cette bouée légère d'une dizaine de kilogrammes est conçue pour une utilisation sur de longues périodes. Ce système multi-paramètres peut mesurer la température, la conductivité (salinité), l'oxygène dissous, la fluorescence et le pH.

- MAREL estran (nke instrumentation ©) : Ce système transportable d'une vingtaine de kilogrammes est installé sur les tables ostréicoles. Il est recouvert d'eau pendant toutes les marées et transmet ses données uniquement à marée basse. Ce système multi-paramètres peut mesurer l'oxygène dissous, la turbidité, la conductivité, la température, le pH, la chlorophylle et les sels nutritifs.
- Bouée MOLIT (Mer Ouverte LITtorale) (nke instrumentation ©) : Ce système en forme de bateau permet un pompage à 2 niveaux : sub-surface et fond. Les paramètres mesurés sont la température, la salinité, l'oxygène dissous, la turbidité, la fluorescence et les sels nutritifs.
- Station MAREL-Carnot : description complète dans la section 1.3.3

Ces stations ont pour objectif la mesure des paramètres physico-chimiques essentiels et indicateurs caractéristiques de l'eau de mer et support à la biologie. L'originalité de ces stations réside dans le système de pompage envoyant l'eau à travers les systèmes où elle est analysée. Lorsqu'il n'y a pas de mesure, une chloration de l'eau par électrolyse protège les capteurs contre le développement de biofouling* et permet aux stations de rester actives sans intervention pendant 3 mois.

Le maintien de ces installations en bon état de fonctionnement a permis aux équipes concernées d'acquérir une solide expérience en maintenance opérationnelle. D'autre part, la multiplication des stations MAREL en France métropolitaine (figure 1.10), regroupées en réseaux locaux a nécessité la mise en place d'une structure chargée d'organiser ces différents réseaux sous la forme d'un projet Ifremer intitulé : « Mise en œuvre et évolution des réseaux de mesure *in-situ* côtier ». Il regroupe les stations : MAREL-Carnot, MAREL Baie de Seine, MAREL Iroise, MAREL Gironde, MAREL GIP Loire, MAREL Rhône (MesuRho), MAREL Vilaine et MAREL Réseau des Iles.

Figure 1.10. Localisation des stations instrumentées du réseau MAREL « Mise en œuvre et évolution des réseaux de mesure in situ côtier ».

1.3.3. MAREL-Carnot

En 2001, le contrat plan état région (CPER) intitulé « Etude et observation de l'écosystème côtier de la Manche orientale : le bloom de *Phaeocystis* et ses effets sur l'écosystème » a été mis en place. L'objectif de cette étude était de savoir comment le bloom printanier de *Phaeocystis* perturbait l'écosystème côtier de la Manche orientale. L'automatisation de l'acquisition des données *in situ* était indispensable pour répondre à cet objectif. C'est pourquoi, la station MAREL-Carnot a été installée dans la rade de Boulogne-sur-Mer en février-mars 2004 et fut inaugurée en novembre 2004 (figure 1.11). Le choix de l'emplacement est un compromis entre les objectifs scientifiques et les contraintes techniques (Lefebvre *et al.*, 2002). Bien que pour certains la zone côtière soit le seul site digne d'intérêt, la solution retenue permet à cette station d'être à la fois sous influence marine (salinité supérieur à 33), et de mesurer des dessalures engendrées par l'arrivée des eaux douces de la Liane lors de l'ouverture du barrage Marguet et est donc sensible aux apports d'origine anthropique (dont les nutriments, éléments essentiels à la croissance du phytoplancton).

La station est constituée d'un tube (15 mètres de long et pesant 12 tonnes) contenant les capteurs installés sur le flotteur afin de suivre les mouvements liés à la marée (figures 1.12 et 1.13).

Figure 1.11. Station de mesure MAREL-Carnot.

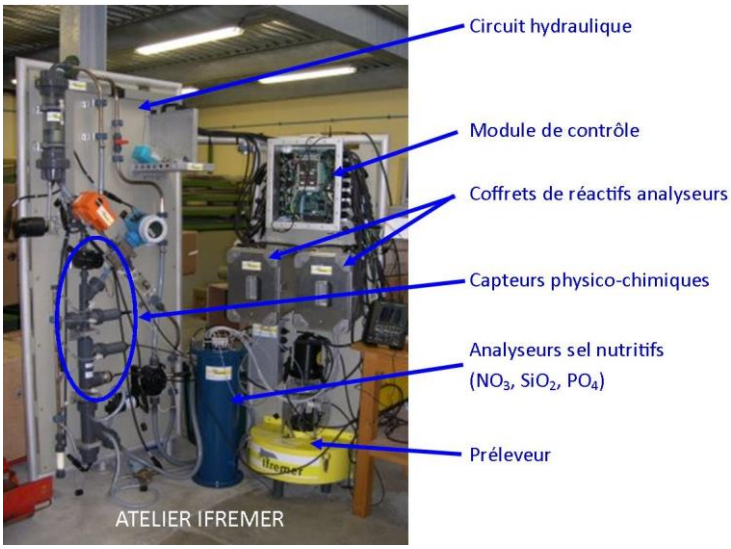


Figure 1.12. Les différentes parties du système hydraulique lors des essais au laboratoire.

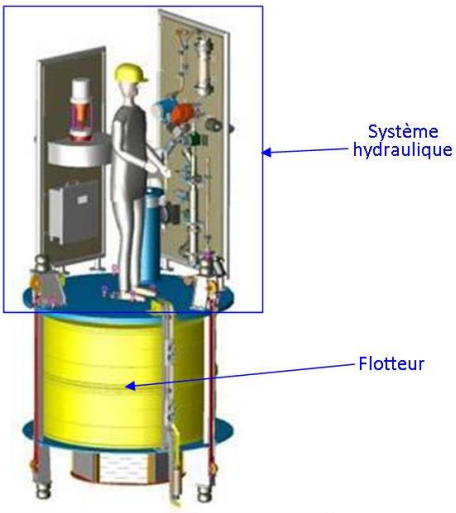


Figure 1.13. Le flotteur situé à l'intérieur du tube.

Le prélèvement de l'eau se fait en sub-surface à 1,5 mètre de profondeur et l'eau est envoyée dans la chambre de passage où elle est redistribuée aux différents capteurs. Un local qui se situe à l'intérieur de la structure de la rade fut aménagé pour recevoir le système de poulies, permettant de gérer les câbles lors des mouvements descendants et ascendants liés à la marée, ainsi que de remonter la station durant les maintenances. Le système GSM envoie les données 2 fois par jour au centre Ifremer Manche Mer du Nord. Des batteries alimentent l'ensemble des instruments.

Cette station multi-capteurs permet de mesurer toutes les 20 minutes la conductivité (mS/cm), la température de l'eau et de l'air (°C), la fluorescence (FFU), la turbidité (NTU), la concentration en oxygène dissous ($mg.L^{-1}$), le P.A.R. (Photosynthetically Active Radiation, µmol de photons $.s^{-1}.m^{-2}$), la direction (degré) et la vitesse du vent ($m.s^{-1}$), la hauteur d'eau (m) et toutes les 12 heures, la concentration en nitrate, en phosphate et en silicate ($\mu mol.L^{-1}$) (tableau 1.1). L'échantillonnage des sels nutritifs est différent car la quantité de réactifs embarqués sur la station est limitée et le temps maximal pour leur utilisation n'excède pas les 3 mois. Une rotation de l'ensemble des capteurs est effectuée trimestriellement afin de contrôler la fiabilité des mesures et de réaliser un recalibrage si besoin.

En plus de ces mesures, certains paramètres sont calculés : le niveau de la mer (marégraphe en mètre), la salinité (PSU), l'oxygène dissous corrigé ($mg.L^{-1}$) et la saturation en oxygène (%) selon les formules suivantes :

- La salinité (notée CSAL1) à partir de la conductivité (notée E_CO1) ; la température étant prise en compte directement dans la mesure du capteur, cette fonction n'est applicable qu'au système MAREL. L'évolution de la salinité est la même que la conductivité, cependant son niveau est moins important de même que son étendue (différence entre le maximum et le minimum du signal) :

$$CSAL1 = 0{,}01230 - 0{,}02984 \times \text{E\_CO1}^{1/2} + 0{,}47731 \times \text{E\_CO1} + 0{,}03563 \times \text{E\_CO1}^{3/2} - 0{,}00230 \times \text{E\_CO1}^{2} + 0{,}0001 \times \text{E\_CO1}^{5/2} \quad (1.1)$$

- L'oxygène dissous corrigé (noté C_O21) à partir de la salinité (notée CSAL1), l'oxygène dissous (noté E_O21) et la température du capteur (TC en °C) ; par conséquent ce calcul est dépendant du capteur MAREL. L'évolution de l'oxygène dissous corrigé est la même que pour l'oxygène dissous non corrigé, cependant son amplitude est moins importante :

$$C\_O21 = E\_O21 \times \exp\left(-\frac{CSAL1}{1{,}80655} \times \left(-0{,}1288 + \frac{53{,}44}{TC + 273{,}15} - 0{,}04442 \times \log(TC + 273{,}15) + 0{,}00071145 \times (TC + 273{,}15)\right)\right) \quad (1.2)$$

- Le pourcentage de saturation en oxygène (noté CSAT1) à partir de l'oxygène dissous (noté E_O21), la température du milieu (notée ETCO1) et la salinité (notée CSAL1). Cependant, il faut calculer la solubilité (noté SO et sans unité) qu'il faudra exprimer en mg.L^{-1} pour obtenir le résultat attendu. Cette saturation en oxygène suit la même évolution que l'oxygène dissous mais le signal possède des amplitudes plus fortes. Cette fonction n'est pas spécifique au système MAREL (Benson et Krause, 1984):

$$SO = \exp\left(-135{,}90205 + \frac{1{,}575701 \times 10^{5}}{ETCO1 + 273.15} - \frac{6{,}642308 \times 10^{7}}{(ETCO1 + 273{,}15)^2} + \frac{1{,}243800 \times 10^{10}}{(ETCO1 + 273{,}15)^3} - \frac{8{,}621949 \times 10^{11}}{(ETCO1 + 273{,}15)^4} - CSAL1 \times \left(0{,}017674 - \frac{10{,}754}{ETCO1 + 273{,}15} + \frac{2140{,}7}{(ETCO1 + 273{,}15)^2}\right)\right) \quad (1.3)$$

Solubilité exprimée en mg.L^{-1} :

$$SO = SO \times 0{,}0319988\, mg.L^{-1} \quad (1.4)$$

Calcul du pourcentage de saturation en oxygène :

$$CSAT1 = 100 \times \frac{E\_O21}{SO} \quad (1.5)$$

Tableau 1. 1. Tableau de synthèse des paramètres de la station MAREL-Carnot avec leur acronyme, leur unité, leur gamme capteur et expert* ainsi que la précision associée à chaque mesure.*

Paramètre	Acronyme	Unité	Gamme capteur*	Gamme expert*	Précision
Oxygène dissous corrigé	C_O21	$mg.L^{-1}$	0 - 20	5 - 20	0,2
Oxygène dissous non corrigé	E_O21	$mg.L^{-1}$	0 - 20	5 - 20	0,2
Saturation en oxygène	CSAT1	%	-	60 - 130	-
Fluorescence	ECHL1	FFU	0 - 150	0 - 50	10%
pH	E_PH1	UpH	6,5 – 9,5	6,5 – 9,5	0,2
Salinité	CSAL1	PSU	0 – 35,5	20 - 40	-
Conductivité	E_CO1	$mS.cm^{-1}$	0 - 70	30 - 60	0,3
Température de l'eau	E__TA	°C	-5 - 30	0 - 30	0,1
Température de l'air	ETCO1	°C	-40 - 60	-5 - 40	0,1
Hauteur d'eau	XMAHH	m	-	0 - 10	-
Vitesse du vent en moyenne	E_VVM	$m.s^{-1}$	-	0 - 41	1
Vitesse du vent en rafale	E_VVR	$m.s^{-1}$	-	0 - 41	1
Direction du vent	E_VDM	degré	-	0 - 360	10
P.A.R.	E_LU1	µmol de photons $.s^{-1}.m^{-2}$	0 - 3 000	0 – 3 000	0,01
Turbidité	E_TU1	NTU	0 – 4 000	0 - 150	10%
Concentration en Nitrate	C_NI1	$\mu mol.L^{-1}$	0 - 100	0 - 100	0,05
Concentration en Phosphate	C_PO1	$\mu mol.L^{-1}$	0 - 10	0 - 100	0,05
Concentration en Silicate	C_SI1	$\mu mol.L^{-1}$	0 - 50	0 - 100	0,05

A la réception des données au centre Ifremer, un niveau de traitement et de qualité leur est attribué (tableau 1.2).

Tableau 1. 2. Les niveaux de traitement et de qualité associés à chaque donnée mesurée par la station MAREL-Carnot.

Niveau de traitement		Niveau de qualité	
T0 :	données brutes non qualifiées	0 :	non qualifiée (niveau de traitement à T0)
T0.5 :	données contrôlées automatiquement (« invisible » de l'utilisateur)	1 :	valeur bonne
		3 :	valeur douteuse
T1.0 :	données contrôlées visuellement	4 :	valeur fausse
T2.0 :	données qualifiées par étalonnage des capteurs	9 :	valeur manquante (la valeur par défaut est de -9999)

A l'intégration de nouvelles données dans la base, le niveau de traitement est à T0 et le niveau de qualité à 0. La procédure effectuée ensuite est la suivante :

- Vérification automatique que les données se situent dans la gamme des capteurs (tableau 1.1-quatrième colonne), et selon les statistiques définies entre 2002 et 2004 (provenant des mesures effectuées durant les différentes campagnes réalisées dans le port de Boulogne-sur-Mer sur cette période), un niveau de qualité est attribué. Passage au niveau de traitement T0.5
- Validation ou modification si besoin du niveau de qualité par un expert (tableau 1.1-cinquième colonne). Passage au niveau de traitement T1.0
- Suite à la rotation des capteurs et à leur étalonnage en laboratoire, le niveau de qualité peut être modifié. Passage au niveau de traitement T2.0

Les statistiques d'aide à la décision n'ont pas évolué depuis la mise en place de la station, c'est pourquoi nous avons choisi de considérer dans nos travaux l'ensemble des données et écarter uniquement les données dont le niveau de qualité est à 9.

1.4. Conclusion

Dans le cadre de l'évaluation et le management de la qualité des eaux côtières et des rivières, le phytoplancton joue un rôle important comme indicateur à court et long terme du changement de la qualité de l'eau.

Dans ce chapitre, nous avons souligné les caractéristiques de notre zone d'étude du point de vue de son :

- Hydrodynamisme : Notre zone d'étude, la côte française de la Manche orientale aux environs de Boulogne-sur-Mer, est sous l'influence d'une structure frontale qui contrôle les échanges entre les masses d'eaux de la côte et du large et en particulier les apports en nutriments à la fois essentiels et potentiellement limitants pour la dynamique du phytoplancton.
- Hydrobiologie : La dynamique phytoplanctonique telle que définie par Margalef (1978), Reynolds *et al.* (2002) et Wyatt (2014) est présentée. Les cellules de phytoplancton sont capables d'intégrer les perturbations naturelles et humaines induites par l'évolution de leur physiologie. La dynamique phytoplanctonique a une dépendance temporelle forte.

Chapitre 2 : Complétion de données

2.1. Introduction

Le chapitre 1 a introduit les zones d'études et ses facteurs / enjeux, notamment la zone côtière boulonnaise et la station de mesures MAREL-Carnot. Avant l'étape de détection et modélisation d'états environnementaux à partir de ces mesures, il est nécessaire de caractériser les données acquises. Cette étape est primordiale, quel que soit la base de données, afin d'extraire l'information utile et la rendre facilement exploitable. Il est notamment intéressant de réaliser une analyse exploratoire des données pour choisir ou générer des algorithmes de traitement automatique des données.

La figure 2.1 reprend le fil conducteur de ce chapitre avec les étapes-clés conduites pour analyser des séries de mesures multi-capteurs :

1. acquisition des données, - alignement temporel des mesures des divers capteurs
2. caractérisation - analyses statistiques et fréquentielles,
3. régularisation, choix de la fréquence et traitement des données manquantes.

Après avoir mis en évidence la problématique du jeu d'études, plusieurs approches de caractérisation des séries de mesures multi-capteurs à valeurs manquantes sont donc présentées et menées. Les séries à long terme présentent l'avantage de contenir une grande connaissance du milieu étudié. C'est pourquoi des analyses statistiques et de composition de ces séries sont nécessaires pour les qualifier.

Le problème de complétion des données est ensuite abordé. Ce domaine étant vaste et actif, nous avons choisi de ne pas faire une présentation exhaustive de l'ensemble des méthodes mais de sélectionner celles les plus adaptées pour notre application. Ces méthodes seront comparées à celles développées durant la thèse.

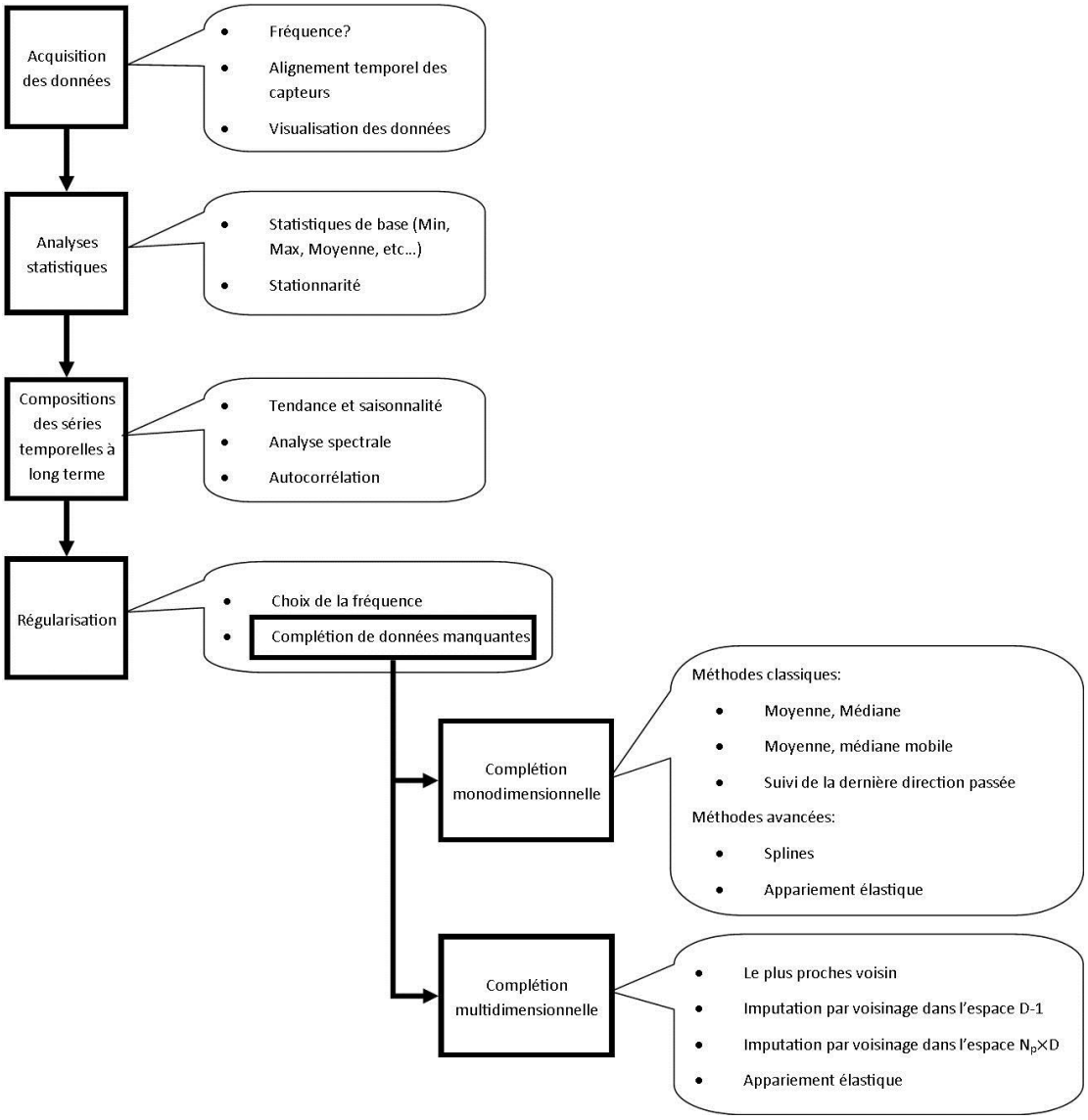


Figure 2.1. Représentation schématique de la structuration du chapitre 2.

2.2. Les données MAREL-Carnot

Les données acquises par la station MAREL-Carnot couvrent principalement la période 2005 à nos jours, l'année 2004 étant la mise en place progressive de certains capteurs et la phase de test. Pour notre étude, nous nous focalisons sur la période 2005- 2009, année 2009 incluse. Ceci représente une base de données de 131 472 instants d'acquisition pour les données physico-chimiques et biologiques (fréquence 20 minutes), et de 7 305 instants pour les concentrations en nutriments (fréquence biquotidienne). Les figures 2.2 et 2.3 illustrent les données réalignées temporellement. Ces paramètres sont présentés en deux groupes au sens du développement phytoplanctonique :

- les paramètres pressions - facteurs de contrôle ;
- les paramètres réponses - effets.

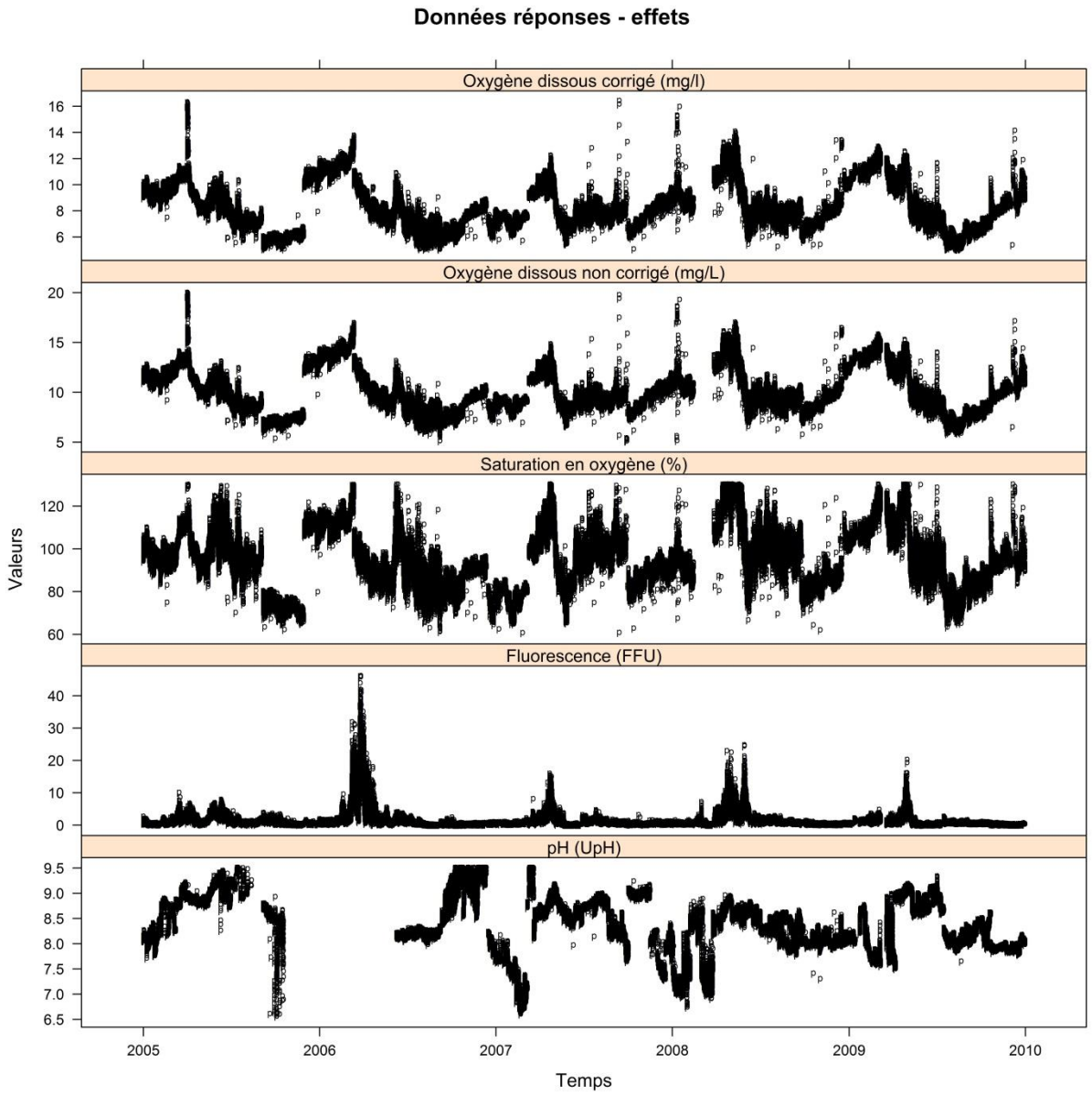


Figure 2.2. Évolution temporelle de l'oxygène dissous corrigé et non corrigé (mg.L^{-1}), de la saturation en oxygène (%), de la fluorescence (FFU) et du pH (UpH) issus de la station MAREL-Carnot au cours de la période 2005-2009.

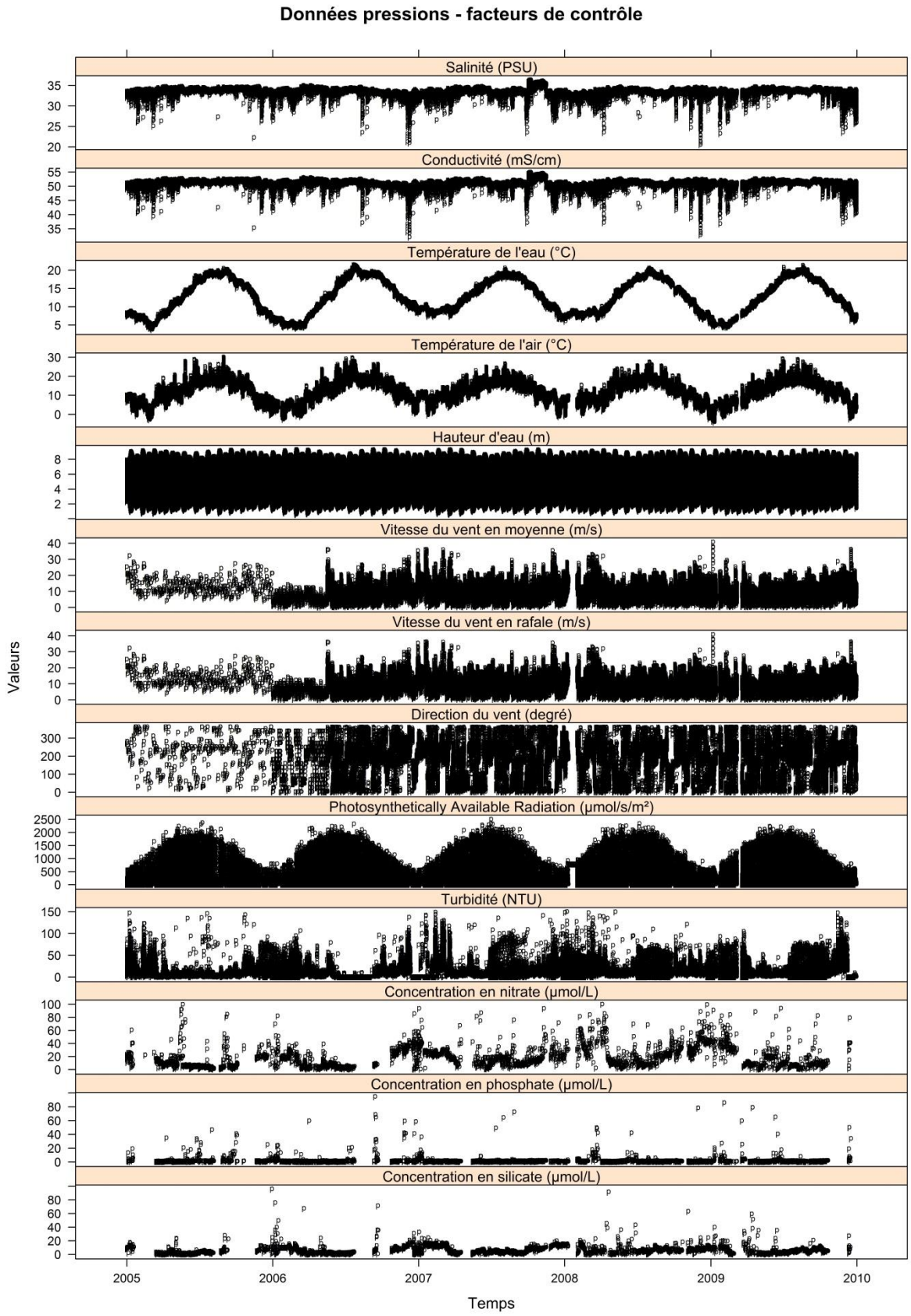


Figure 2.3. Du haut vers le bas : évolution temporelle de la salinité (PSU), la conductivité (mS.cm^{-1}), la température de l'eau et de l'air (°C), la hauteur d'eau (m), la vitesse du vent en moyenne et en rafale (m.s^{-1}), la direction du vent (degré), le PAR (µmol de photons .s^{-1}.m^{-2}),

la turbidité (NTU), les concentrations en nitrate (µmol.L^{-1}), phosphate (µmol.L^{-1}) et silicate (µmol.L^{-1}) issus de la station MAREL-Carnot au cours de la période 2005-2009.

Nous remarquons premièrement que les paramètres sont bruités et que certains possèdent des cycles très visibles comme les températures et le paramètre de rayonnement P.A.R. (de l'anglais Photosynthetically Active Radiation). Deuxièmement, les données possèdent des valeurs manquantes épisodiques ou continues sur des périodes variables. En 2005-2006, aucune donnée de pH n'est disponible pendant un trimestre continu. La courbe concentration en silicate présente de nombreux trous de 2005 à 2009 de taille variable, quelques jours à quelques mois. Après la vérification que les données se situent bien dans la gamme expert*, le pourcentage de valeurs manquantes pour ces cinq années varie de moins de 1 ‰ à plus de 62 % (tableau 2.1).

Tableau 2.1. Nombre et pourcentage de valeurs manquantes pour chaque paramètre pression – facteur de contrôle et réponse - effet de la station MAREL-Carnot dans la période 2005 à 2009 inclus.

Paramètre	Nombres de valeurs manquantes	Pourcentage de valeurs manquantes
Données pression – facteurs de contrôle		
Salinité	16 440	12,50
Conductivité	16 439	12,50
Température de l'eau	18 833	14,32
Température de l'air	16 438	12,50
Hauteur d'eau	1	7.10^{-4}
Vitesse du vent en moyenne	12 343	9,39
Vitesse du vent en rafale	12 343	9,39
Direction du vent	12 417	9,44
P.A.R.	17 501	13,31
Turbidité	17 177	13,07
Concentration en Nitrate	4 376	59,90
Concentration en Phosphate	4 544	62,20
Concentration en Silicate	4 296	58,81
Données réponses - effets		
Oxygène dissous corrigé	21 868	16,63
Oxygène dissous non corrigé	21 814	16,59
Saturation en oxygène	23 764	18,08
Fluorescence	16 182	12,31
pH	35 789	27,22

2.3. Caractérisation

L'analyse statistique des séries temporelles acquises par la station MAREL-Carnot permet d'obtenir des informations qualifiées et précises de l'évolution des phénomènes environnementaux physico-chimiques et biologiques. Dans cette partie, les méthodes usuelles appliquées à ces données sont décrites. Elles sont toutes applicables sur des séries provenant de systèmes à haute fréquence.

2.3.1. Statistiques de base

2.3.1.1. Les statistiques de base et la fonction de densité de probabilité

Les statistiques de base dites descriptives ou exploratoires, permettent de résumer et de synthétiser l'information contenue dans une série afin de mettre en évidence ses propriétés essentielles (minimum, maximum, moyenne, médiane, écart-type, …). Avec une visualisation graphique adaptée (histogrammes de densité, de fréquence et boîte à moustaches), il est possible d'obtenir l'étendue de la série, les gammes de valeurs les plus récurrentes, ainsi que de détecter les valeurs atypiques (extrêmes, erreurs). L'utilisation des histogrammes normalisés sous la forme d'une courbe continue ou discontinue permet d'estimer la fonction de densité de probabilité. On connaît ainsi la probabilité associée à chaque valeur ou intervalle de valeurs d'une variable aléatoire quantitative. Millot (2011) définit dans son ouvrage que les lois de probabilités peuvent être discrètes (loi binomiale, loi multinomiale, loi de Pascal, loi géométrique, loi de Poisson) ou continues (loi normale, loi exponentielle, loi gamma, loi de χ^2, loi de Fisher-Snedecor, loi de Student). Il arrive cependant qu'une variable aléatoire ne suit pas de loi connue, on parle alors de distribution de probabilité (distribution de probabilité de Mann-Whitney ou de Wilcoxon).

L'interprétation des statistiques de base a été conduite sur l'ensemble de la base de données, les résultats figurent en Annexe 1. Nous nous focaliserons uniquement ici sur deux paramètres de la base de données : hauteur d'eau et turbidité. Ces deux paramètres retenus ont été choisis ainsi :

- La hauteur d'eau de forme simple ne possède qu'un instant manquant entre 2005 et 2009. Nous pouvons donc considérer que nous avons une vérité terrain quasi-totale, cas idéal pour qualifier celui-ci et ensuite quantifier sa complétion ;
- La turbidité est un paramètre pression, très structurant des efflorescences.

Exemple 1 : La hauteur d'eau

La hauteur d'eau avec une étendue de 8,74 m possède des valeurs qui varient entre 0,52 et 9,26 m. Sa moyenne est de 4,91 m avec un écart-type σ de 2,19 m. L'erreur standard de la moyenne est quasi nulle $6{,}05.10^{-3}$ m, le nombre d'échantillons N étant grand soit 131 472 mesures. En effet, cette erreur, définie en divisant l'écart-type par la racine carrée du nombre d'échantillons $\sigma/\sqrt{N}$, caractérise la justesse des estimations : plus grande est la taille de notre échantillon, plus la confiance dans les statistiques calculée est forte. (tableau 2.2 et figure 2.4).

L'histogramme en fréquence met en évidence que la forme de sa distribution est voisine de la somme de deux gaussiennes dont les deux maxima locaux se situent respectivement à 2,5 et 7,5 m (figure 2.5).

Tableau 2.2. Statistiques de base de la hauteur d'eau (m) mesurée par la station MAREL-Carnot au cours de la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,52	2,96	4,86	4,91	6,86	9,26	2,19	$6{,}05.10^{-3}$

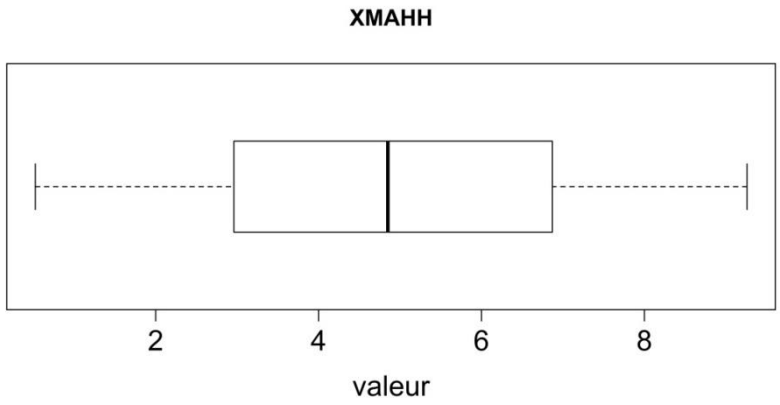


Figure 2.4. Boîte de dispersion de la hauteur d'eau (m) mesurée par la station MAREL-Carnot au cours de la période 2005-2009.

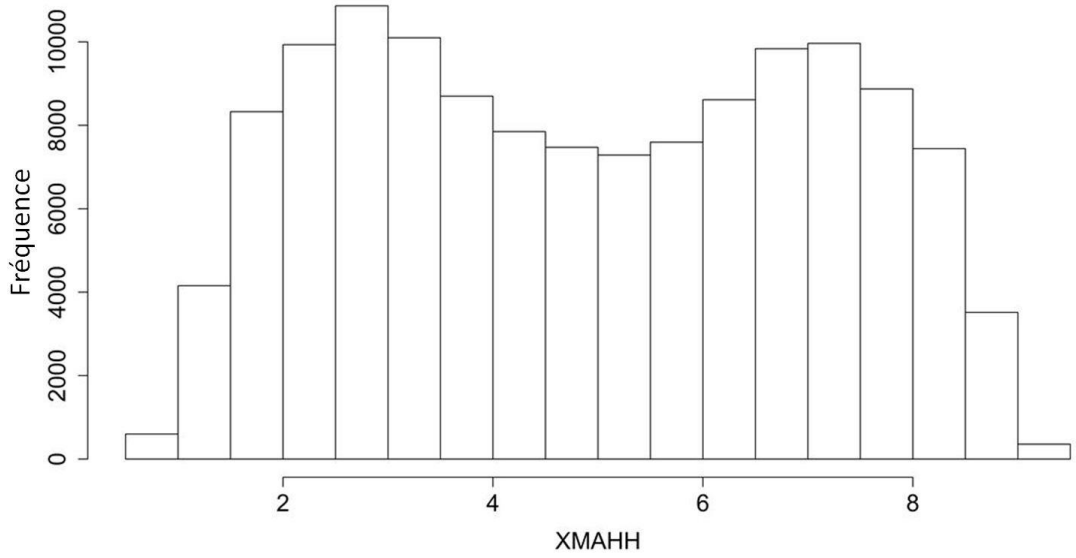


Figure 2.5. Histogramme en fréquence absolue de la hauteur d'eau mesurée par la station MAREL-Carnot au cours de la période 2005-2009.

Exemple 2 : la turbidité

La turbidité est un paramètre dont les valeurs varient entre 0 et 148,9 NTU, soit une large étendue. Sa moyenne se situe près de 12,31 NTU (13 % de données manquantes). Son écart-type apparaît légèrement supérieur à la moyenne, en effet est pris en compte l'ensemble des données notamment les fortes valeurs atypiques illustrées par les ronds de la boîte de dispersion, figure 2.6.

Son erreur standard de la moyenne est faible soit de 0,04 NTU. (tableau 2.3 et figure 2.6). L'histogramme en fréquence sur les données acquises est proche d'une distribution de type χ^2 (p-value*** selon le test de Pearson) (figure 2.7).

Tableau 2.3. Statistiques de base de la turbidité (NTU) mesurée par la station MAREL-Carnot au cours de la période 2005-2009 avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	4,30	7,70	12,31	14,50	148,90	14,27	0,04

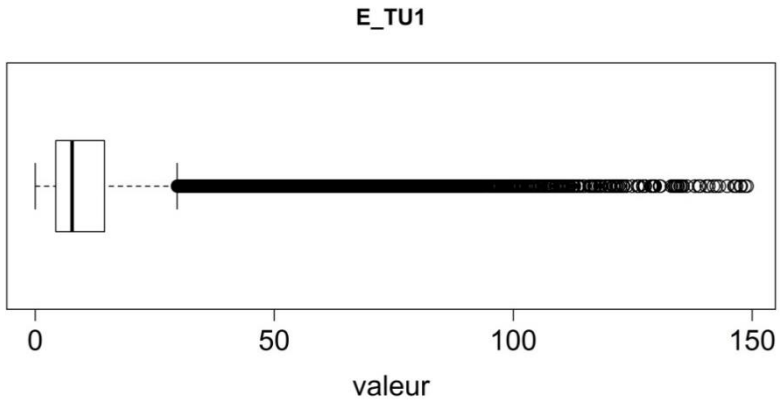


Figure 2.6. Boîte de dispersion de la turbidité (NTU) mesurée par la station MAREL-Carnot au cours de la période 2005-2009.

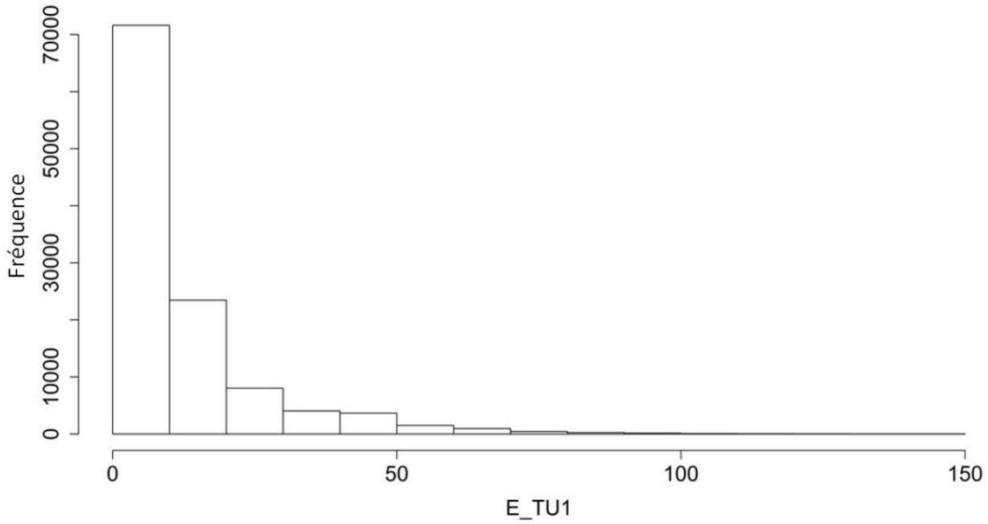


Figure 2.7. Histogramme en fréquence absolue de la turbidité mesurée par la station MAREL-Carnot au cours de la période 2005-2009.

Des précisions supplémentaires sur la base de données, comme la stationnarité ou non des paramètres, peuvent être apportées à partir de ces informations. De plus, la dynamique de chaque série temporelle (tendance, saisonnalité ainsi que les périodes des cycles intra et interannuels s'ils existent) peut être recherchée. Toutes ces informations permettent de mieux apprécier le jeu de données utilisé mais également de réaliser la régression la plus adéquate si elle est nécessaire.

2.3.1.2. Stationnarité d'un signal

On dit qu'un signal $X = \{X_1, X_2, \dots, X_t\}$ est stationnaire si ses propriétés statistiques (moyenne et / ou variance) sont invariantes par translation dans le temps. Il existe deux types de définition de stationnarité (Brockwell and Davis, 2002; Chatfield, 2004) : au sens faible et au sens fort. Ces deux types sont définis comme suit :

1. Stationnarité faible :
   - La moyenne μ et la variance σ^2 sont indépendantes du temps :

Critère 1. $\mu(t) = E[X(t)] = \mu$

Critère 2. $\sigma^2(t) = Var[X(t)] = E\left[X - E[X(t)]\right]^2 = \sigma^2$

   - La fonction d'autocovariance $\gamma(\tau)$ est indépendante du temps pour chaque décalage τ :

Critère 3. $\gamma(\tau) = E\{[X(t) - \mu][X(t+\tau) - \mu]\} = Cov[X(t), X(t+\tau)] = \gamma$

2. Stationnarité forte :

Critère 4. Définit que $\{X_1, X_2, \dots, X_n\}$ et $\{X_{1+h}, X_{2+h}, \dots, X_{n+h}\}$ doivent avoir la même distribution pour tout entier h et $n > 0$.

Il existe différentes méthodes statistiques permettant de vérifier ces deux types de définition. Nous pouvons citer la méthode de Kwiatkowski-Phillips-Schmidt-Shin (KPSS (Kwiatkowski et al., 1992)) (la série test est stationnaire) et la méthode de Dickey-Fuller (la série test est non stationnaire (Dickey et Fuller, 1979)).

Or, le problème de ces méthodes est qu'elles nécessitent des séries complètes. C'est pourquoi, une série haute fréquence (échantillonnage supérieur à une donnée par jour) est souvent réduite à un échantillonnage journalier (moyenne de la journée). Cependant, la réduction de la période d'échantillonnage fait perdre l'information haute fréquence des signaux.
Pour pallier ce problème, nous avons testé les critères ci-dessus indépendamment les uns des autres. L'Annexe 2 reprend les tests de stationnarité au sens faible via une mesure de l'écart-type de chacun des critères. Aucun critère n'est respecté, par conséquent aucune des séries de la station MAREL-Carnot n'est stationnaire au sens faible et donc au sens fort. Cependant, même si un signal est non stationnaire sur l'ensemble de sa partie, il peut localement être stationnaire. Cette partie stationnaire pourra donc être utilisée avec des algorithmes de traitement à fenêtre mobile.

Exemple 2 : la turbidité

La moyenne de la turbidité est calculée sur un intervalle de temps $[1, i]$, avec i allant de 2 à N et est projetée sur la figure 2.8. La moyenne de la série entière, égale à 12,31, est visible en bleu sur la figure. Nous pouvons constater que la moyenne de la turbidité n'est pas constante en fonction du temps et n'est donc pas stationnaire.

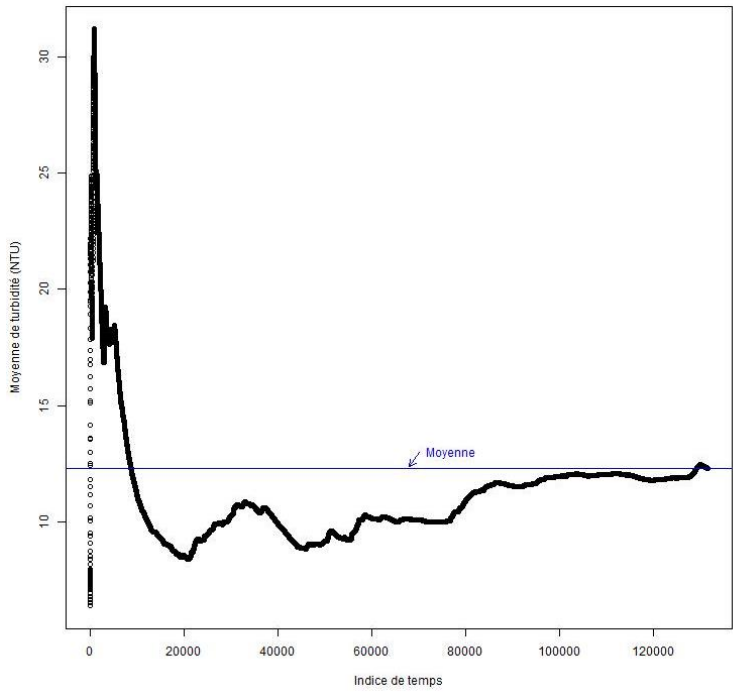


Figure 2.8. Evolution de la moyenne de la turbidité calculée sur un intervalle de temps [1,i], avec i allant de 2 à N. La moyenne de la série entière, égale à 12,31, est représentée par un trait bleu.

2.3.2. Composition de séries temporelles à long terme

2.3.2.1. Analyse de la tendance et de la saisonnalité

Les séries temporelles à long terme possèdent habituellement 3 composantes plus ou moins marquées qui sont (Brockwell and Davis, 2002) :

- une tendance m_t représente l'évolution à long terme d'une série (fonction monotone croissante ou décroissante) ;
- un cycle s_t, ou saisonnalité, est une portion de la série qui se répète de façon régulière. Par exemple, sur une série représentant une année, une période égale à douze représente un cycle mensuel ;
- des résidus Y_t regroupent tout ce qui n'est pas pris en compte par la tendance et la saisonnalité : perturbations irrégulières.

La décomposition d'une série temporelle peut s'effectuer selon deux modèles :

- Additif : $X_t = m_t + s_t + Y_t$;
- Multiplicatif : $m_t \times (1 + s_t) \times (1 + Y_t)$.

A noter que le passage en logarithme d'un modèle multiplicatif permet de le rendre additif :

$$\log\big(m_t \times (1 + s_t) \times (1 + Y_t)\big) = \log(m_t) + \log(1 + s_t) + \log(1 + Y_t)$$

<u>Exemple 1 :</u> La hauteur d'eau

Les données sont moyennées sur la journée. L'analyse de tendance et de résidus permet d'obtenir la figure 2.9. La tendance (« model » sur la figure) est quasi nulle et les résidus possèdent un cycle annuel. Une décomposition des résidus est une solution pour supprimer l'impact lié à ce cycle.

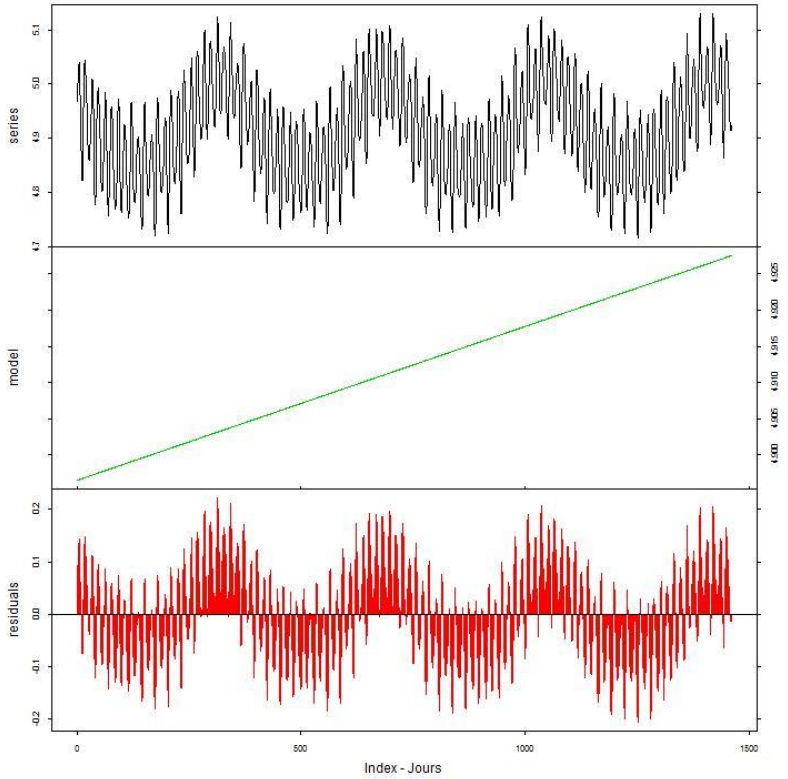


Figure 2.9. Décomposition des données journalières (moyenne) de la hauteur d'eau issue de la station MAREL-Carnot sur la période 2005 à 2008 à partir de la librairie "Pastecs" de R.

Considérant un échantillonnage journalier, l'ensemble de nos signaux sans données manquantes possèdent des cycles propres. La méthode des résidus a été volontairement écartée vu le nombre d'échantillons manquants pour un échantillonnage dit haute fréquence.

2.3.2.2. Analyse spectrale

L'analyse spectrale dite aussi fréquentielle permet également de détecter les périodes caractéristiques d'un signal (cycles). Par ailleurs, elle permet de décomposer ou au moins approximer un signal en somme de signaux élémentaires (constante, sinusoïdes) (Legendre et Legendre, 1998).

Huang et Schmitt (2014) sont repartis de cette approche pour décomposer certains signaux de la station MAREL-Carnot en utilisant une Décomposition Modale Empirique (EMD) suivie d'une transformée de Hilbert-Huang (Huang *et al.*, 1998). Selon les signaux analysés, le nombre de modes (décomposition du signal) diffère. Il est difficile de sélectionner le nombre de modes nécessaires à l'interprétation biologique des efflorescences.

Ces méthodes permettent de s'affranchir des données manquantes contrairement aux méthodes précédentes. Elles multiplient cependant considérablement le nombre de signaux à traiter. Par exemple, le signal d'oxygène dissous est décomposé en 17 modes donc 17 signaux sont à traiter au lieu d'un seul.

2.3.2.3. Autocorrélation

Le calcul de l'autocorrélation fournit une indication importante sur les propriétés d'une série temporelle comme la détermination des fréquences et des amplitudes. Il est ainsi possible de trouver les périodes principales d'un signal à partir d'un corrélogramme. En effet, lorsque le coefficient de corrélation tend vers 1, on peut dire que le décalage τ correspond à une période. Ce coefficient $\rho(\tau)$ est défini via le rapport des fonctions du coefficient d'autocovariance $\gamma(\tau)$ (Chatfield, 2004) :

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} \qquad (2.1)$$

Il est à noter que lorsque le signal est stationnaire, possédant une variance σ^2, le coefficient d'autocorrélation devient :

$$\rho(\tau) = \frac{\gamma(\tau)}{\sigma^2} \qquad (2.2)$$

Si les coefficients sont égaux à zéro, alors ce signal est indépendant du temps.

Exemple 1 : La hauteur d'eau

Le calcul de l'autocorrélation de la hauteur d'eau sur la période 2005-2009 avec un pas de temps de 20 minutes (figure 2.10), montre un décalage de 37 instants pour un coefficient de corrélation quasi égal à 1 (0,996). Ce décalage temporel représente 12 heures. On peut ainsi conclure que le signal de la hauteur d'eau présente un cycle caractéristique de 12 heures, soit la fréquence caractéristique de l'alternance des pleines mers et des basses mers.

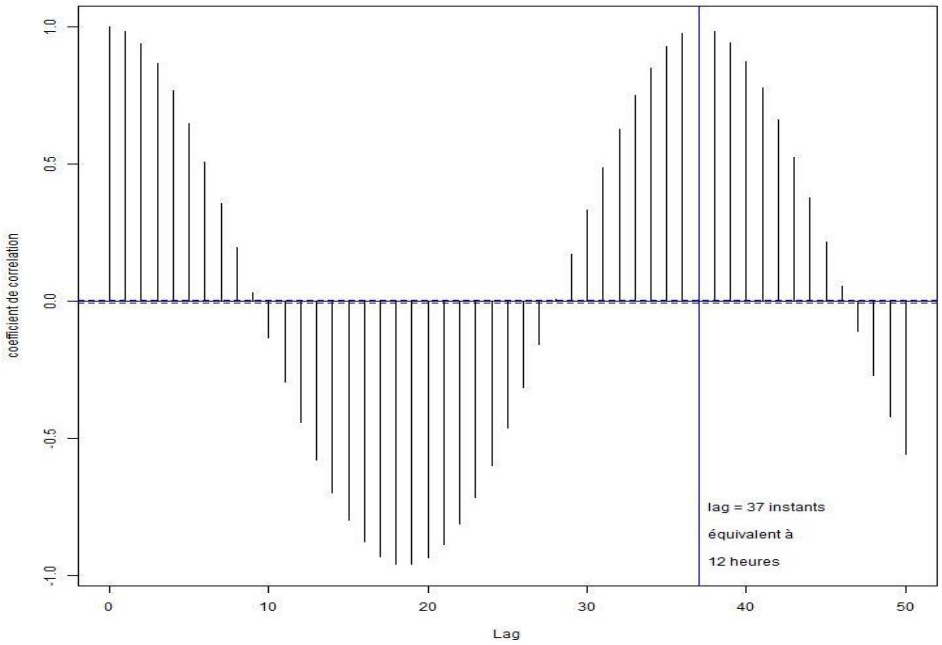


Figure 2.10. Corrélogramme de la hauteur d'eau issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 51 pas de temps.

Selon le seuil accepté d'autocorrélation, ce calcul met en évidence plusieurs cycles caractéristiques. Par exemple, pour un seuil de 0,97 on obtient les cycles suivants : $\tau = \{37; 2050\}$ équivalent à $\{12\ heures, 28\ jours\}$ (figure 2.11). Or, ce signal possède trois cycles d'intérêt connus :

1. 12 heures : alternance Pleine-mer / Basse-Mer ;
2. 28 jours : alternance Morte-eau / Vive-eau ;
3. 6 mois : marée d'équinoxe.

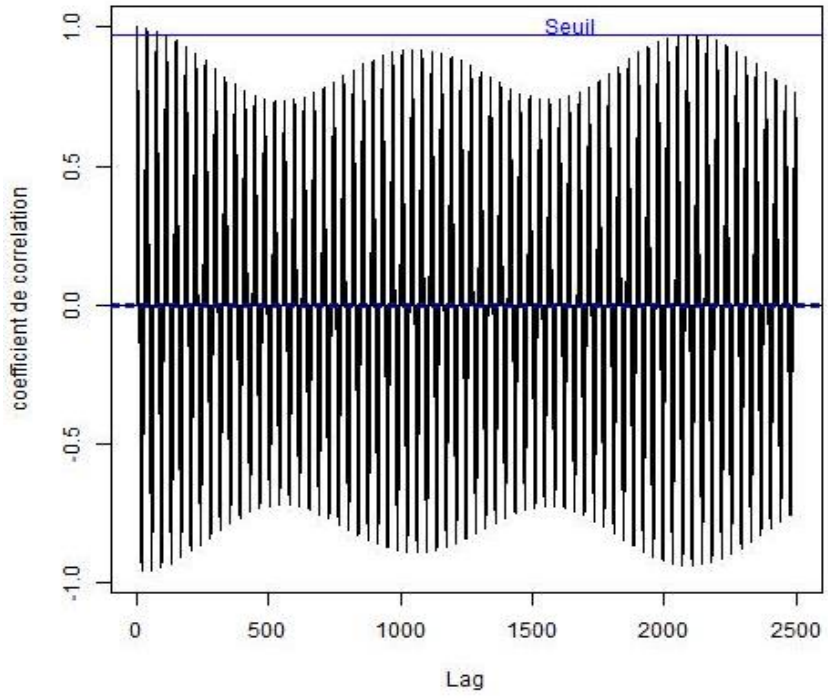


Figure 2.11. Autocorrélation de la hauteur d'eau avec fixation d'un seuil à 0,97.

2.4. Complétion de données

La plupart des études des séries temporelles s'effectuent sur des séries régulières. Cependant, les valeurs manquantes sont un problème récurrent dans les bases de données, l'information fournie est donc incomplète et les analyses sont moins fiables. Ces données manquantes peuvent apparaître lors des maintenances, d'une panne des instruments de mesures ou des appareils de transmission des données.

La régularisation de séries temporelles permet de travailler sur un pas de temps régulier ou de compléter les valeurs manquantes (Grosjean et Ibanez, 2002). Lorsque l'on travaille sur des systèmes basses fréquences avec un échantillonnage mensuel, il est simple à partir d'une régression linéaire ou polynomiale d'ordre 2 de compléter la série. Le problème survient lors de la complétion de séries haute fréquence. Pour le système MAREL-Carnot avec un pas de temps d'échantillonnage de 20 minutes, l'absence d'une journée équivaut à 72 points, une semaine 504 points et un mois 2 200 points environ. Ajoutée à cela la variabilité (et le bruit) due à la haute fréquence, la complétion devient complexe.

Plusieurs solutions sont envisageables pour faire face aux données manquantes :

1. Retirer l'ensemble des instants où il y a au moins une donnée manquante : les analyses sont donc réalisées sur les données valides ;
2. Utiliser des données provenant d'un autre appareil (réseau basse fréquence à proximité par exemple) ;
3. Réaliser une imputation simple qui consiste à remplacer une donnée manquante par rapport à sa série. Le remplacement des données peut se faire par deux types de méthodes :
    - Méthodes classiques :
        - Moyenne, médiane ;
        - Moyenne mobile, médiane mobile ;
        - Suivi de la dernière direction passée.
    - Méthodes avancées :
        - Spline ;
        - Appariement élastique.
4. Réaliser une imputation multiple qui utilise l'ensemble des D paramètres de façon conjointe et non indépendamment les uns des autres afin de compléter au mieux les données de taille $Np \times D$:
    - Imputation par le plus proche voisin dans l'espace $D-1$, espace où la dimension de la valeur manquante a été enlevée ;
    - Imputation par voisinage dans l'espace $D-1$ réduit par classification non supervisée ;
    - Imputation par voisinage dans l'espace $Np \times D$ d'une base réduite par classification non supervisée ;
    - Appariement élastique.

Glasson-Cicognani et Berchtold (2010) ont comparé quelques-unes de ces méthodes entre elles : moyenne et médiane (mobile ou non), k plus proches voisins. Nous avons étendu la comparaison à l'ensemble des méthodes ci-dessus.

2.4.1. Protocole de comparaison de ces méthodes.

Afin de déterminer la complétion la plus optimale, les méthodes ont été testées sur les données de hauteur d'eau issues de la station MAREL-Carnot (XMAHH) sur la période 2005 à 2008 inclus (exemple 1 de la section 2.3.1.1). Le nombre d'instants N_p est de 105 192. Une séquence complète de $N_s = 150$ points a été supprimée pour simuler des données manquantes à l'indice $t = 52\,596$ correspondant au milieu de la période 2005-2008. (figure 2.12).

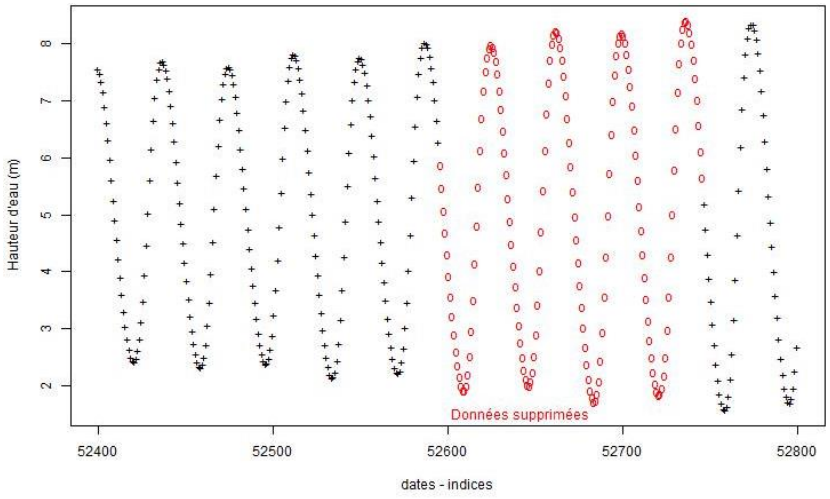


Figure 2.12. Suppression d'une séquence de 150 données (en rouge) de la hauteur d'eau mesurée par la station MAREL-Carnot (XMAHH, mètre) sur la période 2005-2008 à l'indice $t = 52\,596$.

Afin d'estimer la qualité de la reconstruction, trois critères usuels de la littérature sont utilisés sur la portion complétée de taille N_s :

- Coefficient de détermination R^2 ;
- Erreur quadratique ;
- Similarité.

Le coefficient de détermination R^2 est un indicateur qui permet d'apprécier la qualité d'une régression. Il se calcule comme le carré de coefficient de corrélation entre deux variables : il faut donc prendre en compte la p-value du coefficient de corrélation.

L'erreur quadratique normalisée (2.3) estime le ratio de l'aire comprise entre la base de connaissance (données d'origine, noté X) et la base de données reconstruite (noté $Y = \hat{X}$) sur l'aire de X, définie par :

$$Err(Y,X) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y(i)-X(i)\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}X(i)^2}} \qquad (2.3)$$

La reconstruction est jugée :

- Satisfaisante lorsque l'erreur tend vers zéro ;
- Mauvaise lorsque l'erreur tend vers un ;
- Très mauvaise lorsque l'erreur est supérieure à un.

La similarité 2.4 évalue la correspondance entre les deux signaux de longueur N_s fonction de l'inverse de l'aire entre les deux courbes :

$$Sim(Y,X) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\big(Y(i)-X(i)\big)^2} \qquad (2.4)$$

La similarité tend vers un lorsque les courbes sont identiques et tend vers zéro lorsque les amplitudes sont fortement différentes.

2.4.2. Imputation simple

L'imputation simple consiste à remplacer une valeur manquante par une valeur existante du signal considéré ou à partir d'une équation. Nous allons présenter plusieurs stratégies d'imputation monodimensionnelle, des plus simples aux plus avancées que nous comparerons entre elles. Toutes les méthodes présentées ont leur référence basée sur les valeurs antérieures aux données manquantes. Il est facile de généraliser ceci aux valeurs futures ou l'ensemble complet de la base disponible.

2.4.2.1. Les méthodes classiques

Un signal est linéaire lorsque celui-ci évolue de façon constante dans la même direction. Cependant, les signaux environnementaux sont rarement linéaires sauf sur de petites portions. Les méthodes utilisées ici sont toutes construites à partir du même algorithme (algorithme 2.1). Chaque méthode consiste à remplacer la valeur manquante située à l'indice temporel j par une valeur calculée ou extraite sur une fenêtre $intervalle$ passée considérée. Lorsque la donnée $x(j)$ est présente, elle est recopiée dans le signal $y(j)$ et la fenêtre est mise à jour. Les différences entre chaque méthode apparaissent au niveau de l'opérateur et de l'intervalle.

Algorithme 2.1. Algorithme principal utilisé comme base pour les méthodes décrites dans cette partie.

Entrée :
    x la série
    T la taille de la fenêtre considérée
Sortie : y
Variables :
    $Début = 1$
    $Fin = T$
Pour j allant de 1 à la longueur de la série x
    Si x(j) non manquant

$y(j) = x(j)$
$Début = j - T$
$Fin = j - 1$
Sinon
$intervalle = f(Début, Fin, j, T)$
$y(j) = opérateur(x(i), i \in intervalle, x(i) \neq NA)$

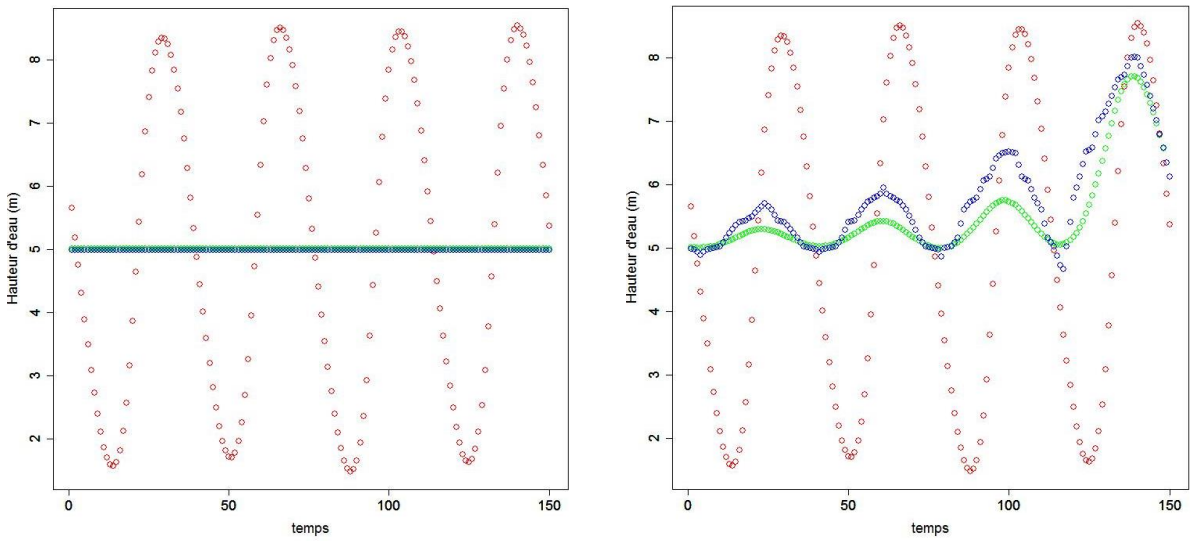
2.4.2.1.1. Imputation par moyenne ou médiane passée

Cette méthode consiste à remplacer la valeur manquante par l'opérateur Moyenne ou Médiane sur une fenêtre de taille T, fixe pour tous les points manquants successifs.
Ces deux méthodes d'imputation ne prennent pas en compte la croissance ou décroissance de la courbe autour du point manquant considéré. Par conséquent, les reconstructions ne sont pas performantes dans notre cas. Pour ces deux méthodes, la similarité entre la portion du signal original et celle complétée est égale à 0,48 avec une erreur quadratique de 0,44 (tableau 2.4).

2.4.2.1.2. Moyenne mobile ou médiane mobile

Cette méthode est une extension de la précédente. La différence se situe dans la mise à jour de la fenêtre considérée. Elle n'est plus figée mais se décale d'un pas lorsqu'une séquence de données manquantes est à compléter. On peut considérer que celle-ci décroit car les valeurs manquantes se situant dans la fenêtre n'interviennent pas dans le calcul. A l'indice j de la donnée manquante, l'intervalle vaut $[j - T, j - 1]$.
L'application de ces méthodes sur des séquences de données manquantes ne permet toujours pas de prendre en compte la dynamique du signal (croissance, courbure) mais réduit l'erreur d'approximation puisqu'elle diminue le voisinage passé considéré du point manquant. Les résultats des calculs d'erreurs pour ces reconstructions sont faibles (tableau 2.4). En effet, pour l'imputation par la moyenne mobile, le coefficient de détermination est égal à 0,12*** avec une erreur quadratique de 0,40 et une similarité de 0,42. Pour l'imputation par la médiane mobile, ces résultats sont respectivement de 0,10***, 0,41 et 0,44.

Malgré les mauvaises reconstructions des méthodes de moyenne sur fenêtre fixe ou mobile sur une période de données manquantes longue (figure 2.13), il est pertinent de réaliser une imputation par la moyenne ou la médiane lorsqu'une valeur ponctuelle est manquante en utilisant les deux points encadrants cette valeur.



(a) (b)

Figure 2.13. Reconstruction par les méthodes de moyenne (en vert) et médiane (en bleu) sur fenêtre fixe (a) ou mobile (b) des données supprimées (en rouge).

2.4.2.1.3. Imputation selon la dernière direction passée

Pour pallier l'absence de prise en compte de la dynamique du signal, le calcul de la direction est introduit dans l'estimation du signal. Ce calcul se base sur le calcul de la tangente de l'angle séparant deux points successifs. Ainsi, la valeur manquante est imputée par l'équation $x(j+1) = 2\,x(j) - x(j-1)$ que l'on va démontrer dans un cadre plus général.

Calcul de la direction dans un cadre général :

Soit t+dt2 l'instant manquant, le calcul de la tangente se fait comme suit :

$$\tan(\Theta) = \frac{x(j) - x(j - dt1)}{dt1} \qquad (2.5)$$

$$\tan(\Theta) = \frac{x(j + dt2) - x(j - dt1)}{dt1 + dt2} \qquad (2.6)$$

En développant (2.6), on obtient :

$$x(j + dt2) = x(j - dt1) + (dt1 + dt2) \times \tan(\Theta) \qquad (2.7)$$

En remplaçant $\tan(\Theta)$ de (2.7) par (2.5), cela nous donne :

$$x(j + dt2) = x(j - dt1) + \frac{dt1 + dt2}{dt1} \times \big(x(j) - x(j - dt1)\big) \qquad (2.8)$$

Dans le cas d'un pas de temps régulier (toutes les 20 minutes pour la station MAREL-Carnot) : $dt1 = dt2 = \Delta$

$$x(j + \Delta) = x(j - \Delta) + 2 \times \big(x(j) - x(j - \Delta)\big) \qquad (2.9)$$

Soit

$$x(j + \Delta) = 2\,x(j) - x(j - \Delta) \qquad (2.10)$$

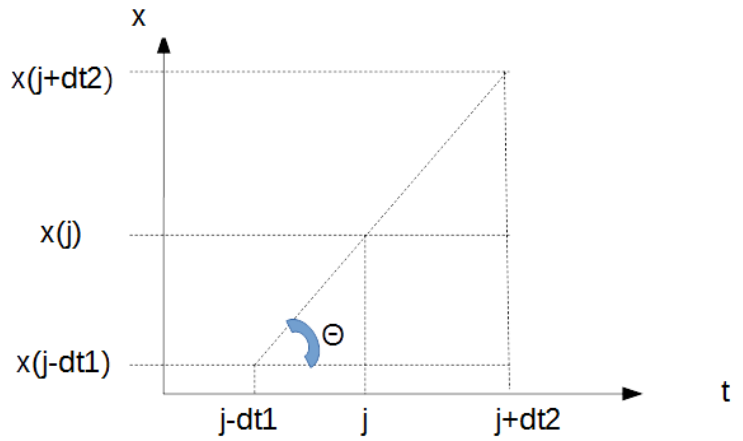
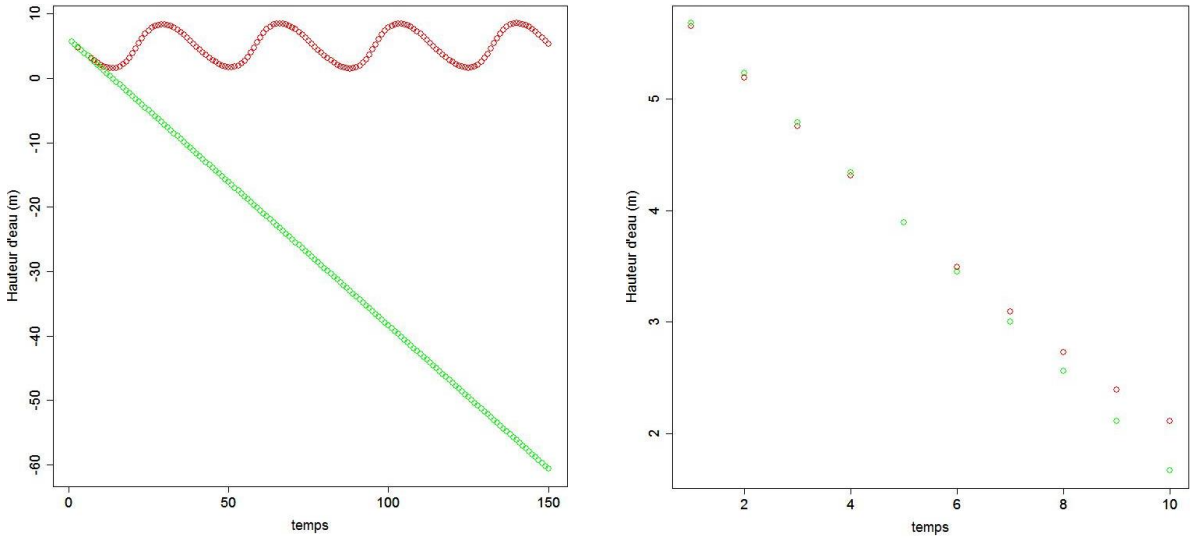


Figure 2.14. Schéma d'aide à la compréhension du calcul, à partir de l'angle Θ, de la direction.

En prenant en compte la dernière direction du signal avant la séquence manquante, il est possible de compléter le signal en conservant cette dernière direction. Cependant, cette complétion n'est valide que pour un signal non bruité ou lissé. En effet, sur un signal bruité la dernière direction calculée ne reflètera pas la direction réelle du signal. Dans l'algorithme 2.1, la fonction *opérateur()* est remplacée par $x(j+1)$ de l'équation (2.10).

Sur notre exemple, cette méthode n'est pertinente que pour des séquences manquantes où la dynamique n'évolue pas : on note ici la présence d'un décrochage (figure 2.15 (a)). En effet, le coefficient de détermination est à 0,03** avec une erreur quadratique forte de 6,83 et une similarité très faible de 0,11 (tableau 2.4). La figure 2.15 (b) illustre le cas d'une séquence sans changement de dynamique, dans ce cas la méthode est appropriée. Sur la hauteur d'eau, la complétion de 10 points permet d'obtenir un coefficient de détermination à 0,99***, une similarité de 0,91, et une erreur quadratique de 0,04 (tableau 2.4).



(a) (b)

Figure 2.15. Imputation selon la dernière direction passée (en vert) sur une séquence (en rouge) possédant une dynamique (a) ou non (b).

2.4.2.2. Méthodes avancées

Nous allons maintenant présenter deux classes de méthodes permettant de prendre en considération la dynamique du signal : soit à partir d'une équation issue d'une étape de régression, soit en recherchant une séquence de dynamique fidèle à celle précédant la séquence manquante.

2.4.2.2.1. Spline

La méthode des splines est à la base une méthode de régression. Ainsi, un nuage de points peut être approximé par un polynôme de degré n défini à l'équation (2.11). La section comportant des valeurs manquantes est imputée par le calcul de ce polynôme à chaque instant manquant.

$$y = a + bt + ct^2 + \cdots + dt^n \qquad (2.11)$$

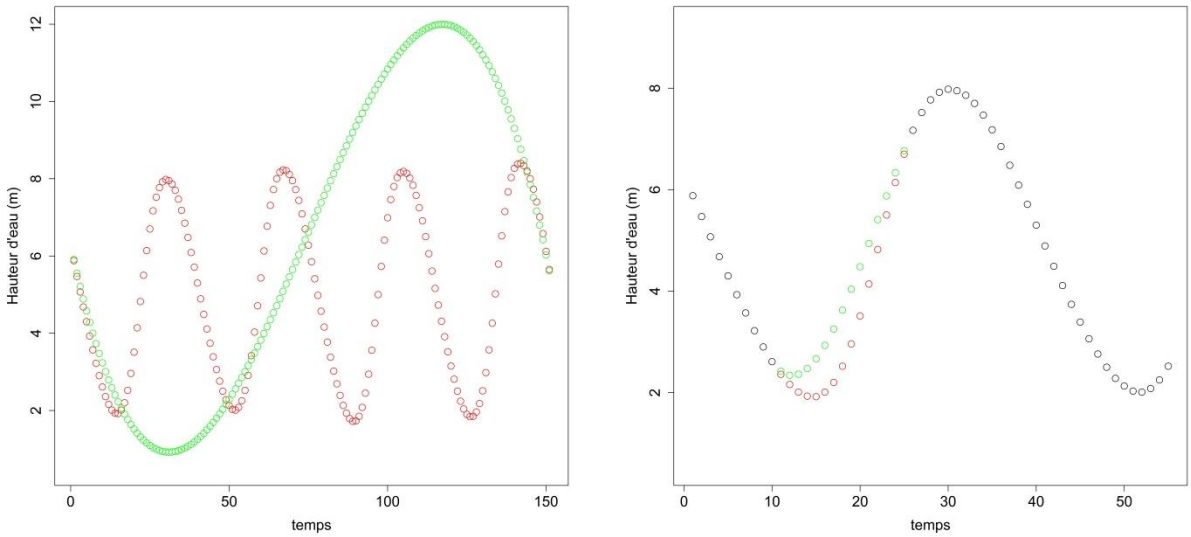
Contrairement aux précédentes méthodes, elle nécessite d'utiliser le voisinage passé et futur de la section manquante. Un polynôme de degré 3 est utilisé afin d'avoir un minimum et un maximum locaux. L'avantage de cette méthode est qu'elle prend en compte les changements de direction ainsi que les points de départ et d'arrivée de cette séquence à compléter.
L'algorithme 2.2 utilisé consiste à toujours recopier le signal dans le cas d'une donnée connue et adapter la fenêtre de voisinage passée et future autour du point manquant. Pour une donnée manquante, l'opérateur spline est utilisé à partir de cette fenêtre de voisinage.

Algorithme 2.2. Algorithme pour l'imputation par la spline cubique.

Entrée :
    x la série
    T la taille de la fenêtre considérée
Sortie : y
Variables :
    $Début = 1$
    $Fin = 1$
    $SéquenceNA = 0$
Pour j allant de 1 à la longueur de la série x
    Si x(j) non manquant
        $y(j) = x(j)$
        $Début = j - T$
        $SéquenceNA = 0$
    Sinon
        $SéquenceNA = SéquenceNA + 1$
        $Fin = j + SéquenceNA + T$
        $y(j) = Spline(x(i), i \in [Début, Fin])$

Quel que soit l'ordre du polynôme dans notre exemple, les oscillations ne sont pas respectées (figure 2.16 (a)). Nous nous retrouvons avec un coefficient de détermination proche de zéro (avec une p-value à 0,854), avec une erreur quadratique de 0,53 et une similarité de 0,41 (tableau 2.4).

Comme précédemment sur une séquence avec un seul changement de dynamique (figure 2.16 (b)), la spline permet d'approximer de façon satisfaisante (la fluctuation du signal est bien prise en compte). Ceci est confirmé par le coefficient de détermination, la similarité et l'erreur quadratique respectivement à 0,95***, 0,84 et 0,19 (tableau 2.4).



(a) (b)

Figure 2.16. Imputation par spline (en vert) sur une séquence (en rouge) avec plusieurs (a) et un seul (b) changement de dynamique.

Le tableau 2.4 reprend l'ensemble des résultats de reconstruction de la séquence manquante du signal de la hauteur d'eau par les méthodes d'imputation simple et spline. Aucune de ces méthodes n'atteint une similarité convenable : toutes inférieures à 0,5 pour la séquence de $N_s = 150$ points, ce qui représente seulement 2 jours de données manquantes. Les figures 2.2 et 2.3 ont montré des zones de trous nettement plus importantes.

Toutes ces méthodes sont limitées dans la complétion de signaux fluctuants. C'est pourquoi nous proposons la méthode basée sur l'appariement élastique.

Tableau 2.4. Récapitulatif des résultats de complétion (R^2, similarité, erreur quadratique) pour chaque méthode d'imputation simple.

Méthode	R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
Moyenne (vert)	NA	0,48	0,44
Médiane (bleu)	NA	0,48	0,44
Moyenne mobile (vert)	0,12***	0,42	0,40
Médiane mobile (bleu)	0,10***	0,44	0,41
Selon la dernière direction passée (avec variation)	0,03**	0,11	6,83
Spline (nombre de variation > 2)	$2{,}23.10^{-4}$ (0,85)	0,41	0,53
Cas particuliers : séquence réduite			
Selon la dernière direction passée (sans variation)	0,99***	0,91	0,04
Spline (nombre de variation $\leq$ 2)	0,95***	0,84	0,19

2.4.2.2.2. Complétion par appariement élastique

2.4.2.2.2.1. Définition

L'appariement élastique ou Dynamic Time Warping (DTW) est une méthode initiée par Sakoe et Chiba (1978). Elle consiste à calculer une distance géométrique entre deux courbes afin de vérifier leur similarité. La méthode accepte une dilatation temporelle et des déformations locales, c'est-à-dire qu'il est possible que les deux courbes ne soient pas de la même longueur. L'algorithme consiste à rechercher la correspondance F entre les paires de points qui minimise une distance euclidienne, c'est-à-dire le coût global de ressemblance qui est défini comme une somme d'écarts d'intensité entre les points appariés. La distance D entre les deux courbes se calcule comme suit (Sakoe et Chiba, 1978) :

Soient la courbe $A = a_1, a_2, \dots, a_i$ et la courbe $B = b_1, b_2, \dots, b_j$ définies sur $\mathbb{R}$ avec i et j pouvant être différents. Le décalage temporel entre les courbes peut être quantifié par $F = \{c(1), c(2), \dots, c(k), \dots, c(K)\}$ l'ensemble des couples appariés où $c(k) =$

$\big(i(k), j(k)\big)$ est la combinaison des indices temporels de chaque courbe et $w(k)$ est le vecteur de pondération du k[ème] appariement. La distance entre les deux courbes est alors :

$$D(A,B) = \min_{F} \left[\frac{\sum_{k=1}^{K} d\big(c(k)\big) \times w(k)}{\sum_{k=1}^{K} w(k)} \right] \qquad (2.12)$$

Les conditions à respecter pour cet appariement sont :

- Appariement des premiers points de chaque signal entre eux ;
- Appariement des derniers points de chaque signal entre eux ;
- Pour les autres points, la matrice de coût est calculée et les affectations se font pour les appariements dont les coûts sont minimaux sans accepter les retours en arrière. Ainsi, lors de la correspondance entre les points, il n'est imposé aucun chevauchement de liaison, c'est-à-dire que si a_k est lié avec b_{k+1} alors a_{k+1} ne pourra être lié qu'avec b_k.

Le calcul du chemin F dans l'espace bidimensionnel $\big((a_1, b_1), \vec{A}, \vec{B}\big)$ rassemblant les similarités les plus élevées permet d'obtenir la correspondance sur l'ensemble du signal (figure 2.17). Cet espace bidimensionnel peut être visualisé comme un graphe pondéré par un coût de déplacement entre chaque point des deux signaux. Chaque nœud du graphe correspond à un coût d'appariement de ces points.

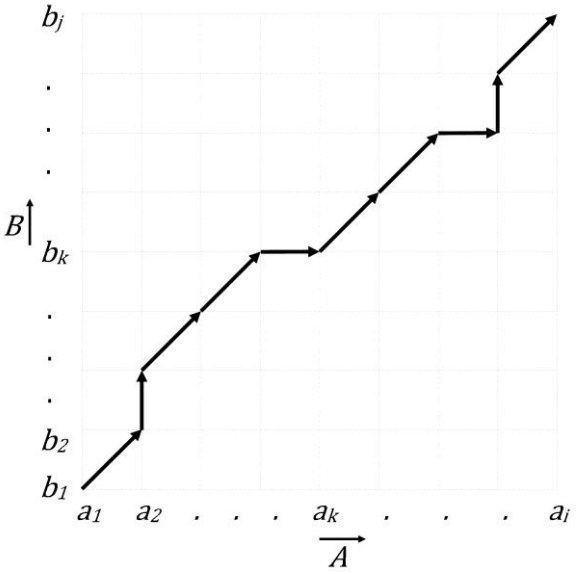


Figure 2.17. Calcul du chemin dans l'espace bidimensionnel des couples (i, j).

Afin que cette distance soit facilement interprétable, Caillault *et al.*, (2009) introduisent une mesure alternative de dissimilarité normalisée entre 0 et 1. Différents types d'appariement élastique (figure 2.18) existent selon les contraintes imposées (Caillault *et al.*, 2009) :

- L'appariement linéaire, l'algorithme impose qu'un point du signal A est apparié au point de B le plus proche temporellement. Si les deux signaux ne sont pas de la même longueur, un échantillonnage sur le signal le plus court est réalisé afin que le nombre de points soit similaire. Le coût associé est la distance temporelle minimale.

- L'appariement élastique libre, l'algorithme impose une distance minimale entre les valeurs.
- L'appariement élastique restreint possède des contraintes à la fois temporelles et d'amplitudes, afin d'éviter que l'appariement ne se fasse sur des points trop éloignés les uns des autres mais dont les valeurs sont cohérentes.

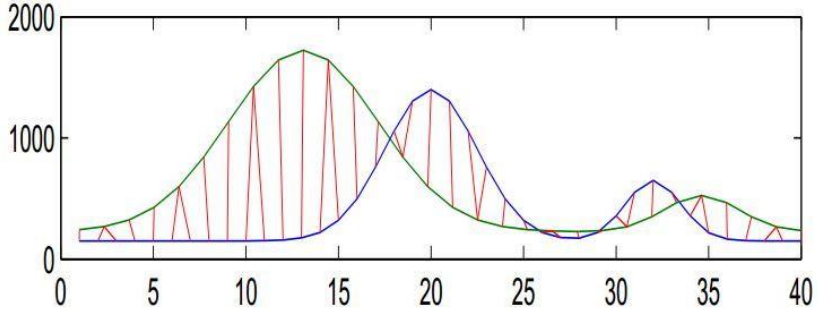
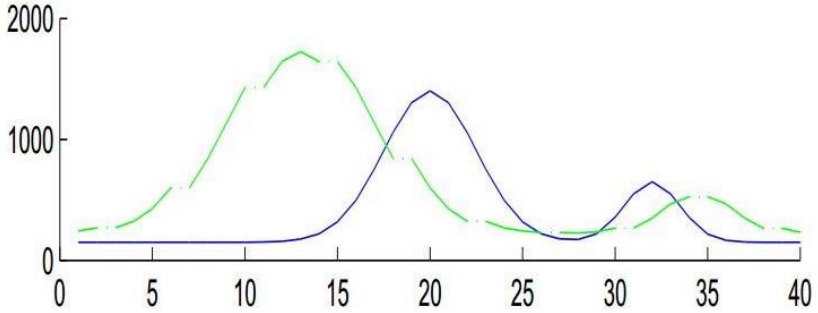
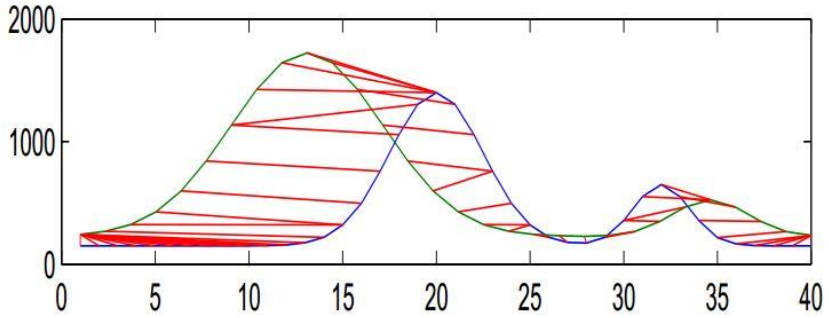
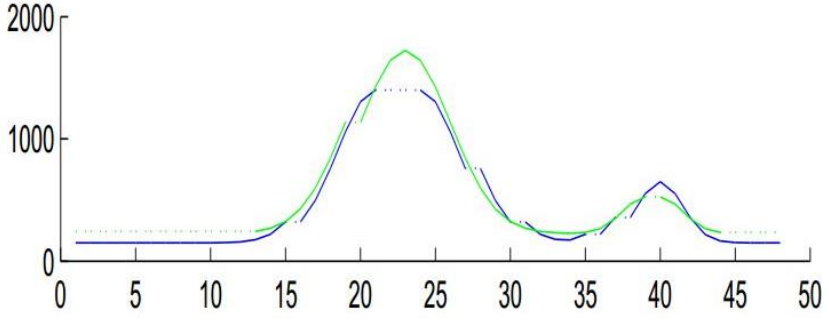
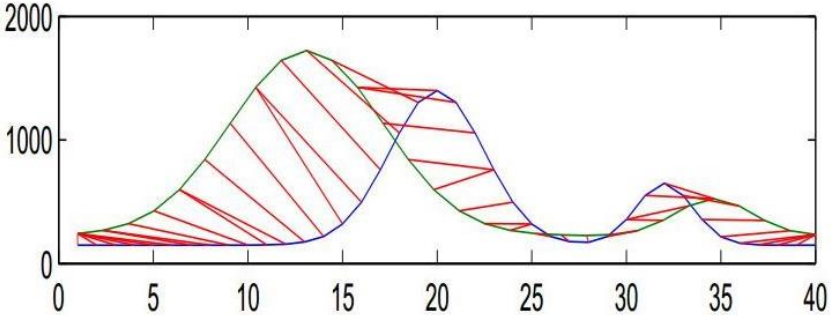
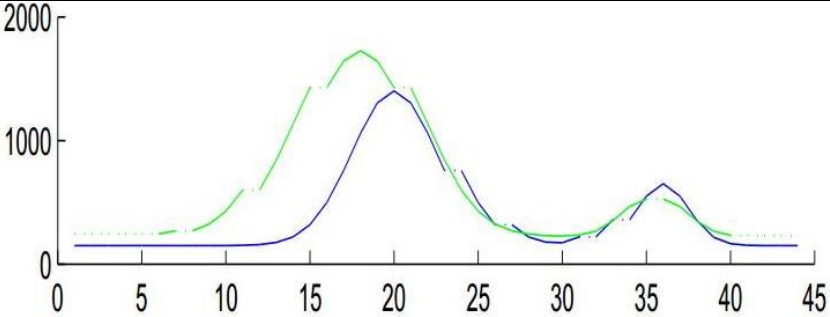
2.4.2.2.2.2. Travaux existants

L'appariement élastique est très utilisé dans les domaines de la reconnaissance de la parole ou d'écriture. Sakoe et Chiba (1978) initie le calcul de cette distance élastique dans la reconnaissance des mots parlés. En effet, un même mot peut être prononcé différemment (exemple : « thèse » et « thèèèse »). C'est pourquoi l'appariement élastique permet de retrouver la similarité entre les différentes prononciations du mot. Ce qui n'est pas possible avec la contrainte d'association linéaire au même instant (appariement linéaire).

En reconnaissance d'écriture, Rath et Manmatha (2003) ont utilisé des images de mots dans leur expérience et ils ont montré que l'appariement élastique était une méthode performante pour prendre en compte la variabilité spatiale du mot.

Le coût d'appariement DTW avec une distance Euclidienne est aussi utilisé en classification de données. Notamment, Petitjean *et al.* (2011) insèrent cette matrice de coût dans l'algorithme de classification non supervisée K-means (Hartigan et Wong, 1979) à la place de la distance Euclidienne couramment utilisée. Ils ont baptisé cette méthode « DTW barycenter averaging » (DBA).

Figure 2.18. Représentation des différents types d'appariements avant et après déformation.

Types	Appariement	Appariement après déformation
Linéaire		
Libre		
Restreint		

2.4.2.2.2.3. Approche de complétion par appariement élastique

La connaissance de la dynamique des données précédant une séquence de valeurs manquantes est le point clé de cette méthode. En effet, l'appariement élastique est utilisé pour la recherche dans la série d'un profil (P) similaire à la requête (R) de taille T, cette requête R étant la séquence précédent la ou les valeurs manquantes. Afin d'avoir la déformation la moins importante possible, l'appariement élastique restreint avec une contrainte temporelle à deux instants est utilisée. Une fois le profil déterminé, la séquence suivant celui-ci est recopiée à l'emplacement des valeurs manquantes (algorithme 2.3).

Algorithme 2.3. Algorithme de complétion par appariement élastique.

Acquisition des paramètres à l'instant t
Si détection d'un paramètre manquant à l'instant t
    Construction d'une requête R :
        Fenêtre temporelle précédent la donnée manquante [t-T-1, t-1], T la taille de la fenêtre
    Comparaison de la fenêtre à la base de connaissances par fenêtre glissante
    Calcul du taux de déformation entre la requête R et la fenêtre de la base analysée, profil P
    Si taux faible, remplacement de la donnée à l'instant t par le vecteur de paramètre suivant P

La recherche du profil P similaire se fait grâce à une fenêtre glissante, ayant la longueur de R, parcourant l'ensemble de la série (figure 2.19).

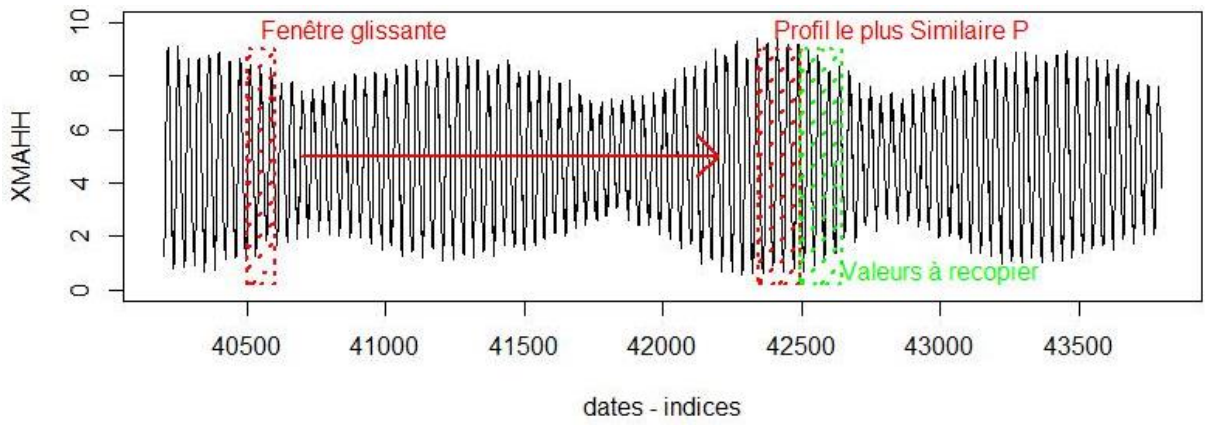


Figure 2.19. Recherche de la portion P par appariement élastique grâce à une fenêtre glissante (en rouge) et détection des valeurs à recopier (en vert).

Afin d'optimiser le temps de calcul, un premier calcul du critère de coût est réalisé sur la première moitié de la séquence à tester. Si ce coût, donc la distance D définie par Sakoe et Chiba (1978) est inférieure à un seuil, alors ce coût est calculé pour l'ensemble de la séquence. Ce seuil est pré-calculé comme étant la distance minimale obtenue sur les dix

premières séquences de profil analysé. Le profil retenu est celui dont le coût de ressemblance avec la requête est le plus bas sur l'ensemble de la série (figure 2.20 et 2.21). La figure 2.20 illustre le chemin *F* associé à notre contrainte de restriction temporelle schématisée par les pointillés. La figure 2.21 montre les paires d'appariement entre la requête et la fenêtre glissante, ici correspondant au profil le plus similaire à la requête.

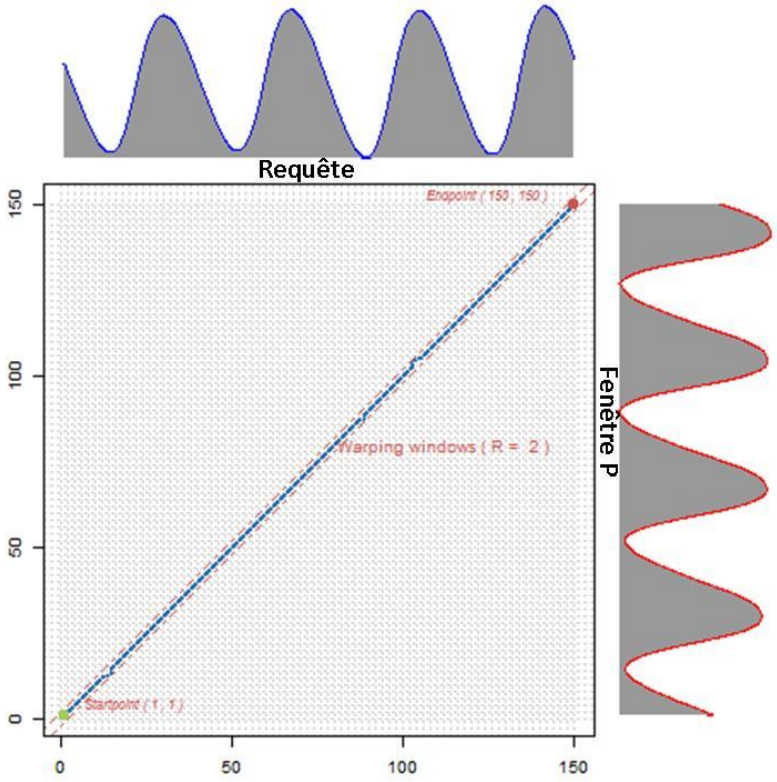


Figure 2.20. Espace bidimensionnel pour le calcul du taux de déformation entre la requête R et la fenêtre P.

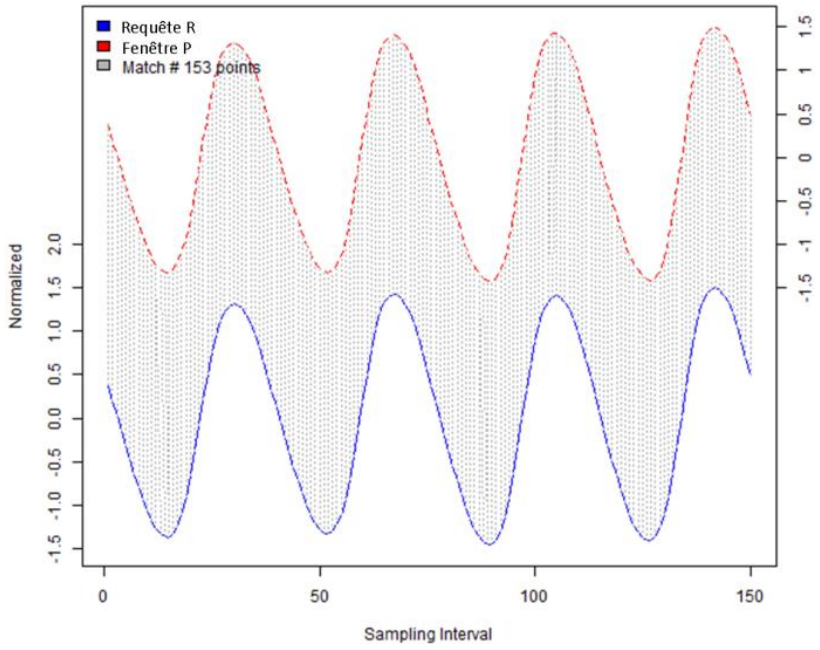


Figure 2.21. Correspondance entre les deux signaux R et P avec le nombre de points appariés.

Les valeurs manquantes sont complétées par la copie des données situées à la suite du profil P (figure 2.22).

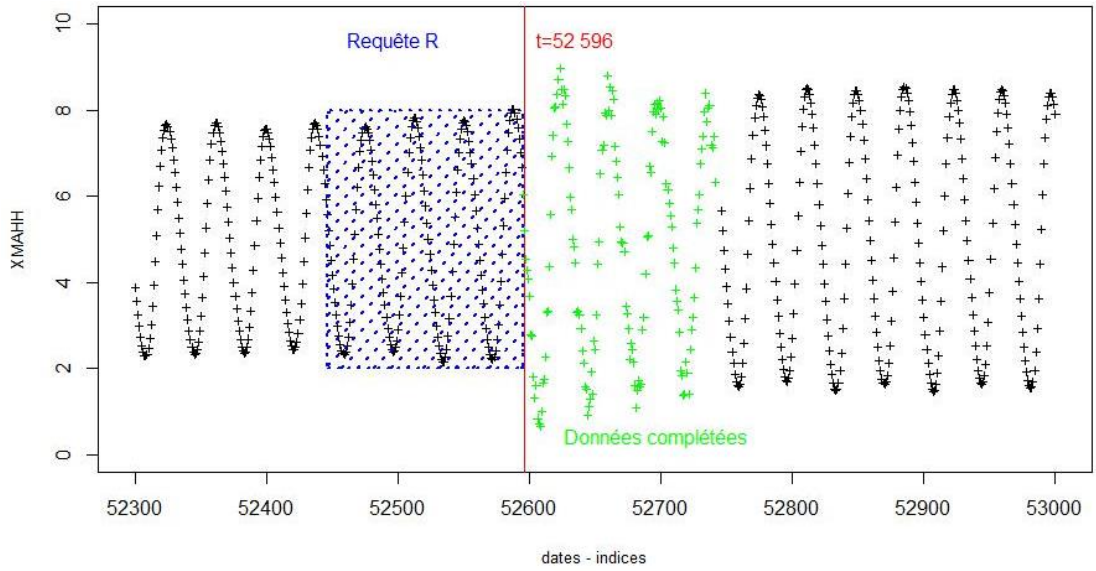


Figure 2.22. Recopie des valeurs (en vert) situées après la portion P à la place des données manquantes à l'instant t=52 596.

Cette méthode n'est applicable que si l'ensemble des événements est acquis sur la série x. C'est pourquoi, il est préférable de travailler sur de grandes bases de données afin d'intégrer un maximum de variabilités du phénomène étudié.

Protocole de test :

Afin de tester la robustesse de cette méthode, quatre expériences ont été réalisées :

E1. La requête R et la série x non bruitées : correspond au test de la série brute de la hauteur d'eau. Les signaux sont rarement aussi lisses, on peut donc assimiler ce signal à un signal filtré.

E2. Requête R bruitée uniquement : ce test peut être assimilé à un filtrage de la série complète x. La requête n'est pas filtrée pour conserver la variabilité existante avant les données manquantes.

E3. Série x bruitée uniquement : par analogie, on peut dire qu'un filtre a été appliqué sur la requête pour améliorer les chances de correspondance.

E4. Requête R et série x bruitées : cette expérience équivaut à tester un cas réaliste de signal totalement bruité.

Le bruit ajouté à la hauteur d'eau équivaut à un ajout d'une nouvelle valeur aléatoire comprise dans l'intervalle $[-0{,}5; 0{,}5]$ à chaque mesure. L'équation en langage R de l'ajout du bruit sur la série x pour chaque instant t s'écrit : $x(t)\ bruitée = x(t) + runif(1, -0.5, 0.5)$. De plus, chacune de ces expériences est testée deux fois :

- La série x précède la séquence à valeurs manquantes (dates : 2005-2007, noté x_{Av})
- La série x succède la séquence à valeurs manquantes (dates : 2007-2008, noté x_{Ap})

En utilisant les mêmes critères que précédemment, analysons maintenant la qualité de l'appariement entre la requête R et le profil P le plus proche :
Sans bruit additionnel (E1), le profil le plus proche dans la série avant la requête (dates : 2005-2007) a une similarité de 0,74 et 0,79 pour le profil trouvé dans la base après la requête (2007-2008) avec une erreur quadratique inférieure à 0,08. Avec un bruit additionnel (E4), le profil le plus proche dans la base avant la requête (2005-2007) a une similarité de 0,62 et 0,69 pour le profil trouvé dans la base après la requête (2007-2008) avec une erreur quadratique à 0,15 (figure 2.23, tableau 2.5, Annexe 2). L'ensemble de ces résultats mettent en avant la puissance de l'appariement élastique dans différentes situations. Rappelons que pour les méthodes présentées précédemment, le coefficient de détermination était proche de zéro, contrairement à ici où ces valeurs sont supérieures à 0,9. Par ailleurs, cet algorithme a testé une séquence de plus de 52 000 points en 5 secondes environ (tableau 2.6) sur un processeur Intel i7 à 2,4 GHz.

Tableau 2.5. Résultats du calcul du coefficient de détermination, de similarité et de l'erreur de déformation moyenne pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur la série $\boldsymbol{x_{Av}}$ entre la requête R et du profil P le plus proche.

Expérience sur x_{Av}	R^2	$Sim(R,P)$	$Err(R,P)$
E1	0,99***	0,74	0,08
E2	0,97***	0,64	0,15
E3	0,99***	0,71	0,10
E4	0,96***	0,62	0,15

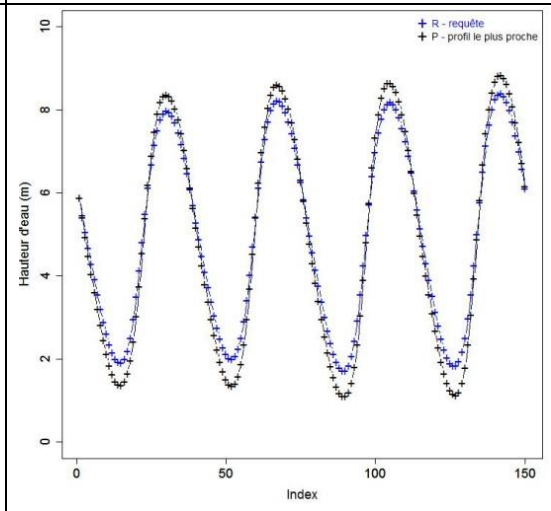
Tableau 2.6. Temps de calcul mis pour réaliser la totalité des appariements élastiques avec la distance associée sur l'ensemble de la séquence.

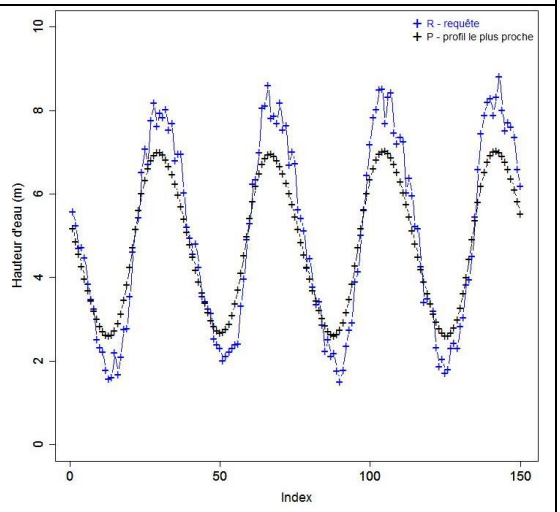
Expérience	Temps de calcul sur la séquence précédent les données manquantes	Temps de calcul sur la séquence succédant les données manquantes
E1	5,039 secondes	5,039 secondes
E2	5,258 secondes	5,241 secondes
E3	5,241 secondes	5,195 secondes
E4	5,216 secondes	5,304 secondes

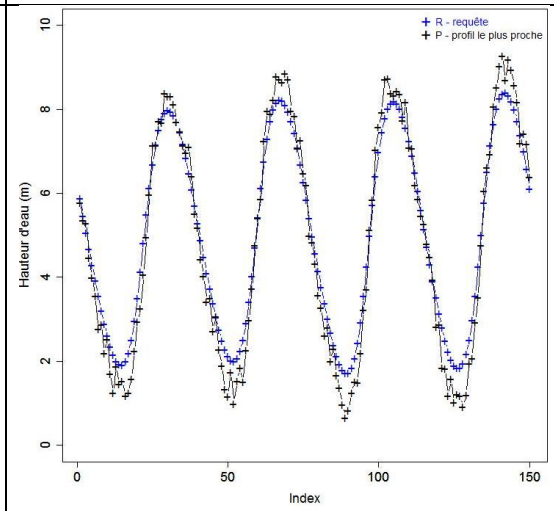
En ce qui concerne la qualité de la complétion, les résultats sont les suivants :
Sur l'ensemble des reconstructions, le coefficient de détermination est supérieur à 0,93***. Les données insérées entraînent une erreur quadratique minimale de 0,08 pour une similarité de 0,74 par rapport aux données d'origine avec la base 2005-2007 (x_{Av}) et une erreur quadratique de 0,07 avec une similarité de 0,78 dans le cas où la base utilisée est x_{Ap} (2007-2008). Si l'on rajoute le bruit (E4), l'erreur quadratique est de 0,12 (la similarité est à 0,68)

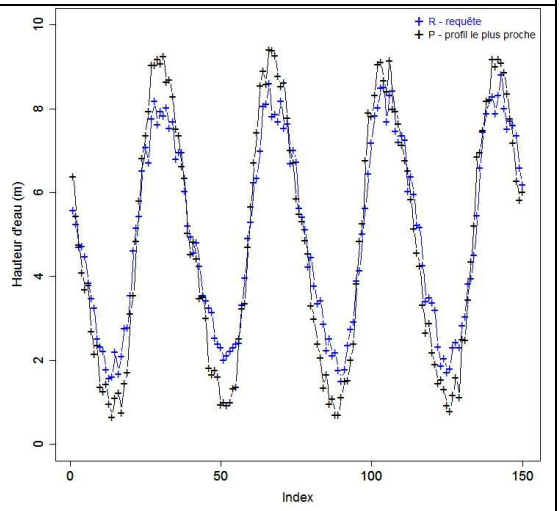
avec la base x_{Av} et de 0,09 (similarité à 0,73) avec la base x_{Ap} (figure 2.24, tableau 2.7 et Annexe 2).

Figure 2.23. Représentation, pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur la série $\boldsymbol{x_{Av}}$ *entre la requête R (en bleu) et du profil P le plus proche (en noir).*

<table>
<tr><th></th><th>Requête R Brute</th><th>Requête R Bruitée</th></tr>
<tr><td>Profil P Brut</td><td> 

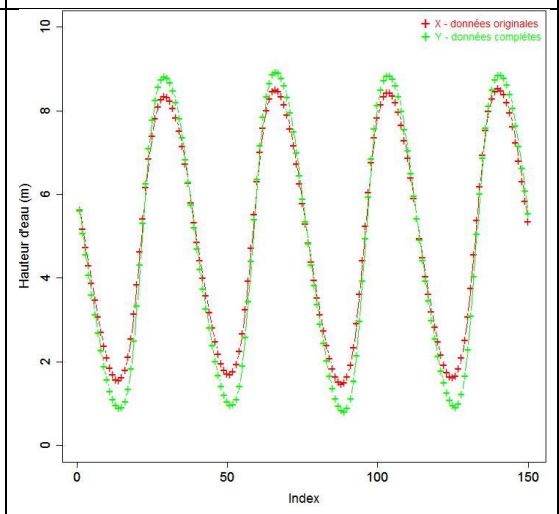
E1 : R vs P</td><td> 

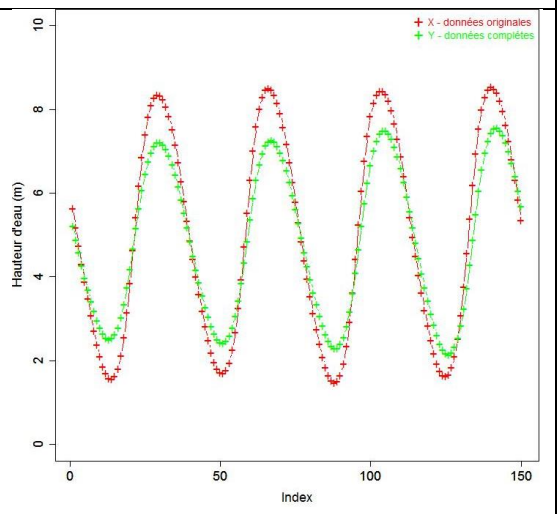
E2 : R vs P</td></tr>
<tr><td>Profil P Bruité</td><td> 

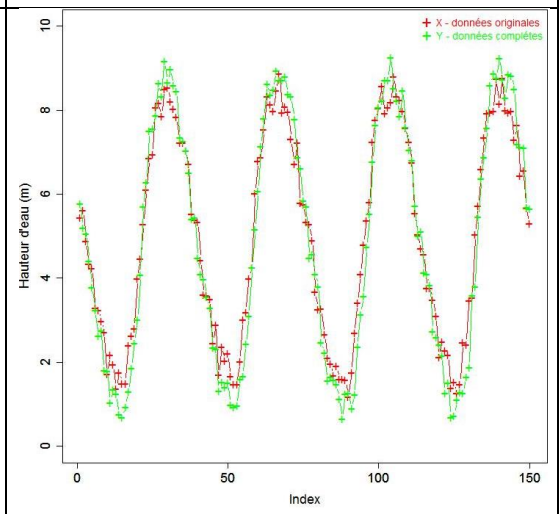
E3 : R vs P</td><td> 

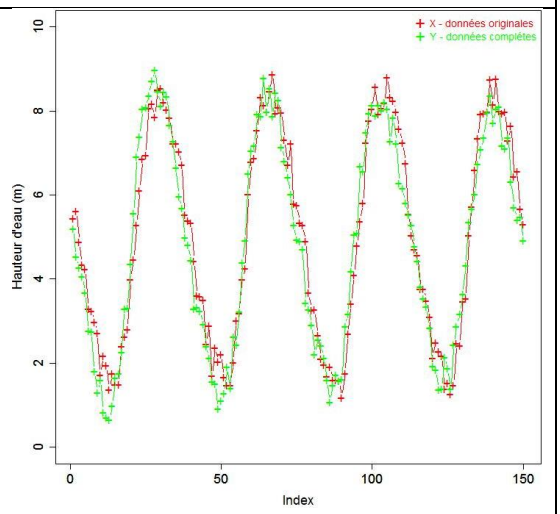
E4 : R vs P</td></tr>
</table>

*Figure 2.24. Représentation, pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur la série $\boldsymbol{x_{Av}}$ entre les données supprimées **X** (en rouge) et les données complétées **Y** (en vert).*

<table>
<tr><th></th><th>Requête R Brute</th><th>Requête R Bruitée</th></tr>
<tr><td>Profil P Brut</td><td> 

E1 : X vs Y</td><td> 

E2 : X vs Y</td></tr>
<tr><td>Profil P Bruité</td><td> 

E3 : X vs Y</td><td> 

E4 : X vs Y</td></tr>
</table>

*Tableau 2.7. Résultats du calcul du coefficient de détermination, de similarité et de l'erreur de déformation moyenne pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur les données précédent les valeurs manquantes entre les données originales **X** et les données complétées **Y**.*

Expérience	R^2	$Sim(Y,X)$	$Err(Y,X)$
E1	0,99***	0,74	0,08
E2	0,98***	0,64	0,13
E3	0,97***	0,70	0,11
E4	0,93***	0,68	0,12

2.4.2.2.2.4. Conclusion

La recherche de séquences similaires par appariement élastique permet de compléter une base de données à valeurs manquantes en respectant autant que possible la dynamique des signaux de la base de connaissances. Les progrès récents dans l'implémentation des méthodes DTW avec listes circulaires permettent de traiter ainsi de grand jeu de données. Les résultats proposés ici sur une série parfaitement connue, XMAHH de la station MAREL-Carnot, avec simulation de données manquantes ouvrent une porte intéressante pour la complétion des données et l'amélioration des traitements liés. La portion complétée a plus de 70 % de similarité avec la courbe originale. De plus, la différence d'amplitude avec les données originales est inférieure à 50 centimètres pour un marnage de 9 mètres. Nous notons également qu'il n'y a pas de déphasage temporel. Ce déphasage est un paramètre important puisque les efflorescences phytoplanctoniques sont très structurées dans le temps.
Plusieurs auteurs proposent d'améliorer les similarités entre deux courbes en intégrant les notions de pente (DDTW (Keogh et Pazzani, 2000)) et courbure (AFBDTW (Xie et Wiltgen, 2010)) ou encore de comparer des séries multidimensionnelles par appariement conjoint (Caillault *et al.*, 2009; Najmeddine *et al.*, 2012).

2.4.3. Complétion multi-conjointe

La complétion conjointe consiste à utiliser les autres paramètres existants pour compléter le paramètre désiré. Nous commencerons par la méthode du plus proche voisin, qui est la méthode la plus instinctive. Puis, nous utiliserons l'imputation par voisinage dans l'espace $D-1$ réduit par classification non supervisée et nous terminerons par une imputation par voisinage dans l'espace $N_p \times (D-1)$ réduit par classification non supervisée.

2.4.3.1. Le plus proche voisin

Une première solution est d'utiliser l'espace $D-1$, où la dimension k du paramètre à compléter n'est pas utilisée. Dans cet espace, nous recherchons la donnée possédant la distance minimale avec la donnée $x(j)$, avec j l'instant où se situe la valeur manquante dans la dimension k. Nous récupérons la valeur de cette donnée retrouvée dans la dimension k afin qu'elle remplace la donnée manquante à l'instant j (algorithme 2.4).

Les figures 2.25 à 2.28 illustrent l'algorithme 2.4 de complétion par le plus proche voisin dans l'espace $D-1$. Nous avons construit un exemple pédagogique, proche du signal de marée et volontairement bruité.
Cet exemple, figures 2.25 et 2.26, est composé d'une séquence de $N=60$ points dans un espace à $D=3$ dimensions. Le point à l'instant j noté $x(j)$ possède une valeur manquante pour la dimension 3 que nous chercherons à compléter par la valeur de son plus proche voisin dans l'espace $[dimension\ 1, dimension\ 2]$. La figure 2.27 schématise cette recherche : calcul de toutes les distances des points avec le point $x(j)$ matérialisé en lignes tiretées grises et sélection du point de distance minimum en vert. Le signal complété est dessiné figure 2.28.

Cette méthode ne permet pas de prendre en compte la dynamique du signal : dans notre cas, la pente croissante de la courbe au voisinage.

Algorithme 2.4. Algorithme pour la complétion par le plus proche voisin.

Entrée :

BD la base de données : $BD = \{bd_{ti}\}_{N\times D}$

x la série à compléter : $x = \{bd_{tk}\}$

Br est la base réduite privée du paramètre à compléter dont toutes les lignes ne possèdent aucun paramètre manquant : $Br = BD\backslash\{x;\ \forall t, bd_{ti} = NA\}$ de taille $N' < N \times (D-1)$

v le vecteur d'indice de données manquantes de x

Sortie : y

$y = x$

Pour tous les indices de v: v_t

$$y_{v_t k} = x_j | j = argmin \left\| Br_K - Br_j \right\|_{K \neq j}$$

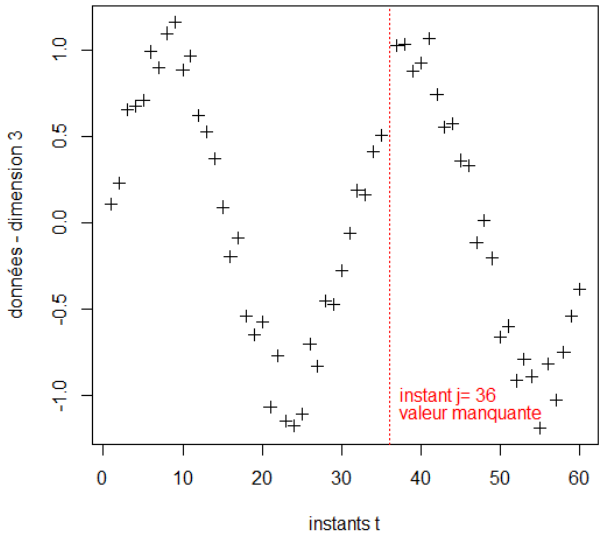


Figure 2.25. Représentation de la dimension 3 du jeu fictif avec à l'instant $j = 36$ une valeur manquante.

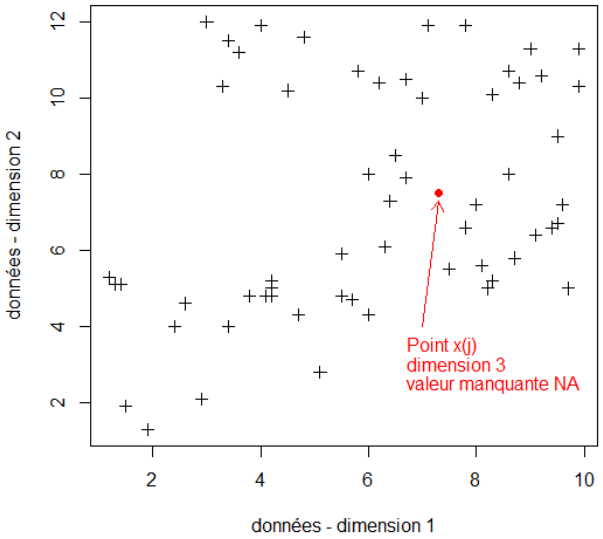


Figure 2.26. Représentation des dimensions 1 et 2 du jeu fictif avec le point $\boldsymbol{x}(j)$ en rouge indiquant que celui-ci est manquant dans la dimension 3.

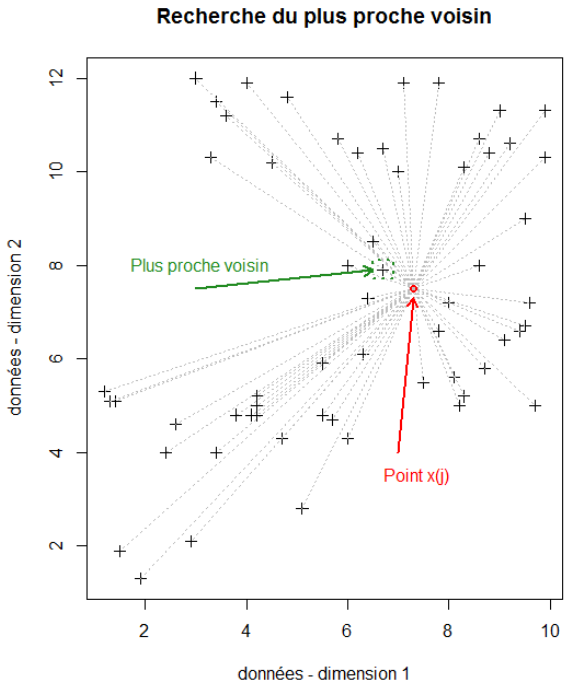


Figure 2.27. Recherche du plus proche voisin du point $\boldsymbol{x}(j)$ dans les dimensions 1 et 2.

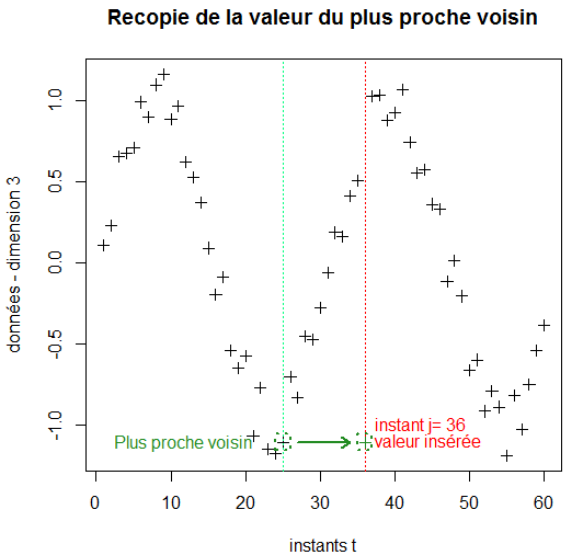


Figure 2.28. Recopie de la valeur du plus proche voisin à l'instant $j = 36$.

L'application de cet algorithme sur notre exemple 1, la hauteur d'eau, confirme que la qualité de la complétion n'est pas satisfaisante (figure 2.29). En effet, le coefficient de détermination est de $5{,}3.10^{-3}$ (avec une p-value de 0,38), la similarité est de 0,56 et l'erreur quadratique est de 0,50.

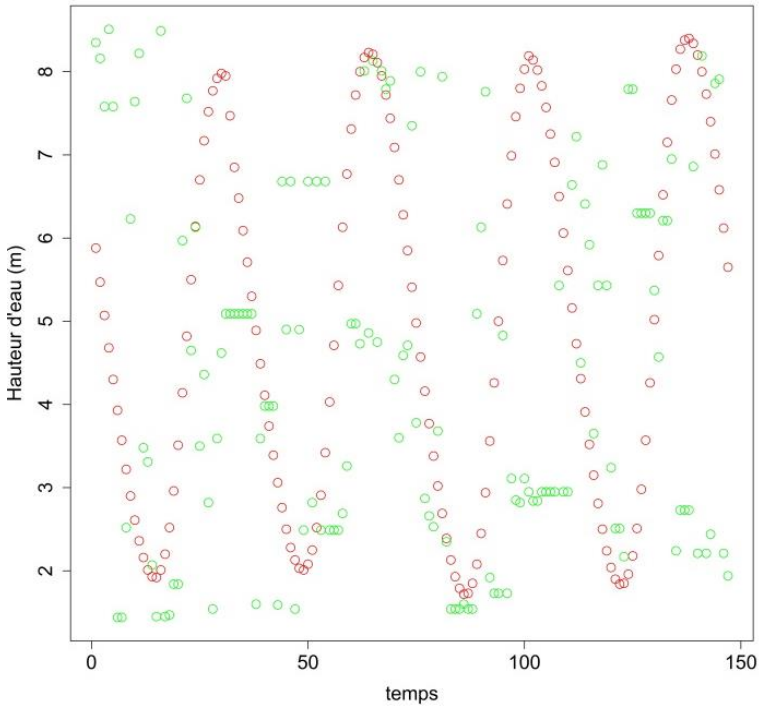


Figure 2.29. Représentation des données supprimées de la hauteur d'eau (en rouge) et du remplacement de celles-ci par leur plus proche voisin (en vert).

Une amélioration possible de cet algorithme 2.4 est d'utiliser la base réduite Br constituée des signaux corrélés au signal à compléter, définie par :

$$Br = \{BD_{tj} | cor(x, BD_{tj}) > 0{,}6\} \qquad (2.13)$$

Notre exemple 1 de la hauteur d'eau n'est pas corrélé aussi fortement avec les autres paramètres, c'est pourquoi cette amélioration ne peut être réalisée sur cet exemple.

2.4.3.2. Imputation par voisinage dans l'espace *D-1* réduit par classification non supervisée

Une autre solution est de réaliser une classification non-supervisée sur une base de données réduite : sans le paramètre contenant les valeurs manquantes. L'algorithme de classification non supervisée utilisé est l'algorithme K-means avec sa version STFKM, nommée Self Tuning Fast K-means. Cet algorithme rapide est basé sur l'algorithme usuel des K-means (Hartigan et Wong, 1979) avec une initialisation des centres efficace pour des grandes bases de points à classer. Nous détaillerons cet algorithme au chapitre 3, il permet ici de réduire le voisinage étudié et de s'affranchir de la détermination du nombre de voisins par la sélection automatique du nombre de centres par la variance expliquée. Le barycentre des données de la série x est calculé pour chaque groupe. Chaque donnée manquante est remplacée par le barycentre du cluster associé à son instant (algorithme 2.5).

Algorithme 2.5. Algorithme de l'imputation par voisinage dans l'espace $N \times (D-1)$ par classification non supervisée.

Entrée :

BD la base de données : $BD = \{bd_{ti}\}_{N \times D}$

x la série à compléter : $x = \{bd_{tk}\}$

Br est la base réduite privée du paramètre à compléter dont toutes les lignes ne possèdent aucun paramètre manquant : $Br = BD \backslash \{x;\ \forall t, bd_{ti} = NA\}$ de taille $N' < N \times (D-1)$

v le vecteur d'indice de données manquantes de x

Sortie : y

$y = x$

Réalisation de l'algorithme STFKM sur Br

Calcul du barycentre des points de chaque classe dans l'espace $N \times D$

Pour tous les indices de v: v_t

$y_{vt} =$ k^{ème} barycentre associé à l'indice vt

Nous reprenons le même exemple à N=60 points. Les figures 2.30 et 2.31 suivantes illustrent l'algorithme 2.5 de recherche du point le plus proche après réduction du nombre de points par l'algorithme STFKM, où K a été calculé tel que la variance expliquée soit supérieure ou égale à 95 %. Le nombre de points dans notre cas est de K=10 centres. Le barycentre des points

appartenant au centre le plus proche obtenu dans l'espace réduit est imputé à la valeur manquante de $x(j)$. Cet algorithme permet de prendre en compte la structure des données au voisinage de $x(j)$ dans l'espace des paramètres disponibles.

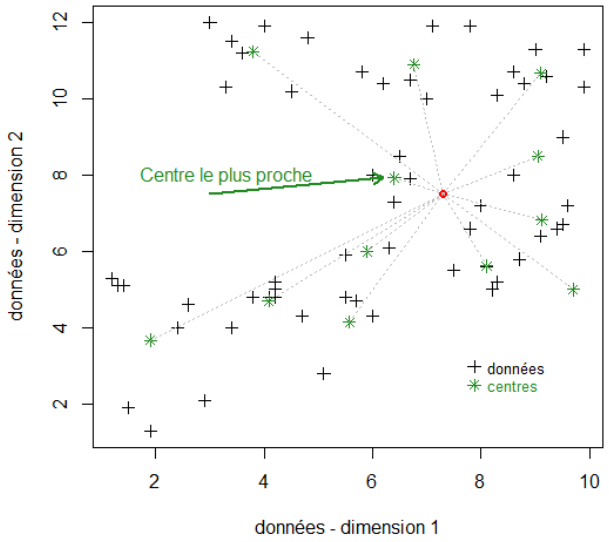


Figure 2.30. Recherche du centre de gravité le plus proche du point $x(j)$ dans les dimensions 1 et 2.

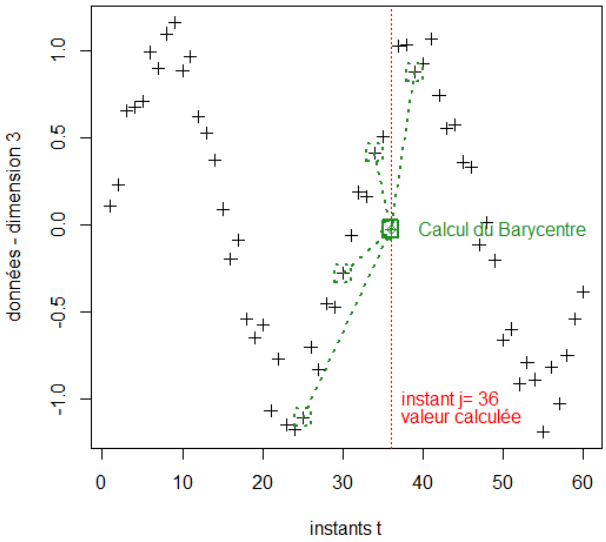


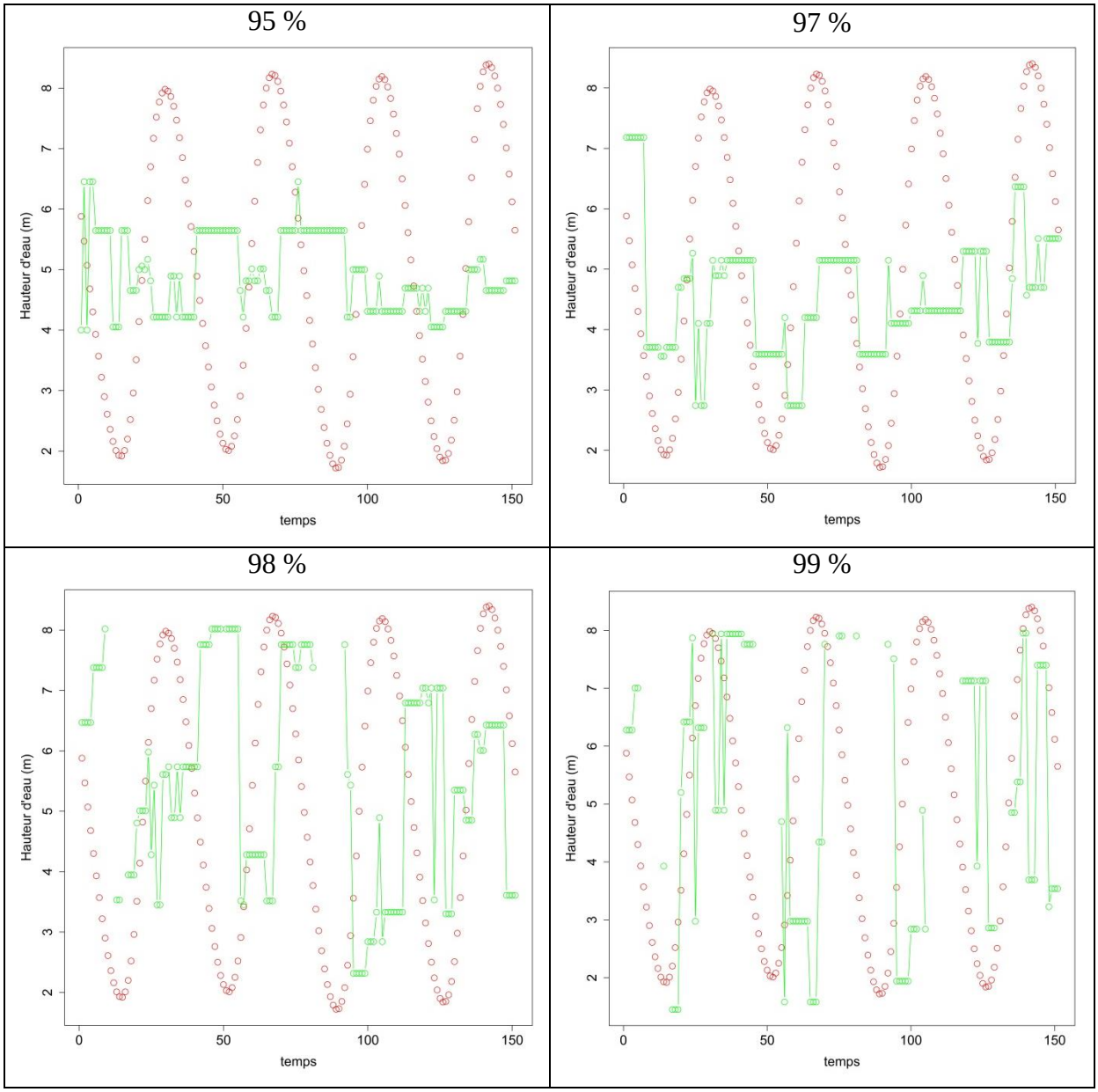
Figure 2.31. Calcul du barycentre dans la dimension 3 des points associés au centre dans les dimensions 1 et 2 (figure 2.30) puis copie de celui-ci à l'instant $j = 36$.

Le nombre d'instants utilisé pour la classification est de 103 616. Le tableau 2.8 regroupe le nombre de centres de gravité en fonction du pourcentage de variance expliquée ainsi que le ratio entre le nombre de données total et le nombre de centres de gravité. Ce nombre de centres de gravité joue un rôle important dans les résultats de la qualité de la complétion de données (tableau 2.9) et nous pouvons constater que la complétion est obtenue avec un pourcentage de variance expliquée de 97 %.

Tableau 2.8. Nombre de centre de gravité et le ratio de ce nombre avec le nombre de données total pour chaque pourcentage de variance expliquée.

Pourcentage de variance expliquée	Nombre de centre de gravité	Ratio nombre total de données – centre de gravité
95 %	1 848	56,07
97 %	4 728	21,92
98 %	8 782	11,80
99 %	20 052	5,17

Figure 2.32. Représentation, pour chaque pourcentage de variance expliquée, des données supprimées de la hauteur d'eau (en rouge) et du remplacement de celles-ci par l'imputation par voisinage dans l'espace D-1 réduit par classification non supervisée (en vert).



Les représentations des reconstructions permettent d'apprécier au mieux les résultats (figure 2.32). En effet, bien que les calculs d'erreurs soient sensiblement identiques, on peut voir que pour 95 et 97 % de variance expliquée, l'amplitude du signal n'est pas respectée. Pour 98 et 99 %, les amplitudes sont en partie respectées mais un déphasage ne permet pas d'avoir des résultats satisfaisants.

Tableau 2.9. Résultats du calcul du coefficient de détermination, de similarité et de l'erreur de déformation moyenne pour chaque reconstruction de la hauteur d'eau pour chaque pourcentage de variance expliquée associé.

Pourcentage de variance expliquée	R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
95 %	$8{,}10.10^{-2}$***	0,48	0,45
97 %	$7{,}57.10^{-2}$***	0,54	0,40
98 %	$6{,}80.10^{-2}$***	0,46	0,54
99 %	$1{,}26.10^{-3}$ (0,74)	0,45	0,53

2.4.3.3. Imputation par voisinage dans l'espace $N_p \times D$ d'une base réduite par classification non supervisée

Afin de réduire la complexité de l'algorithme, une classification non-supervisée (sans aucune connaissance a priori) est réalisée sur la base de données de l'espace d'origine $N \times D$. En effet, après l'étape de classification, une simple recherche du plus proche voisin permet de compléter les données (algorithme 2.6).

Algorithme 2.6. Algorithme d'imputation par voisinage dans l'espace $N_p \times D$ d'une base réduite par classification non supervisée.

Entrée :

BD la base de données : $BD = \{bd_{ti}\}_{N \times D}$

x la série à compléter : $x = \{bd_{tk}\}$

v le vecteur d'indice de données manquantes de x

Sortie : y

$y = x$

Réalisation de l'algorithme STFKM sur BD

Projection des centres dans l'espace $N_p \times (D-1)$ privé de la coordonnée k : C'

Recherche du plus proche voisin de x sur C' : centre g

Copie de la coordonnée k du centre g à la place de la valeur manquante

Sur notre exemple à $N = 60$ points, la projection des centres dans les dimensions 1 et 2 permet de visualiser les distances entre le point $x(j)$ et la base réduite : centres calculés par la classification non supervisée. Le centre le plus proche de la donnée $x(j)$ est repéré puis le barycentre, associé à ce centre dans la dimension 3, est copié à l'instant manquant $j = 36$.

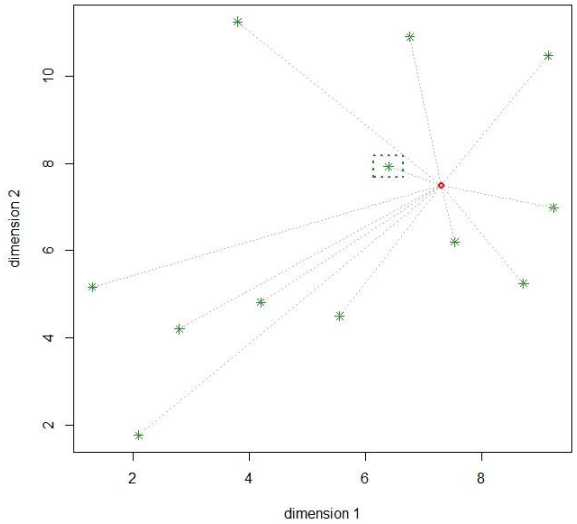


Figure 2.33. Projection des centres dans les dimensions 1 et 2, puis recherche du centre de gravité le plus proche du point $\boldsymbol{x}(j)$.

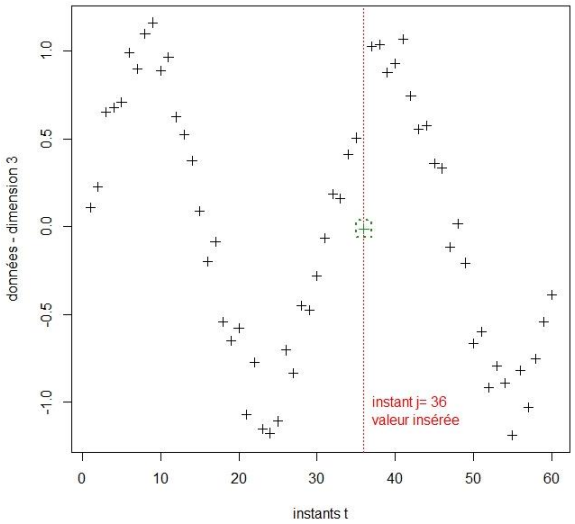


Figure 2.34. Copie du barycentre de la dimension du centre le plus proche de $\boldsymbol{x}(j)$ dans les dimensions 1 et 2 à l'instant $j = 36$.

Cet algorithme 2.6 est réalisé sur notre exemple de hauteur d'eau, la variance expliquée est de 0,99 (figure 2.35). Analysons maintenant la qualité de la complétion : le coefficient de détermination est de $1{,}36.10^{-2}$ (avec une p-value à 0,16), la similarité avec les données réelles est de 0,54 et l'erreur quadratique de 0,55.

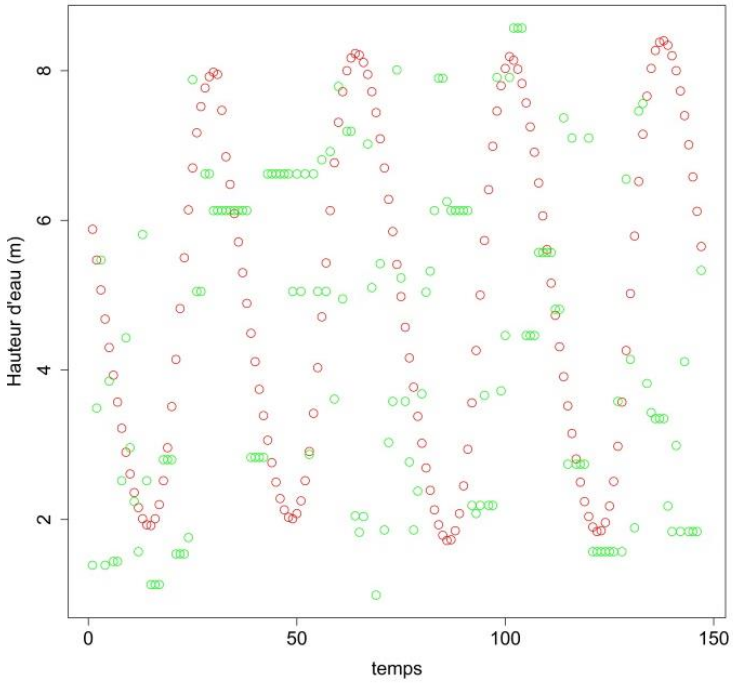


Figure 2.35. Représentation des données supprimées de la hauteur d'eau (en rouge) et du remplacement de celles-ci par l'imputation par voisinage dans l'espace Np×D d'une base réduite par classification non supervisée (en vert).

Dans les algorithmes précédents, les méthodes de classification peuvent être remplacées par des méthodes plus complexes et robustes à la forme des données : mélange de densités, classification spectrale.

2.4.3.4. Complétion pour le paramètre de température de l'eau

L'ensemble de ces méthodes a aussi été testé pour compléter les trous de la température de l'eau dont la variabilité est beaucoup moins marquée à court terme que pour la hauteur d'eau. Le même protocole de suppression a été opéré aux mêmes instants soit une fenêtre N_s de 150 points à l'indice $t = 52\,596$.

Ce paramètre est fortement corrélé avec la température de l'air et l'oxygène dissous. Nous utilisons donc cette base constituée de ces quatre paramètres pour imputer les valeurs manquantes de la séquence.

La synthèse des résultats obtenus sont dans le tableau 2.10. Les graphiques associés sont en Annexe 2. Au niveau de la similarité et du calcul d'erreur quadratique, l'algorithme d'imputation à partir d'une quantification des données donne les meilleurs résultats (similarité de 0,84 et une erreur quadratique de 0,07). Cependant, le coefficient de détermination est très faible (0,02 (0,12)). En effet, les valeurs de remplacement sont très proches des valeurs réelles (différence inférieure au degré Celsius), mais la variabilité des données n'est pas conservée : les valeurs de remplacement sont quasi-constantes alors que les valeurs réelles ne le sont pas.

Tableau 2.10. Résultats du calcul du coefficient de détermination, de similarité et de l'erreur de déformation moyenne de la complétion de la température de l'eau.

Méthode de la section	Variance expliquée	R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
Le plus proche voisin	-	0,02 (0,129)	0,59	0,27
Imputation par voisinage dans l'espace *D-1* réduit par classification non supervisée	0,95	0,03**	0,46	0,23
	0,97	0,03**	0,49	0,21
	0,98	0,19**	0,58	0,18
	0,99	0,01 (0,29)	0,55	0,19
Imputation par voisinage dans l'espace $Np \times$ D d'une base réduite par classification non supervisée	0,99	0,02 (0,12)	0,84	0,07

2.5. Conclusion

La première partie de ce chapitre nous informe sur les différentes méthodes permettant de caractériser une série temporelle (statistiques de base, tendance, saisonnalité, autocorrélation, etc…). Le point faible de ces méthodes est que les signaux doivent être réguliers, c'est-à-dire sans valeurs manquantes. La seconde partie de ce chapitre est donc dédiée à la complétion de données.

Nous avons vu un certain nombre de méthodes pour la complétion de données. Analysons maintenant les plus performantes d'entre elles en fonction des autres sur notre exemple 1 : la hauteur d'eau (tableau 2.11). Afin que la comparaison soit réalisée sur un pied d'égalité, le résultat retenu pour la complétion par appariement élastique (DTW) provient de l'expérience E1 : utilisation du signal brut non bruité artificiellement.

La méthode permettant d'obtenir la meilleure complétion en monodimensionnel est l'imputation par appariement élastique. C'est la seule méthode qui permet de préserver la dynamique de signaux complexes lorsqu'elle possède une banque complète des phénomènes possibles. Cependant, une extension dans un cadre multi-conjoint, qui semble être la meilleure voie pour pallier les autres approches ne prenant pas en compte la dynamique temporelle des signaux, reste à développer.

La méthode actuellement la plus fiable d'un point de vue opérationnel et contrôle vis-à-vis d'une vérité terrain est la moyenne mobile. C'est pourquoi, dans la suite de ce manuscrit, la complétion de données se fera à partir de cette méthode.

Tableau 2.11. Résultats du calcul du coefficient de détermination, de similarité et de l'erreur de déformation moyenne pour l'ensemble des méthodes de complétion les plus performantes.

Méthode	R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
Moyenne mobile	0,12***	0,42	0,40
Spline	$2{,}23.10^{-4}$ (0,85)	0,41	0,53
DTW : E1 sur x_{Ap}	0,99***	0,78	0,07
Le plus proche voisin	$5{,}30.10^{-3}$ (0.38)	0,56	0,50
Voisinage base réduite $N \times (D-1)$, variance expliquée : 0,97	$7{,}57.10^{-2}$***	0,54	0,40
Voisinage base réduite $N_p \times D$, variance expliquée : 0,99	$1{,}36.10^{-2}$ (0,16)	0,54	0,55

Chapitre 3 : Construction d'un modèle Markovien caché non supervisé par classification spectrale

3.1. Introduction

Dans ce chapitre, nous nous intéressons ici à modéliser la dynamique des efflorescences phytoplanctoniques à partir de signaux multidimensionnels et sans connaissance *a priori* c'est-à-dire sans connaissance de la succession saisonnière des taxons du phytoplancton et de la biomasse en général. Modéliser la dynamique des efflorescences signifie être capable de segmenter finement les observations en une suite d'états distincts, par opposition aux approches de segmentation biclasse permettant de détecter la présence ou absence d'une efflorescence. Cette volonté d'augmenter la compréhension est d'une part liée aux stratégies opérationnelles d'échantillonnage et d'autre part au besoin de déterminer la dynamique du phytoplancton de manière plus précise, en exploitant au maximum l'information contenue dans la base de données haute fréquence.

L'écologie numérique connaît un engouement fort vers des techniques explicites d'apprentissage automatique des états contenus dans une base de données, notamment les arbres de décision (Borcard *et al.,* 2011; Chen et Mynett, 2006; Gorsky *et al.*, 2010; Holiday, 2009). Ces derniers ont l'avantage d'offrir une vue synthétique et compréhensible par un public non spécialiste du traitement des données. Le sommet de l'arbre correspond à l'ensemble des observations, les branches aux critères de segmentations et ses feuilles à la segmentation finale. Les articles proposant des techniques d'apprentissage discriminante utilisent souvent ces arbres dans leurs comparatifs pour situer leurs travaux ou interpréter leurs résultats (Millie *et al.*, 2006; Zighed et Rakotomalala, 2000).

Les feuilles (sorties) d'un arbre hiérarchique correspondent à des états caractéristiques d'observations d'inertie minimale basées sur des hypothèses de seuils multivariés. Or cette approche par seuil ne semble pas correspondre à la dynamique des efflorescences. Comme nous pouvons le voir sur la figure 3.1, les niveaux maximums de fluorescence durant une efflorescence printanière ne sont pas les mêmes d'une année à l'autre. Pour l'année 2005, l'efflorescence printanière avec un niveau moyen (proche de 4 FFU) peut être confondue avec le niveau de l'efflorescence suivante. De part ces variabilités annuelles et interannuelles, il n'est par conséquent pas plausible de fixer un seuil sur l'unique connaissance de la fluorescence d'où l'importance de considérer une approche multi-paramètres.

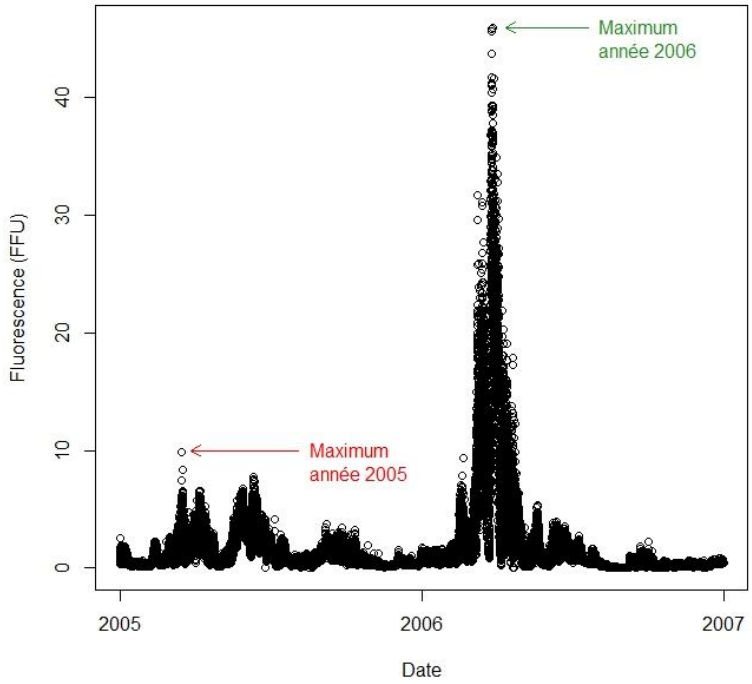


Figure 3.1. Évolution temporelle de la fluorescence (FFU) mesurée par la station MAREL-Carnot au cours de la période 2005-2006. Le maximum de chaque année est pointé par une flèche: rouge pour 2005, verte pour 2006.

Conformément à l'évolution normale du phytoplancton mise en évidence par les travaux de Margalef (1978) et Reynolds *et al.* (2002) (Chapitre 1), la succession des phases d'une efflorescence phytoplanctonique peut être vue comme un chemin à travers des états environnementaux guidés à la fois par les observations et leur enchaînement. Nous pouvons ainsi représenter la dynamique via un graphe connecté dont un nœud représente un état environnemental et un arc de connexion la possibilité de passer d'un état à un autre. Les états ne sont pas des événements directement observables contrairement aux paramètres physico-chimiques et biologiques. Nous pouvons retirer aussi des travaux cités précédemment que la biomasse phytoplanctonique est contrainte par un niveau élevé de dépendance entre la succession des observations. L'utilisation d'un Modèle de Markov Caché (MMC) ergodique semble alors l'approche naturelle pour caractériser la dynamique d'une efflorescence phytoplanctonique à partir des seules observations que sont les paramètres physico-chimiques et biologiques.

Les modèles de Markov cachés (Rabiner, 1989) ont montré leur intérêt en reconnaissance de la parole et de l'écriture où de larges bases sont étiquetées et permettent de construire ces modèles de manière supervisée. Dans Caillault *et al.* (2005) et Jaeger *et al.* (2001), une hybridation entre un réseau de neurones et un MMC est réalisée pour reconnaître de l'écriture cursive ou hors-ligne. Un mot peut être modélisé par un MMC, en effet un mot est une séquence de lettre structurée par des probabilités de transitions où chaque lettre est représentée par plusieurs graphèmes avec des probabilités d'apparition. La construction d'un

modèle de Markov caché nécessite d'estimer l'ensemble de ses paramètres. Les paramètres du MMC à définir sont :

- Le nombre d'états ;
- Les lois de transition entre états et les lois d'émission de ces états ;
- La caractérisation de ces états.

Habituellement, les paramètres du MMC sont appris avec une base de données labellisée ou fixée avec une information a priori.

Beaucoup de travaux ont été menés sur la structure interne d'un MMC (maillage de MMC, topologie 2D ou 3D (Alexandrov et Gerstein, 2004; Won *et al*., 2006), MMC avec des états hiérarchisés (Fine *et al.*, 1998), contrainte de durée dans un état, ...) pour modéliser des séquences à partir de bases de connaissance étiquetées ; de même plus récemment ils ont été étendus à l'alignement de séquences (technique dite « Pair-HMM ») (Arribas Gil, 2007; Shao *et al.*, 2004).

Ici, nous abordons la question de la prédiction des efflorescences phytoplanctoniques en utilisant un modèle de Markov caché hybride non supervisé construit à partir d'une importante base de données multidimensionnelle. La construction d'un MMC sans apprentissage est une piste récente.

Avant d'aborder dans la suite les différentes techniques et nos choix pour construire un MMC de façon non supervisée, nous rappelons sa définition et précisons les notations. Puis nous détaillerons notre hybride.

3.2. Modélisation de séries temporelles par un modèle de Markov caché construit par apprentissage non supervisé

3.2.1. Présentation d'un modèle de Markov caché

Un MMC est un processus markovien discret dont les états du modèle et les événements (qui peuvent être observés) sont séparés. Il est noté $\lambda = \lambda(N, M, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ et est défini à partir d'un couplet de structure (N, M) et un ensemble de trois paramètres probabilistes $(\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ (figure 3.2) que nous rappelons comme suit (Rabiner, 1989) :

- N le nombre fini d'états distincts $\boldsymbol{S} = \{s_1, s_2, \dots, s_N\}$ du modèle, ces états sont les nœuds du graphe. Par rapprochement avec notre application nous pourrions vouloir détecter les états environnementaux suivants : la période non productive, la pré-efflorescence, l'efflorescence, la post-efflorescence, et d'autres événements rares tels que des ouvertures de barrage, des failles capteurs, etc.....
- M le nombre de paramètres/symboles distincts $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M\}$ caractérisant un état s.
- $\boldsymbol{\pi} = \{\pi_i\}$ de taille N, définit la distribution des états initiaux. $\pi_i = P(s(t = 1) = s_i)$ est la probabilité que l'état initial à l'instant $t = 1$ soit l'état s, noté également état i.

Lorsque aucune information sur l'état prédomine durant l'acquisition des données, les états initiaux sont considérés *a priori* équiprobables.

- $\boldsymbol{A} = \{a_{ij}\}$ de taille $N \times N$, décrit les probabilités de déplacement entre états (arcs), avec $a_{ij} = P(s(t) = s_i | s(t-1) = s_j)$ la probabilité conditionnelle du passage d'un état s_i à un état s_j. La matrice de transition $\boldsymbol{A}$ est normalisée en ligne.
- $\boldsymbol{B} = \{b_{ik}\}$ de taille $N \times M$, définit les probabilités d'émission soit la distribution des probabilités pour chaque symbole par état. $b_{ik} = P(\boldsymbol{v}(t) = \boldsymbol{v}_k | s(t) = s_i)$ correspond à la probabilité conditionnelle d'être à la fois dans l'état s_i et dans un symbole $\boldsymbol{v}_k$. La matrice $\boldsymbol{B}$ est normalisée telle que $\sum_{i,j} b_{ij} = 1$.

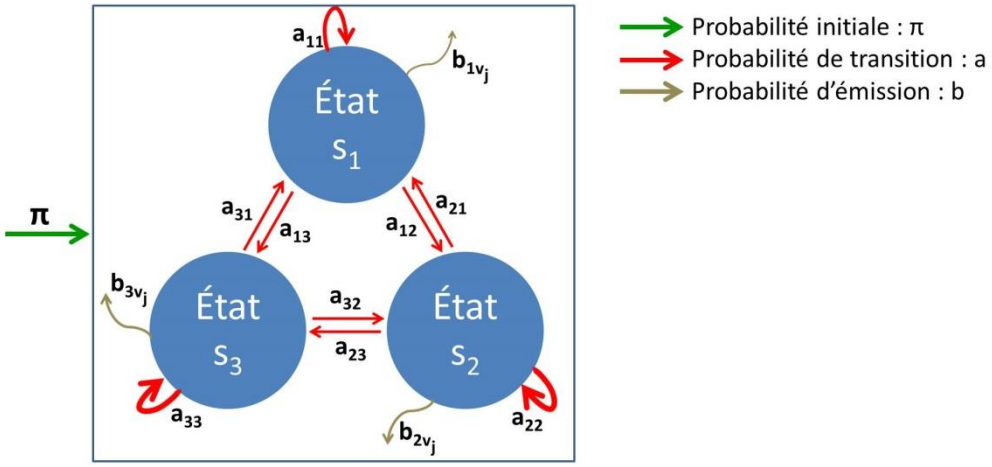


Figure 3.2. Représentation d'un modèle de Markov caché réduit ici à trois états. Les boules bleues correspondent aux états ($\boldsymbol{s_i}$), la flèche verte à la probabilité initiale de rentrer dans un des trois états. Les flèches rouges représentent les probabilités de transition entre états ($\boldsymbol{a_{ij}}$) et, les flèches marrons les probabilité d'émission d'un état par rapport à son symbole ($\boldsymbol{b_{ik}}$).

3.2.2. Analyse des approches usuelles pour déterminer les paramètres

Les paramètres MMC $\lambda(N, M, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ peuvent être estimés par Maximum A Posteriori (MAP) selon l'équation $\lambda_{MAP} = \arg\max_\lambda \big(L(\boldsymbol{X}, \lambda) \times P(\lambda)\big)$ où $L(\boldsymbol{X}, \lambda) = P(\boldsymbol{X}|\lambda)$ est la vraisemblance du modèle vis-à-vis des observations $\boldsymbol{X}$ et $P(\lambda)$ la distribution *a priori* des paramètres MMC.

En faisant l'hypothèse que les paramètres à estimer suivent une distribution *a priori* uniforme, l'estimateur MAP est équivalent à l'estimateur par maximum de vraisemblance (ML pour Maximum Likelihood) : $\lambda_{ML} = \arg\max_\lambda \big(\log\big(L(\boldsymbol{X}, \lambda)\big)\big)$. L'optimisation de cette vraisemblance est difficile à calculer globalement. Pour cette raison, il est usuel de poser certaines hypothèses tel que la détermination indépendante des paramètres notamment N et M avec leurs paramètres associés et d'utiliser des processus itératifs.

Détermination de N

Le nombre d'états est très souvent fixé par un expert en lien avec les applications envisagées. Les autres paramètres $\theta = (M, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ sont alors déterminés selon un critère de maximum de

vraisemblance des observations $\boldsymbol{X}$ contraint à ce choix de N et une partition des observations associée. Une autre possibilité est d'adapter itérativement les paramètres en alternant une phase d'adaptation de la structure des états des MMC avec une phase d'adaptation des paramètres $\theta = (M, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$(Ait-Mohand *et al.*, 2010)). Ainsi la seconde adaptation, celle de $\theta = (M, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, utilise une méthode basée sur le MAP ou le maximum de vraisemblance $P\left(s|\boldsymbol{X}, \hat{\lambda} = \hat{\lambda}(N, \theta)\right)$.

Sans observations labellisées, la combinaison (nombre d'états, choix de l'algorithme de partitionnement des observations en état, modèle $\theta = (M, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$) peut être vu comme des modèles MMC différents. Ainsi, une solution itérative est d'analyser le maximum de vraisemblance de ces modèles. La forte complexité opératoire de cette approche incite les auteurs à fixer *a priori* le nombre d'états.

Détermination itérative de $M, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}$

Dans le cas le plus simple, l'observation des symboles correspond aux sorties du système ou à une loi d'émission. Le choix de la distribution de cette loi dépend du problème et de la complexité calculatoire et d'interprétation choisie. Il est usuel d'utiliser une loi d'émission gaussienne dont les caractéristiques sont faciles à exprimer ou des lois non elliptiques (Chatzis, 2010; Volant *et al.*, 2012). Le calcul explicite de ces lois requiert une connaissance complète de nos données.

Pour une séquence d'observations finie avec une labellisation partielle et un nombre d'états N donné, les symboles du MMC, les matrices de transition et d'émission peuvent être adaptées itérativement (Liao *et al.*, 2002) en utilisant une approche Expectation-Maximization (EM) dans le but de maximiser la vraisemblance de l'état par rapport à l'observation $\boldsymbol{X}$ et du modèle $\log P(s|\boldsymbol{X}, \lambda)$. Partant d'une initialisation des paramètres MMC (i=1), l'algorithme EM itère les étapes suivantes jusqu'à obtenir la convergence du critère :

- Phase E : calcul d'un paramètre de vraisemblance des observations et de l'état à partir du MMC, $\lambda = (N, \theta_i)$;
- Phase M : redéfinition des nouveaux paramètres du MMC qui optimise ce critère en utilisant l'algorithme Baum-Welch, θ_{i+1}

Dans le but de réduire la complexité du modèle et de choisir celui qui donne la description optimale des données, il est usuel d'utiliser une approche EM basée sur un critère de vraisemblance pénalisé. La vraisemblance du modèle est contrainte par la minimisation du nombre de paramètres libres du modèle, noté ν_0. Le choix du nombre de paramètres libres peut être déterminé par le critère de redondance minimax (Rissanen, 1984), basé sur le nombre de bits moyen nécessaire pour coder un message de taille N, $R = \frac{\nu_0}{2} \log N$.

Ainsi on cherche à maximiser le critère suivant $\log L(\boldsymbol{X}, \lambda|N) - \frac{\nu_0}{2} \log N$.

Quel que soit le critère utilisé et ses dérivés (comme par exemple le critère d'information Bayésien $BIC(M) = \log P\left(\boldsymbol{X}, M, \hat{\lambda}\right) - \frac{\nu_0}{2} \log N$), les performances de l'EM dépend de l'étape

d'initialisation, soit de la sélection d'un modèle pré-estimé tant pour les états que les symboles.

Figueiredo et Jain (2002) ont montré l'impact de cette initialisation et l'intérêt d'estimer et sélectionner le modèle par un unique algorithme dans le cadre de données multi-classe représentées par des mélanges de gaussiennes.

Détermination de M par quantification

Pour s'affranchir de cette initialisation, nous avons opté pour des algorithmes dont le calcul explicite des paramètres ne requiert aucune connaissance. Rousseeuw *et al.* (2013b) et Warren Liao (2005) ont proposé une méthode non hiérarchique sur le processus itératif K-means (détaillé section 3.3.1). Pour notre application, cela se justifie pleinement puisqu'un état environnemental n'est pas caractérisé par un vecteur de paramètres physiques unique ; deux sorties du système peuvent appartenir à plusieurs états (Exemple : les symboles sont représentés par les mois et les états par les saisons).

Baudry *et al.* (2010) ont préféré utiliser un critère d'entropie pour enlever la contrainte de groupes de formes sphériques imposée par l'algorithme K-means et utiliser une combinaison hiérarchique des paramètres. Il a été décidé d'utiliser une méthode de classification non supervisée robuste, la classification spectrale, décrite à la section 3.3.2 qui permet elle aussi de s'affranchir de la forme de la distribution des paramètres afin de conserver la structure interne des données.

Ainsi, nous construirons notre modèle hybride basé sur les principes suivants :

- Définir automatiquement la meilleure structure du MMC en utilisant une approche unique pour définir les paramètres N, M, la topologie des états et des symboles sans hypothèse sur leur distribution.
- Minimiser le nombre de paramètres libres du modèle $\nu_0 = card(\lambda)$. La taille des paramètres $\boldsymbol{\pi}$ $(N \times 1)$, $\boldsymbol{A}$ $(N \times N)$ et $\boldsymbol{B}$ $(N \times M)$ est fonction des deux paramètres N et M et par conséquent $card(\lambda) = N + M + N \times 1 + N \times N + N \times M$. Cette réduction devra conserver au mieux la structure interne des données.
- Calculer les paramètres probabilistes du MMC sans processus itératif.
- Maximiser la vraisemblance de l'état obtenu par rapport à une nouvelle observation et le modèle construit.

3.3. Construction de notre Modèle de Markov Caché non supervisé

L'architecture du système MMC-NS proposé à partir du MMC hybride est schématisée sur la figure 3.3. Les données collectées $\boldsymbol{X}$ à haute résolution temporelle de 2005 à 2008 (① de la figure 3.3), de taille $N_p \times D_p$, sont tout d'abord prétraitées ② (chapitre 2). Ensuite, une étape de classification non supervisée est appliquée afin de détecter les états environnementaux et de les caractériser en symboles ③. L'étape finale s'appuie sur l'information temporelle entre ces états pour développer le modèle de la dynamique des efflorescences du phytoplancton (il

s'agit de suivre l'évolution de la biomasse du phytoplancton) ④. Le modèle alors construit est utilisé pour estimer une nouvelle efflorescence phytoplanctonique à venir ou des états spécifiques ⑤.

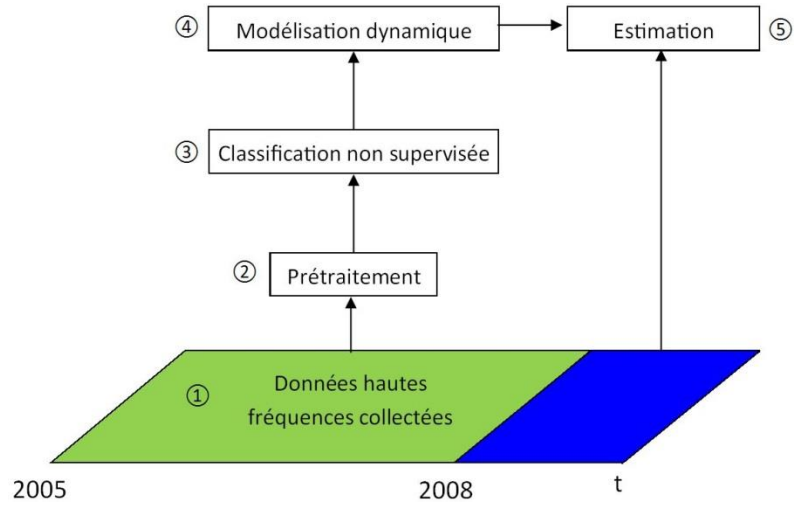


Figure 3.3. Architecture du système hybride constituée de 5 étapes : Prélèvement, prétraitement, classification non supervisée, modélisation dynamique et estimation des états.

Le MMC-NS hybridé $\lambda(\boldsymbol{S}, \boldsymbol{V}, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$ (figure 3.4), est construit à partir des étapes suivantes :

- Génération des symboles $\boldsymbol{V} = \{v_k\}$, matrice de M points de taille N, par quantification vectorielle à partir d'une base d'observations $\boldsymbol{X} = \{x_i(t)\}$ où x_i est la i^{ème} composante de la donnée x acquise à l'instant t.
- Génération des états $\boldsymbol{S} = \{s_i\}$, $s_i = i$ le label associé à une observation. Cette génération est basée sur une classification non supervisée des symboles. La base de données MAREL-Carnot contient $26\,280 \times 19$ valeurs par an ($N_p = 26\,280$ instants et $D_p = 19$ paramètres). Pour découvrir les états sous-jacents à cette grande base de données, une sélection de prototypes est nécessaire. L'étape de quantification vectorielle nous permet de réduire le nombre de données et ainsi de pouvoir appliquer un algorithme de classification non supervisée possédant des contraintes de coût calculatoire ou mémoire, tel que la classification spectrale.
- Calcul de $\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}$

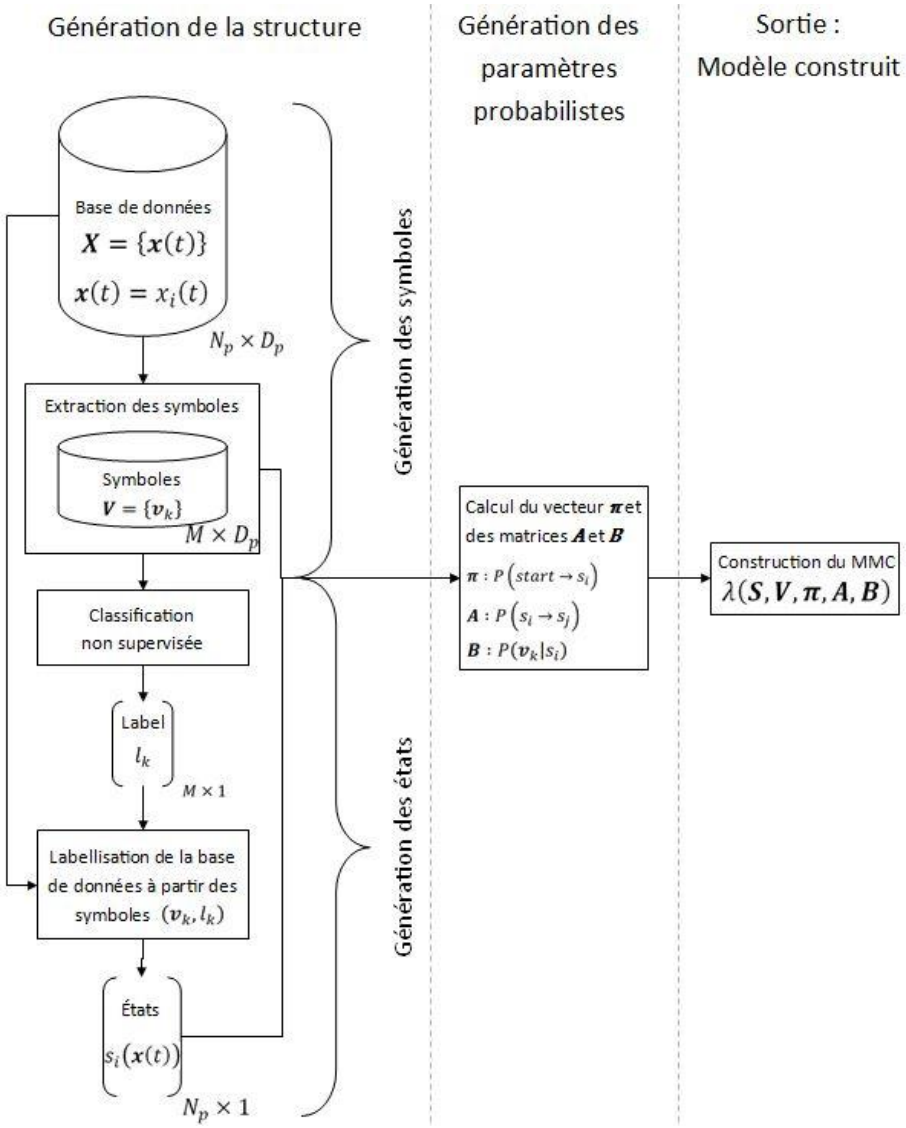


Figure 3.4. Système basé sur un Modèle de Markov Caché hybridé. Trois parties sont séparées par des lignes en pointillées : génération de la structure, génération des paramètres probabilistes et la sortie (modèle construit).

3.3.1. Génération des symboles

Les symboles $\boldsymbol{V} = \{v_1, \dots, v_M\}$ du MMC caractérisent un état s_i. Un état environnemental n'est pas caractérisé par un unique représentant (figure 3.1) : l'état d'efflorescence printanière de 2006 présente des niveaux totalement différents de l'année 2005. Par conséquent, nous avons opté pour un codage des symboles (codebook) par quantification vectorielle (Debyeche *et al.*, 2007; Ko *et al.*, 2008; Koo *et al.*, 1992). L'idée principale est de construire un vecteur de prototypes à partir de l'ensemble des observations notées $\boldsymbol{X} = \{x(1), \dots, x(N_p)\}$ de N_p points de données conservant l'information haute fréquence. L'algorithme K-means est bien adapté à la méthode de quantification vectorielle et est populaire dans la classification de données (Jain, 2010).

Rappelons l'idée principale de cet algorithme. La technique des K-means est de partitionner un ensemble de données $\boldsymbol{X} = \{x(1), \dots, x(N_p)\}$ en K groupes $G = \{g_1 \dots g_K\}$ selon un critère de minimisation des distances intra groupe J.

$$J(X, G) = \sum_{k=1}^{K} \sum_{i, x(i) \in g_k}^{N_p} \|x(i) - \mu_k\|^2 \qquad (3.1)$$

Où $\mu_k = \sum_{i, x(i) \in g_k}^{N_p} \frac{x(i)}{card(g_k)}$ est le barycentre du groupe g_k

Ce critère est optimisé par le processus itératif suivant :

1) Initialisation de K centres ;
2) Affectation de chacun des points $\boldsymbol{X}$ à son centre le plus proche ;
3) Ré-estimation des centres ;
4) Calcul du critère J, retour à 2) si le critère n'est pas respecté.

Plusieurs algorithmes existent, dont notamment l'algorithme de Hartigan et Wong (1979) qui impose qu'aucun groupe n'est de cardinal nul. Cet algorithme nécessite plusieurs itérations pour converger (Warren Liao, 2005) et la connaissance du paramètre libre K.

Lorsque le jeu de données est très important (supérieur à 20 000 points) vis-à-vis des machines calculatoires actuelles, l'algorithme K-means rapide (Fast K-means) est utilisé (Shindler *et al.*, 2011). La différence réside dans l'initialisation des centres. Usuellement, l'initialisation des centres est aléatoire parmi les données. Dans le Fast K-means, un sous-échantillonnage de la base est opéré. Sur cet ensemble réduit est effectué un premier K-means dont les centres finaux seront utilisés comme centres initiaux d'un second K-means. Nous avons modifié cet algorithme en incluant l'automatisation de la recherche du nombre de groupes K basé sur le critère d'Elbow noté vE au lieu du critère initial de dispersion des points par rapport au critère J :

$$vE = \frac{J_B}{J_{Total}} = \frac{J_B}{J_B + J_W} \qquad (3.2)$$

Où J_B est la dispersion inter groupe et J_w la dispersion intra groupe :

$$J_B = \sum_{k=1}^{K} \sum_{i,j}^{N_p} \|x(i) - x(j)\|^2 \ avec\ x(i) \in g_k, x(j) \notin g_k \qquad (3.3)$$

$$J_w = \sum_{k=1}^{K} \sum_{i,j}^{N_p} \|x(i) - x(j)\|^2 \ avec\ x(i)\ et\ x(j) \in g_k \qquad (3.4)$$

Le nombre de groupes K est incrémenté jusqu'à ce que le pourcentage de variance expliquée ou le nombre $Kmax$ de prototypes retenus (c'est-à-dire les symboles) soit atteint. Par ailleurs, pour accélérer ce processus, il est aussi possible de procéder à une recherche de K dans l'ensemble $K = 1, \dots, Kmax$ par dichotomie. Le principe de l'algorithme proposé, nommé algorithme K-means rapide auto-réglé (STFKM acronyme de Self Tuning Fast K-means), est décrit dans l'algorithme 3.1. L'initialisation des centres (Étape 1) pour les grandes bases de

données (nombre de points supérieur à 20 000) est très importante ici pour accélérer le processus de convergence. $Kmax$ est le nombre maximum de points réduits spécifié par l'utilisateur ou, dans le cas par défaut, le nombre de mesures de la série temporelle. $varExplained$ est la variance expliquée désirée par l'utilisateur ; par défaut, ce nombre est fixé à 95 %.

Algorithme 3.1. Aperçu de l'algorithme K-means rapide auto-réglé noté STFKM utilisé pour la génération des symboles.

Procédure STFKM$\left(\boldsymbol{X}_{N_p \times D_p}, Kmax = N_p, varExplained = 0{,}95\right)$
Variables : $k = 1$, $vE = 0$
tant que $k < Kmax$ ou $vE < varExplained$ **faire**
$k = k + 1$
Étape 1 : Initialisation des k centres
Découper les données en n sous-échantillons de 20 000 points
Partitionner chaque sous-échantillon E_n en k groupes selon K-means ($K = k$)
Sélectionner les k centres des groupes de E_n selon le critère $J(E_n)$ maximal
Étape 2 : Affecter chaque point de $\boldsymbol{X}$ des N_p points à son centre le plus proche.
Étape 3 : Ré-estimer les k centres des groupes
Étape 4 : Si aucun mouvement d'affectation
Alors passer à l'étape 5
Sinon retourner à l'étape 2
Étape 5 : Calcul de $vE = J_B/(J_B + J_w)$
fin du **tant que**
retourner les k centres obtenus.
fin de la **procédure**

Soit l'exemple pédagogique (figure 3.5) composé d'un cercle d'une taille de 2000 points et une boule inscrite de même taille. On remarque que la réduction de $\boldsymbol{X}$ en K symboles est fidèle à la structure des données (réduction de 4 000 à 260 points en accord avec le critère Elbow : variance expliquée de 95 %).

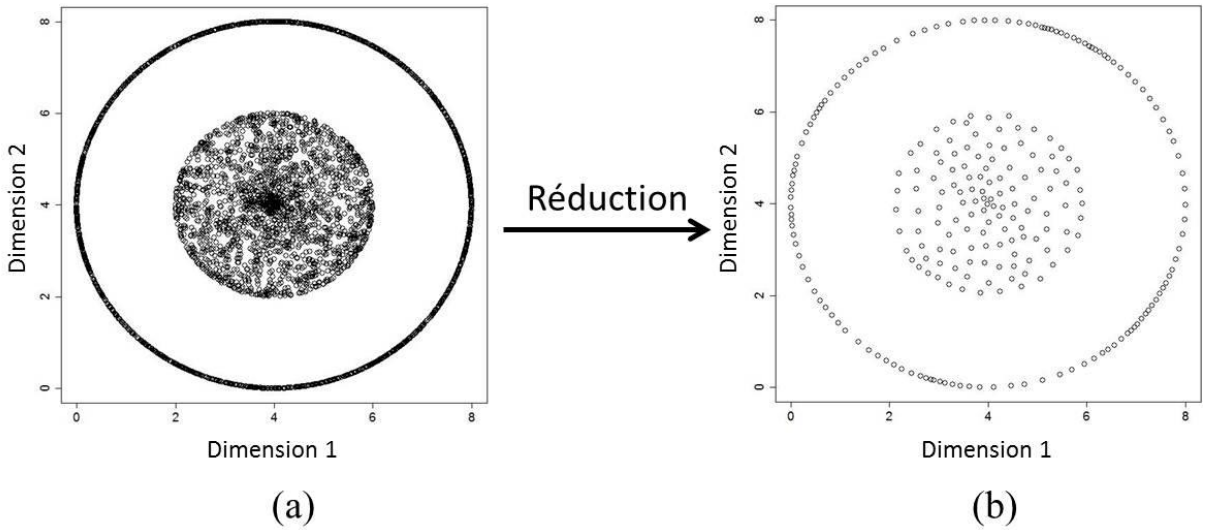


Figure 3.5. Étape de génération des symboles par STFKM : représentation graphique du (a) jeu de données de l'exemple pédagogique brut, et (b) le jeu de données échantillonné avec le critère d'Elbow fixé à 95 % de variance expliquée.

L'algorithme STFKM nous fournit K centres qui seront nos $M = K$ symboles, réduisant le plus fidèlement notre base $\boldsymbol{X}$.

3.3.2. Génération des états

A partir de ces M symboles, N états sont calculés par classification non supervisée. L'algorithme K-means, ou sa version modifiée, peut être utilisé pour générer nos états (Rousseeuw *et al.*, 2013a, 2013b).

Cet algorithme est adapté au traitement de données volumineuses (K-means Rapide). Simple et rapide à implémenter, sa complexité (quantité de ressources en temps et en espace mémoire nécessaire pour la résolution de problèmes au moyen de l'exécution d'un algorithme) est en $O(IKN_D)$ avec I le nombre d'itérations, K le nombre total de groupes et N_D le nombre total de données (lignes x colonnes). L'algorithme est également très sensible aux évènements rares, au bruit et aux données aberrantes, ce qui est, dans notre cas d'étude, une force. En effet, la dynamique d'un écosystème n'est pas constante, il faut donc une méthode permettant de détecter des phénomènes inhabituels intervenant sur le système étudié.

Cependant, l'une des principales faiblesses de cet algorithme est qu'il ne permet pas de traiter des ensembles de données de structures complexes non convexes et non-linéairement séparables (Singh et Chauhan, 2011). Par conséquent, pour un jeu de données dont la représentation graphique est similaire à celle de la figure 3.5 (a), le résultat de la classification n'est pas optimal (figure 3.6). En effet, le résultat d'un K-means dans l'espace initial des données montre une séparation du cercle et de la boule inscrite selon leur diamètre en deux classes. Or, il est intuitif de distinguer les deux éléments géométriques : le cercle extérieur représente le premier groupe et la boule inscrite représente le second.

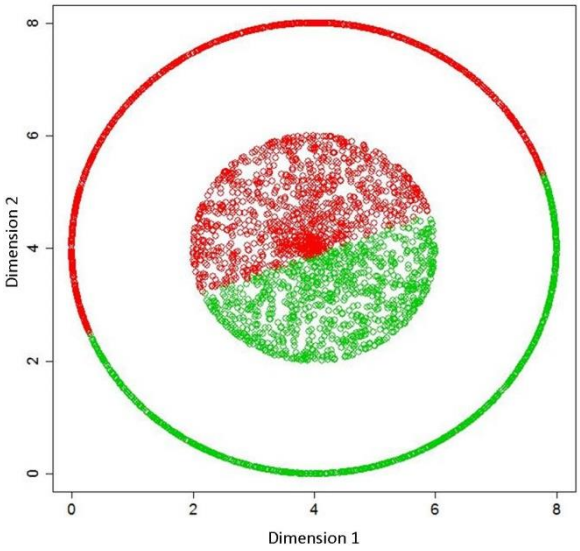


Figure 3.6. Projection du résultat de classification par l'algorithme K-means sur l'exemple pédagogique.

Dans notre étude, les paramètres physico-chimiques de la station MAREL-Carnot suivent des processus stochastiques, non-linéaires et non stationnaires (Chapitre 2). Ils ne suivent pas de distributions gaussiennes et la caractérisation en état environnemental est inconnue. Par conséquent, nous avons choisi la méthode de classification spectrale afin de lever ces hypothèses sur la forme des données.

L'idée clé de la classification spectrale est de transformer l'espace initial des données d'entrées dans un nouvel espace où les données seront linéairement séparables. Ainsi, un algorithme tel que celui du K-means pourra donc facilement classer les données projetées. Cette technique de classification découle de la théorie des graphes où chaque donnée peut être assimilée à un nœud et chaque arc de connexion entre nœuds à une mesure de similarité entre deux données. Pour partitionner les données, un critère de coupe est choisi en fonction de l'application (coupe intra-groupe, coupe intergroupe, coupe normalisée selon le cardinal ou le volume des groupes obtenus (Shi et Malik, 2000)). L'ensemble de ces coupes est calculé à partir des similarités du graphe, celles-ci dépendent de la métrique choisie notée w.

Une coupe entre deux groupes g_k et g_l est définie par l'équation :

$$Cut(g_k, g_l) = \sum_{i=1, x(i)\in g_k}^{N_p} \sum_{j=1, x(j)\in g_l}^{N_p} w\big(x(i), x(j)\big) \qquad (3.5)$$

La coupe intra-groupe ($Cut_w(g_k)$) est définie par la somme des similarités à l'intérieur du groupe.

$$Cut_w(g_k) = Cut(g_k, g_k) \qquad (3.6)$$

La maximisation de la somme des coupes intra-groupes pour une partition $G = \{g_1, \dots, g_k\}$ est équivalent à minimiser le critère de dispersion J_w (équation (4)). De même, minimiser la somme des coupes inter-groupes ($Cut_b(g_k)$) est parallèle du critère de dispersion J_b à maximiser.

$$Cut_b(g_k) = \sum_{l=1}^{K} Cut(g_k, g_l) \qquad (3.7)$$

Pour notre application, nous avons choisi d'optimiser la partition sur un critère d'Elbow, soit minimiser la coupe inter-groupe tout en maximisant la coupe intra-groupe, ce qui revient à optimiser un critère de type :

$$\sum_{k=1}^{K} \frac{Cut_b(g_k)}{Cut_b(g_k) + Cut_w(g_k)} \qquad (3.8)$$

En reprenant la définition d'un volume d'un groupe

$$vol(g_k) = Cut_b(g_k) + Cut_w(g_k) \qquad (3.9)$$

soit la somme de tous les arcs issus des nœuds du groupe, nous obtenons alors le critère $NCut$ comme étant la somme sur chaque groupe de la partition de la coupe inter-groupe sur le volume du groupe :

$$NCut(\boldsymbol{X}, G) = \frac{1}{K} \sum_{k=1}^{K} \frac{Cut_b(g_k)}{vol(g_k)} \qquad (3.10)$$

Ce critère $NCut$ peut être réécrit de manière matricielle sous la forme d'un quotient de Rayleigh (Shi et Malik, 2000) faisant intervenir :

- La matrice de similarité $\boldsymbol{W}$ des arcs du graphe de données $\boldsymbol{W} = \{w(x(i), x(j))\}$, la matrice est symétrique semi-définie positive (matrice de Gram) ;
- La matrice $\boldsymbol{D}$, la matrice diagonale où d_{ii} est la somme en ligne de $\boldsymbol{W}$, soit le volume de chaque nœud (dit aussi degré) et $d_{ij} = 0$ lorsque $i \neq j$;
- La matrice $\boldsymbol{Z}$ issue de la décomposition en vecteurs propres noté z_k.

Luxburg (2007) introduit les vecteurs indicateurs f_k fonction du volume du groupe g_k et tel que $\boldsymbol{Z} = \boldsymbol{D}^{1/2}\boldsymbol{F}$, où $\mathbf{Z}$ (respectivement $\mathbf{F}$) est la matrice formée des vecteurs z_k (f_k) en colonnes.

$$NCut(\boldsymbol{X}, G) = \frac{\boldsymbol{F}^T(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{F}}{\boldsymbol{F}^T\boldsymbol{D}\boldsymbol{F}} = \frac{\boldsymbol{D}^{-1/2}\boldsymbol{Z}^T(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{Z}\boldsymbol{D}^{-1/2}}{\boldsymbol{Z}^T\boldsymbol{Z}} \qquad (3.11)$$

Pour minimiser ce critère, on impose que $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}$ identité ainsi on sélectionne les k vecteurs propres, ordonnés selon les valeurs propres décroissantes, solution du problème généralisé

$$\boldsymbol{L}z_k = e_k\boldsymbol{D}z_k \qquad (3.12)$$

avec $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ et e_k étant la valeur propre associée au vecteur propre z_k.

Les vecteurs propres $\boldsymbol{Z} = \{z_k\}$ représentent les nouvelles caractéristiques spatiales où les données pourront être simplement classées par l'algorithme K-means. Ainsi, via cette projection des données dans l'espace spectral, il est capable de classer des distributions de données aussi bien convexes que non convexes. La classification spectrale permet de classer des données localement connectées mais pas nécessairement toutes connexes. L'algorithme de la classification spectrale a été exploité dans plusieurs applications : segmentation d'image,

reconnaissance de la parole, récupération de l'information, etc. (Jain *et al.*, 1999). Récemment, des algorithmes ont été développés pour éviter les exigences de réglages : pour définir une fonction de similarité (Kong *et al.*, 2013; Zelnik-Manor et Perona, 2004) et pour déterminer le nombre de groupes. Les travaux de Yan *et al.* (2009) et Chen et Cai (2011) permettent de traiter les applications avec un volume élevé de données, ils reprennent l'algorithme de classification de Ng *et al.* (2001) présentée dans l'algorithme 3.2 qui consiste à sélectionner les k vecteurs propres de $\boldsymbol{L_{Ng}} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ ordonnés selon leurs valeurs propres croissantes.

Algorithme 3.2. Algorithme de classification spectrale.

Procédure Classification Spectrale($\boldsymbol{W}_{M\times M}, K$)
Variables : $\boldsymbol{W}, \boldsymbol{D}, \boldsymbol{L_{Ng}}, \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{l}$
$\boldsymbol{W}_{ij} = 0$
$\boldsymbol{D}$ la somme en ligne de $\boldsymbol{W}$ et $d_{ij} = 0$ avec $i \neq j$ → matrice des degrés
$\boldsymbol{L_{Ng}} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ → matrice Laplacienne
Extraire les K plus grand vecteurs propres z de $\boldsymbol{L_{Ng}}$
Former la matrice $\boldsymbol{Z} = [z_1, z_2, \dots, z_k] \in \mathbb{R}^{M\times K}$
Normaliser **Z** en ligne pour former **Y** tel que $y_{ij} = w_{ij}/\left(\sum_j z_{ij}^2\right)^{1/2}$
Appliquer K-means sur les lignes de $\boldsymbol{Y}$: $\boldsymbol{l} = Kmeans(\boldsymbol{Y}, K)$
Assignation du label $\boldsymbol{l}_i$ à chaque point original $\boldsymbol{x}(i)$ dans l'espace initial
retourner le vecteur de label : $\boldsymbol{l}$
fin de la **procédure**

Le nombre de groupes K en entrée de l'algorithme de la classification spectrale et la manière de construire la matrice de similarité de Gram $\boldsymbol{W}$ ont des effets significatifs sur le résultat de la classification. La fonction noyau gaussien est la plus largement utilisée pour la construction de $\boldsymbol{W} = \{w_{ij}\}$ définie par :

$$w_{ij} = exp\left(-\frac{\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2}{2\sigma^2}\right) \qquad (3.13)$$

Le paramètre de dispersion σ aide à creuser la matrice et tend à obtenir un cas idéal avec une décomposition en vecteurs propres exacte (dans un cas idéal, les K premières valeurs propres sont égales à un). Cependant, un mauvais choix de σ amène à une mauvaise classification. Zelnik-Manor et Perona (2004) ou Kong *et al.* (2013) ont proposé un paramètre de dispersion locale σ_i pour chaque donnée $\boldsymbol{x}(i)$ basé sur son voisinage, plutôt que de sélectionner un paramètre constant σ. La matrice de Zelnik-Manor et Perona de similarité $\boldsymbol{W}$ est choisie avec un voisinage à z voisins ($\boldsymbol{x}(nz)$ le $z^{\text{ème}}$ voisin du point $\boldsymbol{x}(i)$).

$$w_{ij} = exp\left(-\frac{\|\boldsymbol{x}(i) - \boldsymbol{x}(j)\|^2}{2\sigma_i\sigma_j}\right) \; avec \; \sigma_i = \|\boldsymbol{x}(i) - \boldsymbol{x}(nz)\| \qquad (3.14)$$

L'estimation du nombre de groupes K par l'analyse peut être déterminée selon plusieurs approches :

- L'analyse des amplitudes des valeurs propres (égal ou proche de un) où l'écart maximal entre celles-ci appelé la technique du gap.
- L'analyse des vecteurs propres :
    - Par recherche d'une base de vecteurs propres orthogonaux robuste (Kong *et al.*, 2013; Zelnik-Manor et Perona, 2004) ;
    - Par processus itératif de sélection des K premiers vecteurs propres tels qu'aucune donnée dans l'espace spectral à K vecteurs soit proche de l'origine (Sanguinetti *et al.*, 2005) ;
    - Par un processus EM sur un critère de vraisemblance basé sur la capacité de chaque vecteur propre à séparer les données (Xiang et Gong, 2008).

Pour sélectionner le nombre d'états N pour la typologie du MMC-NS, la technique du gap maximal entre valeurs propres successives est utilisée : elle est une des plus simples à implémenter et la plus rapide.

La projection des données dans l'espace des vecteurs propres du Laplacien permet d'avoir une représentation des données linéairement séparables où l'algorithme K-means permet d'obtenir efficacement et rapidement la classification voulue (figure 3.7).

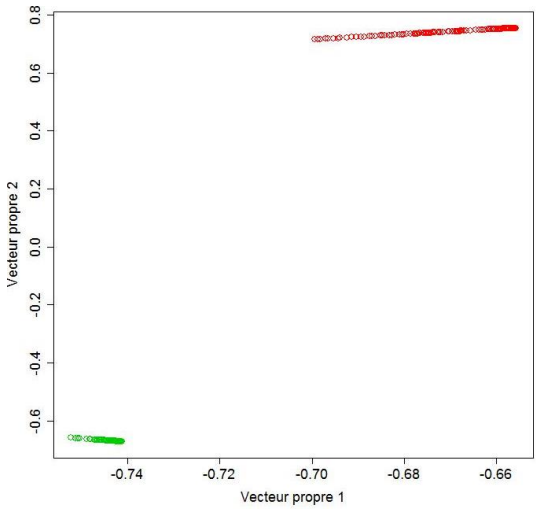


Figure 3.7. Représentation des données dans l'espace des vecteurs propres du Laplacien de l'exemple pédagogique précédent où l'algorithme K-means a été réalisé sur les données avec $k = 2$.

A partir de la classification spectrale des M symboles $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M\}$ issus de l'étape STFKM (figure 3.8), les labels $s_i = l_k$ sont affectés selon leur centre de gravité (c'est-à-dire le symbole) $\boldsymbol{v}_k$, aux données observées $\boldsymbol{X} = \{x(1), x(2), \dots, x(N_p)\}$ (figure 3.9) : le résultat de la classification est optimal. En effet, la projection des labels obtenus par classification spectrale dans l'espace initial des données montre une séparation du cercle et de la boule inscrite selon leur géométrie : le cercle extérieur représente le premier groupe et la boule inscrite représente le second.

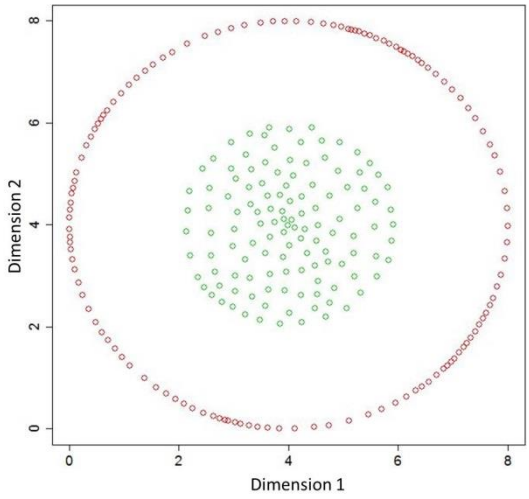


Figure 3.8. Projection de la classification spectrale sur les prototypes/symboles de l'exemple de la figure 3.5 (b).

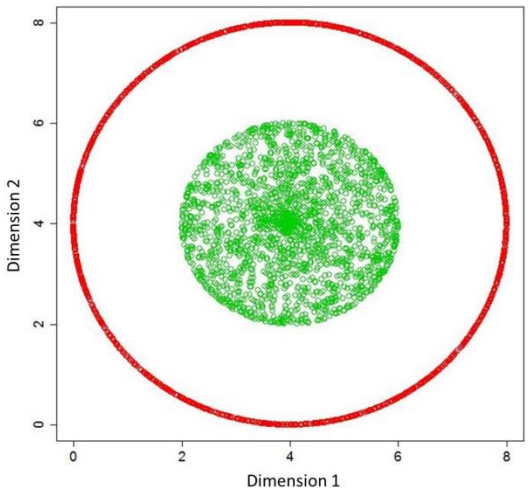


Figure 3.9. Affectation aux données brutes du jeu de l'exemple pédagogique des labels $\boldsymbol{s_i} = l_k$ selon leur symbole $\boldsymbol{v_k}$.

3.3.3. Calcul du vecteur π et des matrices A et B

Les états et les symboles ont été générés. Les matrices et le vecteur associés au MMC-NS sont maintenant calculés :

- La matrice de transition $\boldsymbol{A}$;
- La matrice d'émission $\boldsymbol{B}$;
- Le vecteur de probabilités initiales $\boldsymbol{\pi}$.

3.3.3.1. La matrice de transition A

La matrice de transition $\boldsymbol{A} = \{a_{ij}\}$, de taille $N \times N$, permet de connaître la probabilité de passer d'un état à un autre, ou de rester dans le même état (figure 3.10).

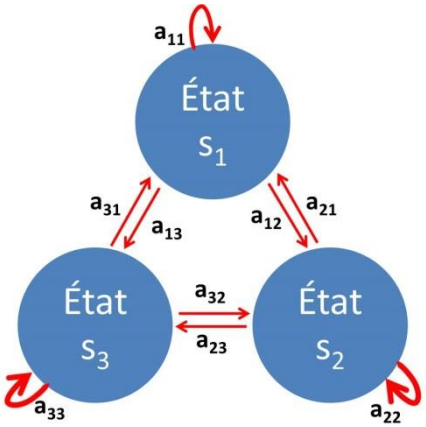


Figure 3.10. Représentation des probabilités de transition $\boldsymbol{a_{ij}}$ (en rouge) entre les états $\boldsymbol{s_i}$ d'un système Markovien.

Nous utilisons un Modèle de Markov Caché d'ordre 1, cela signifie que l'état à l'instant t ne dépend uniquement que de l'état à l'instant précédant $t-1$.

$$a_{ij} = P\left(s_j(t)|s_i(t-1), \dots, s_u(1)\right) = P\left(s_j(t)|s_i(t-1)\right), 1 \leq i, j \; et \; u \leq N \quad (3.15)$$

Pour obtenir cette probabilité, le nombre de transitions d'un état s_i à un état s_j sur l'ensemble de nos observations **X** est sommé. Lorsque cette somme est terminée, la matrice est normalisée en ligne.

$$\sum_{j=1}^{N} a_{ij} = 1 \quad (3.16)$$

3.3.3.2. La matrice d'émission B

La matrice d'émission $\boldsymbol{B} = \{b_{ik}\}$, de taille $N \times M$, permet de connaître la probabilité d'être dans un état s_i et dans un symbole $\boldsymbol{v}_k$ au même instant t.

$$b_{ik} = P\left(\boldsymbol{v}_k(t)|s_i(t)\right) \quad (3.17)$$

Un état étant constitué d'un seul ou d'un ensemble de symboles pouvant être commun avec d'autres états (figure 3.11), cette probabilité est obtenue en sommant ces apparitions simultanées sur la base **X**.

$$\sum_{i=1}^{N}\sum_{j=1}^{M} b_{ij} = 1 \quad (3.18)$$

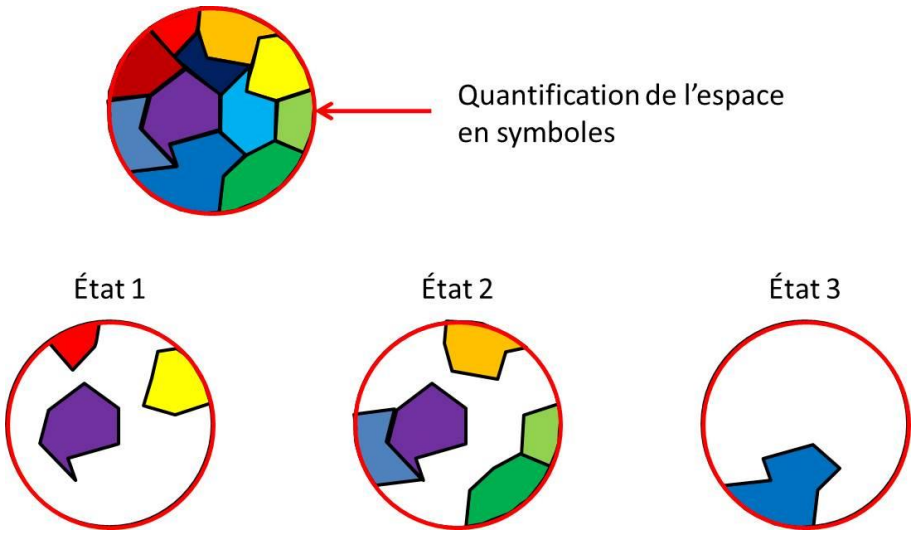


Figure 3.11. Un état (le cercle) est constitué d'un ou de plusieurs symboles (polygone à l'intérieur du cercle).

3.3.3.3. Le vecteur de probabilités initiales π

Le vecteur de probabilités initiales, de taille N, permet d'estimer l'état le plus probable dans lequel sera le premier instant d'une nouvelle séquence $\boldsymbol{X}(T)$. Lorsque cette séquence est la suite temporelle de la base de données $\boldsymbol{X}$, la dernière donnée de $\boldsymbol{X}$ avec son état associé peuvent être utilisés (figure 3.12). Ainsi, si dans un système à 2 états, l'état associé à la dernière donnée de $\boldsymbol{X}$ est s_1, alors $\boldsymbol{\pi} = \{P(s_1) = 1, P(s_2) = 0\}$.

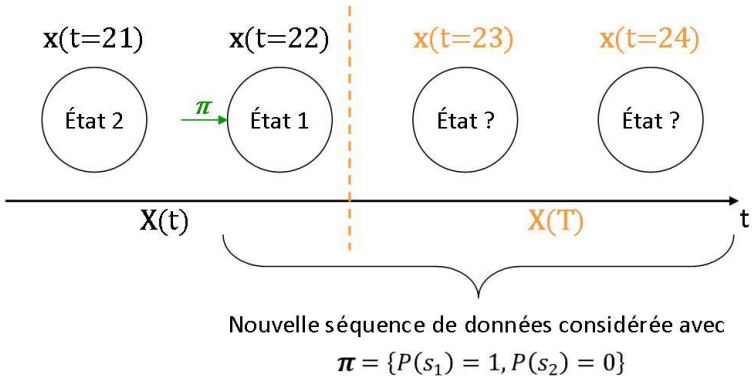


Figure 3.12. Estimation du vecteur de probabilités initiales $\boldsymbol{\pi}$ lorsque les nouvelles données acquises sont dans la continuité temporelle de la base de données initiales.

La séquence de données $\boldsymbol{X}(T)$ n'étant pas obligatoirement la suite temporelle de la base de données où a été construit le modèle λ, le vecteur de probabilités initiales $\boldsymbol{\pi}$ possède des valeurs équiprobables. Dans un système à 2 états, $\boldsymbol{\pi} = \{P(s_1) = 0{,}5\,, P(s_2) = 0{,}5\}$ (figure 3.13).

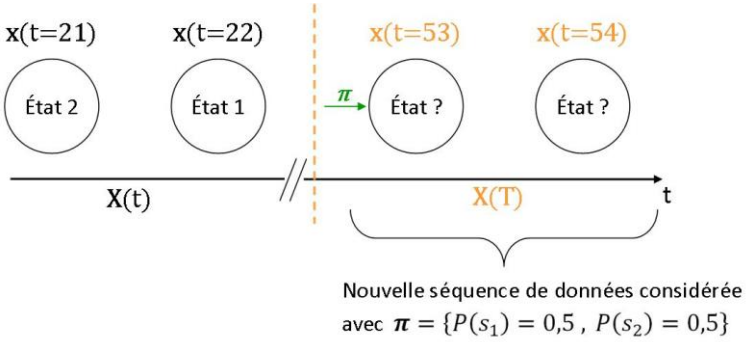


Figure 3.13. Estimation du vecteur de probabilités initiales π lorsque les nouvelles données acquises ne sont pas dans la continuité temporelle de la base de données initiales.

3.4. Prédiction

Les paramètres du Modèle de Markov Caché ont été calculés sur la base des observations $\boldsymbol{X}$. A partir de ce modèle λ, il est possible de prédire / estimer l'état dans lequel se trouve une nouvelle donnée ou les états d'une séquence de nouvelles données.

Nous recherchons à maximiser la probabilité d'être dans un état sachant les observations et le modèle λ :

$$\begin{aligned} &\arg\max_i P(s(T) = s_i, x(1) \dots x(T)|\lambda) \sim \arg\max_i P_\lambda\big(s(T) = s_i, x(1) \dots x(T)\big) \\ &\sim \arg\max_i P_\lambda(x(1) \dots x(T)|s(t) = s_i) \times P(s(T) = s_i) \end{aligned} \qquad (3.19)$$

L'optimisation de cette équation nécessite deux étapes :

1. Assignation du symbole le plus proche à chaque donnée $x(t)$ de la séquence dans l'espace initial pour la partie gauche de l'équation.
2. Calcul par l'algorithme de Viterbi (algorithme 3.3) (Forney, 1973; Viterbi, 1967) du séquencement d'états le plus probable sachant λ et le symbole le plus proche.

Algorithme 3.3. Algorithme de Viterbi.

Entrée :

Le modèle $\lambda(\boldsymbol{S}, \boldsymbol{V}, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$

La séquence $\boldsymbol{C} = \{c(1), \dots, c(T) \; ou \; c(t) = \boldsymbol{v}_j \; avec \; j = 1, \dots, M\}$ de longueur T des symboles $\boldsymbol{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_M\}$ associés à chaque nouvelle donnée

Sortie :

Le chemin d'états optimal τ

Variables :

$\boldsymbol{\delta}$ la matrice de taille $N \times T$

h un tampon de mémorisation

Initialisation :

$\delta_i(t = 1) = \pi_i \times b_{ic(t=1)}$ avec $i = 1, \dots, N$

Itération :

Pour tout t allant de 2 à T :

$h = \max(\delta_j(t-1) + a_{ji})$ avec $i, j = 1, \dots, N$

$\delta_i(t) = b_{ic(t)} + h$ avec $i = 1, \dots, N$

Recherche du chemin optimal $\boldsymbol{\tau}$:

$\tau(t = T) = \arg\max_i(\delta_i(t = T))$ avec $i = 1, \dots, N$

Pour tout t allant de $(T-1)$ à 1 :

$\tau(t) = \arg\max_i(\delta_i(t) + a_{i\tau(t+1)})$ avec $i = 1, \dots, N$

L'étape 1 correspond à l'affectation des symboles aux nouvelles données entrantes ; elle est réalisée par l'algorithme du plus proche voisin : chaque donnée est associée à son symbole le plus proche dans l'espace initial des données. Soit la séquence de symboles $\boldsymbol{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_M\}$ associés à chaque nouvelle donnée $x(T)$ et son assignation notée :

$$\boldsymbol{C} = \{c(1), \dots, c(T) \; ou \; c(t) = \boldsymbol{v}_j \; avec \; j = \{1, \dots, M\}\} \qquad (3.20)$$

Nous avons donc :

$$P_\lambda(x(1) \dots x(T) | s(T) = s_i) = P_\lambda(\boldsymbol{C} | s(T) = s_i) \qquad (3.21)$$

La seconde étape recherche la séquence d'états optimale appelée le chemin d'états $\boldsymbol{\tau}$.

$$P(s(T) = s_i) = P(s(t) = s_i | s(t-1) = s_j) \times P(s(t-1) = s_j) \qquad (3.22)$$

Par récurrence, on obtient :

$$P(s(T) = s_i) = \prod_{t=T}^{2} P(s(t) = s_i | s(t-1) = s_j) \times P(s(t=1) = s_i) \qquad (3.23)$$

Ce protocole du système de décision, soit d'estimation d'un ou plusieurs états, peut être résumé par la figure 3.14 :

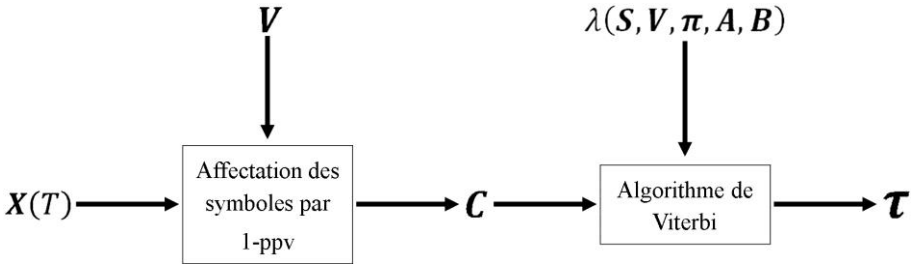


Figure 3.14. Système de décision : estimant une séquence d'états $\boldsymbol{\tau}$ pour de nouvelles données $\boldsymbol{X}(T)$ en utilisant le MMC-NS et l'algorithme de Viterbi.

Prenons un exemple pédagogique, notre MMC hybride a été réalisé sur un jeu de données. Il en est sorti 2 états. Trois nouvelles données sont acquises ($T = 3$) : la séquence $\boldsymbol{C}$ est calculée. La matrice $\boldsymbol{\delta}$ de taille $N \times T$, correspondant au score maximal de passage d'un état à un autre prenant en compte les transitions et les émissions, doit maintenant être construite (figure 3.15). On commence par l'étape d'initialisation où chaque δ_{i1} est calculé uniquement à partir sa probabilité initial π_i et la probabilité d'émission associée à l'instant 1 et à l'état i, avec $i = 1\ ou\ 2$. Pour les instants suivants, δ_{it} représente la probabilité d'émission associée à l'instant t et à l'état i additionné au maximum de la somme entre $\delta_{j(t-1)}$ et la probabilité de transition a_{ji} avec j l'état de départ de cette transition.

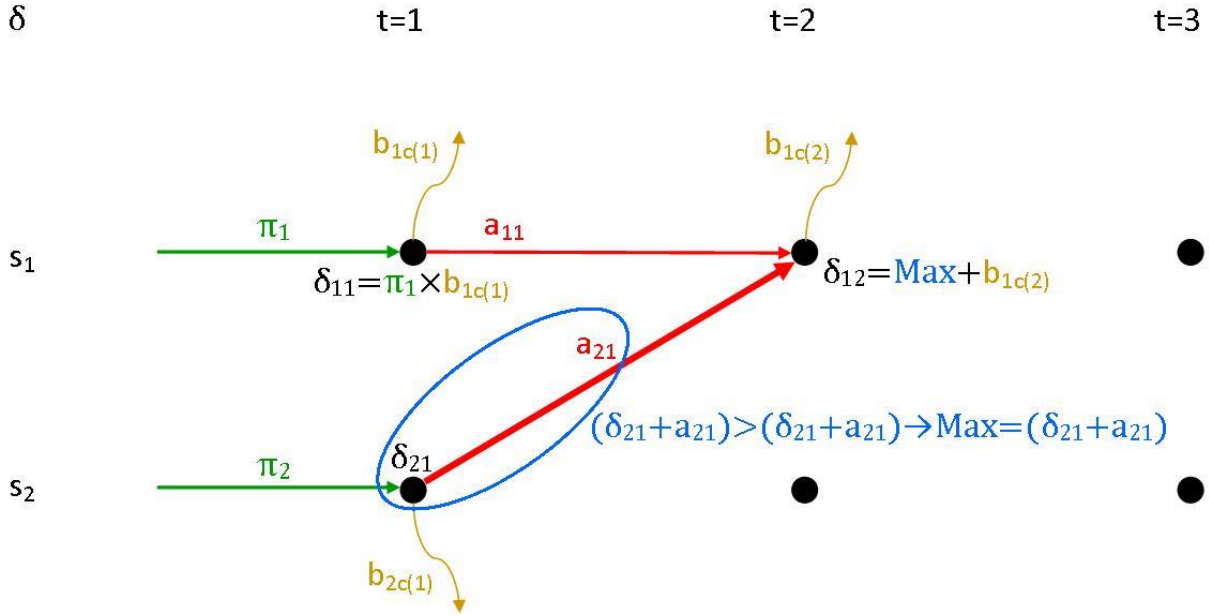


Figure 3.15. Construction de la matrice $\boldsymbol{\delta}$ correspondant aux étapes d'initialisation et d'itération de l'algorithme 3.3 avec les probabilités initiales $\boldsymbol{\pi_i}$ (vert), de transitions $\boldsymbol{a_{ij}}$ (rouge), et d'émissions $\boldsymbol{b_{ic(k)}}$ (marron).

Maintenant que notre matrice δ est construite, le chemin d'états optimal $\boldsymbol{\tau}$ peut être recherché (figure 3.16). Contrairement à la première partie de cette algorithme, la recherche du chemin se réalise en commençant par le dernier instant acquis ($t = T = 3$). L'état i possédant la plus

grande valeur de δ_{iT} est associé à $\tau(t=T)$: δ_{13} dans notre exemple. A partir de ce moment, les autres valeurs δ_{uT} avec $u \neq i$, ne sont plus utilisées : δ_{23} dans notre exemple. Le prochain état à estimer se situe à l'instant $t = T - 1$. L'état i à l'instant t possédant la plus forte valeur $(\delta_{it} + a_{ij})$, avec j l'état associé à $\tau(t+1)$, est alloué à $\tau(t)$. Comme précédemment, les autres valeurs δ_{uT} avec $u \neq i$, ne sont plus utilisées. Dans notre exemple, nous avons donc $(\delta_{22} + a_{21})(\delta_{12} + a_{11}) \rightarrow \tau(t=2) = s_2$. Cela continue jusqu'à $t = 1$ inclus et nous avons pour les nouvelles données acquises $\boldsymbol{X}(T)$ la séquence d'états correspondant au chemin d'état optimal $\boldsymbol{\tau}$.

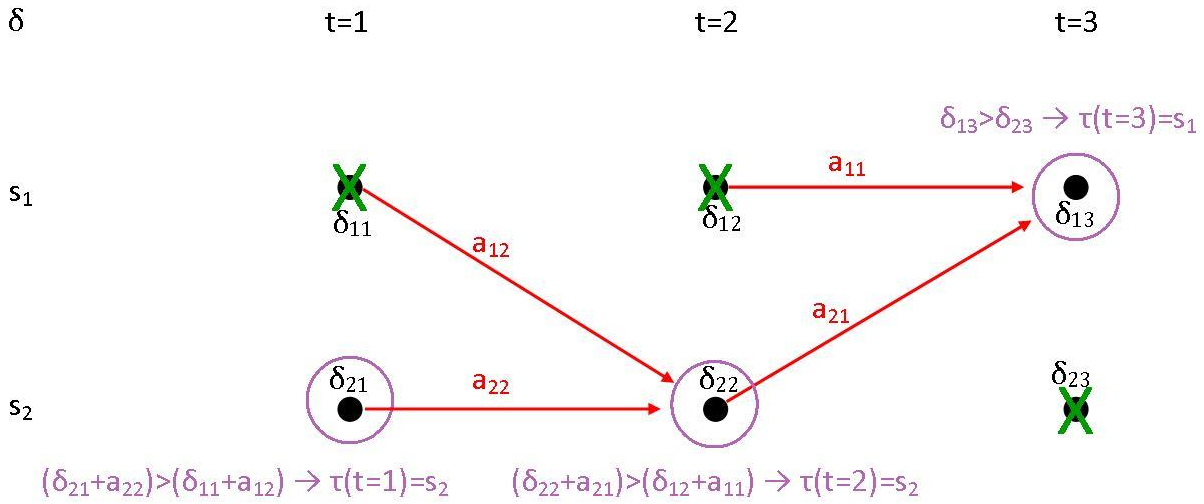


Figure 3.16. Recherche du chemin d'états optimal $\boldsymbol{\tau}$ (cercle violet) en fonction de la matrice δ et des probabilités de transitions $\boldsymbol{a_{ij}}$ (rouge).

Les nouvelles données acquises $\boldsymbol{X}(T)$ peuvent contenir des valeurs manquantes. Ces valeurs ne sont pas prises en compte dans l'algorithme 3.3. C'est pourquoi, lors de la recherche des états des nouvelles observations, certaines erreurs peuvent apparaitre. En effet, si la donnée à $t = 6$ est manquante, alors l'algorithme utilisera la probabilité de transition a_{ij} correspondant au passage de l'état i à l'instant $t = 5$ à l'état j à l'instant $t = 7$. Pour résoudre ce problème, un découpage des nouvelles données $\boldsymbol{X}(T)$ en plusieurs séquences ne contenant aucune valeur manquante peut être réalisé.

3.5. Conclusion

La dynamique des efflorescences du phytoplancton est un processus complexe, l'état environnemental courant ne dépend pas uniquement des observations (paramètres physiques) mais de la dépendance de leur succession. Nous avons donc proposé, pour modéliser ce phénomène, de caractériser cette dynamique par un Modèle de Markov Caché ergodique Non Supervisée (MMC-NS) noté $\lambda(S, V, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$, c'est-à-dire sans aucune connaissance *a priori* sur nos données. La particularité de notre approche est de pouvoir proposer une méthodologie de construction des paramètres d'un MMC de façon automatique et non supervisée.

La construction de la structure a été basée sur un principe de minimisation optimal du nombre de paramètres du MMC-NS et une caractérisation des états et symboles fidèle à la distribution des données tout en conservant la nature haute fréquence de celles-ci.

Les états $\boldsymbol{S} = \{s_1, s_2, \dots, s_N\}$ sont calculés par la classification spectrale des symboles. Cette classification transforme l'espace initial des données d'entrées dans un nouvel espace où les données sont linéairement séparables.
Les symboles $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_M\}$ sont générés via une quantification vectorielle. L'algorithme des K-means est bien adapté à la méthode de quantification vectorielle et est populaire dans la classification non supervisée. Cet algorithme est modifié pour inclure l'automatisation de la recherche du nombre de groupes / symboles M basé sur le critère d'Elbow appelé algorithme Self Tuning Fast K-means (STFKM). Les paramètres probabilistes du MMC-NS sont estimés par un unique comptage et normalisés sur la base des observations.

A partir de ce modèle construit, un système de décision a été proposé. Découpé en deux phases, ce système recherche l'assignation d'une nouvelle donnée à un symbole appris puis détermine l'état de cette donnée par maximum de vraisemblance.
Le système de décision fournit l'estimation de l'état d'une nouvelle donnée uniquement lorsque l'observation de cette nouvelle donnée est cohérente vis-à-vis du symbole assigné dans la première étape du système. Il rejettera toute nouvelle donnée éloignée de la quantification de l'espace appris afin de ne pas donner de fausses estimations.

L'ensemble des choix de notre méthodologie, sélection des symboles par l'algorithme STFKM et la génération des états par classification spectrale est évalué dans le chapitre suivant. Pour cela nous avons tout d'abord fixé un nombre d'états et établi un vecteur d'états de référence selon un découpage réalisé par la DCE - 2000/60/CE. Nos méthodes peuvent ainsi être comparées aux approches usuelles de la littérature. Le séquencement des états des données mesurées par la station MAREL-Carnot, estimé par MMC-NS, sera confronté avec une interprétation d'experts.

Chapitre 4 : Application aux données de la station de mesure MAREL-Carnot

4.1. Introduction

Le chapitre précédent décrit la méthodologie proposée pour modéliser un phénomène physique complexe par un graphe probabiliste d'états finis à partir d'observations temporelles multi-paramètres sans autre connaissance *a priori*. Les différentes étapes de la construction de ce graphe basé sur un Modèle de Markov Caché ergodique hybridé (MMC-NS) ont été menées de manière totalement non supervisée. Pour analyser la pertinence de l'approche proposée et notamment la robustesse des générations de symboles et des états du MMC, il est important de comparer, dans un premier temps chaque étape par des algorithmes classiques et dans un second temps de comparer le résultat du modèle vis-à-vis des applications envisagées. Cette étape de validation sera menée sur l'application principale de cette thèse et notamment devra répondre aux questions suivantes :

- Le modèle construit est-il capable de modéliser la dynamique des efflorescences phytoplanctoniques dans les eaux de Boulogne-sur-Mer via les séries de données à haute résolution temporelle enregistrées par la station MAREL-Carnot ?
- Le modèle est-il capable de prédire un état environnemental particulier tel qu'une efflorescence printanière ou automnale, un état particulier (notion d'événement extrêmes) ?

La validation des algorithmes proposés dans un cadre non supervisé sans labellisation est difficile ; un partitionnement sera alors jugé sur des critères géométriques comme l'homogénéité d'un groupe. Dans ce sens, une première étude sera conduite via une labellisation arbitraire biclasse permettant de comparer les algorithmes non supervisés aux techniques de classification supervisée usuelles. La validation du modèle MMC construit à N états fixés sera quantifiée sur la capacité du système à estimer une nouvelle observation par reconstruction guidée par la sortie du MMC-NS.
Ensuite, un modèle MMC-NS construit de manière totalement non supervisée (N états non fixés) sera qualifié par une interprétation à dire d'experts.

Ce chapitre est donc découpé de la manière suivante. La première section reprend la caractérisation des données de la station MAREL-Carnot et les prétraitements nécessaires (Chapitre 2). La deuxième section présente et valide la topologie obtenue à partir du MMC-NS à 2-états fixés comparées à des techniques d'apprentissage supervisé. Dans la troisième section, les états obtenus par un modèle MMC-NS à N-états sont interprétés du point de vue de leurs significations écologiques. Pour finir, un MMC-NS a été construit pour chaque année

afin de comparer les résultats à ceux obtenus à partir du modèle intégrant la période 2005-2008.

4.2. Prétraitement des données MAREL-Carnot

Les étapes essentielles de prétraitement des données avant toute étape de classification automatique sont détaillées au chapitre 2. Nous reprenons ici les différents prétraitements et le protocole associés aux données utilisées, les mesures multi-capteurs issues de la station MAREL-Carnot, détaillées au chapitre 1, pour :

- Corriger les valeurs hors gamme et les paramètres bruitées ;
- Aligner les capteurs sur une échelle de temps identique ;
- Compléter les données manquantes par une moyenne mobile contrainte ;
- Extraire les paramètres pertinents par une étude des paramètres corrélés.

4.2.1. Extraction et correction des données

4.2.1.1. Extraction des données

Nous rappelons que nous utilisons les données acquises durant la période 2005-2009. La base initiale contient 19 paramètres physico-chimiques : 16 de ces paramètres sont prélevés à une fréquence de 20 minutes et 3 à une fréquence de 12 heures (concentrations en nutriments).

Sur ces 19 paramètres disponibles, trois ne seront pas utilisés :

- Le pH possède un nombre de données non valides important (défaut de capteur). Aucune correction ne peut être apportée à cette mesure ;
- La hauteur d'eau mesurée : le signal mesuré possède une composante de bruit très importante liée à un problème technique ;
- La direction du vent est une variable circulaire variant de 0 à 360°. Numériquement le passage de 359 à 1 est très grand, alors qu'il ne correspond qu'à une différence de deux degrés et indique un vent de direction Nord. Pour essayer de contourner ce problème, le signal a été découpé en deux composantes, sinus et cosinus, mais cela n'apporte pas de réelle solution puisque qu'aucune relation entre ces deux composantes n'est utilisée lors de la classification. Ainsi, bien que certains organismes, dont Météo France, utilisent le signal brut en rose de 360 degrés, nous avons préféré le supprimer pour ne pas fausser notre classification : aucune solution de remplacement n'a été trouvée.

4.2.1.2. Correction des données

Sur l'ensemble des paramètres sélectionnés, ne sont retenues que les valeurs comprises dans la gamme de mesure du capteur et /ou définie comme pertinente par l'expert (Chapitre 1) : une valeur dite manquante (NA) est attribuée aux autres.

Une vérification supplémentaire est appliquée au paramètre de fluorescence. Elle consiste à valider ou non de la présence d'un effet quenching entrainant une diminution des valeurs de fluorescence.

L'effet quenching

Nous avons vu dans le chapitre 1 que la fluorescence est une émission lumineuse provoquée par l'excitation des pigments algaux. Cette excitation est due à l'absorption de photons à une certaine longueur d'onde dans le cadre du processus de la photosynthèse. Cependant, une limitation de celle-ci peut être observée. Elle régule et protège la cellule phytoplanctonique lorsque l'énergie lumineuse absorbée excède la quantité de stockage d'énergie utilisable. Une diminution de la fluorescence est observée : c'est l'effet quenching (figure 4.1) (Halverson et Pawlowicz, 2013; Müller *et al.*, 2001). Lorsqu'il apparaît, il est nécessaire de corriger les données de fluorescence. Halverson et Pawlowicz (2013) supposent que la biomasse phytoplanctonique évolue quasi-linéairement durant une journée. De ce fait, la correction des données de fluorescence consiste à effectuer une interpolation linéaire entre les mesures effectuées avant et après le pic de luminosité (représentée par le trait bleu sur la figure 4.1).

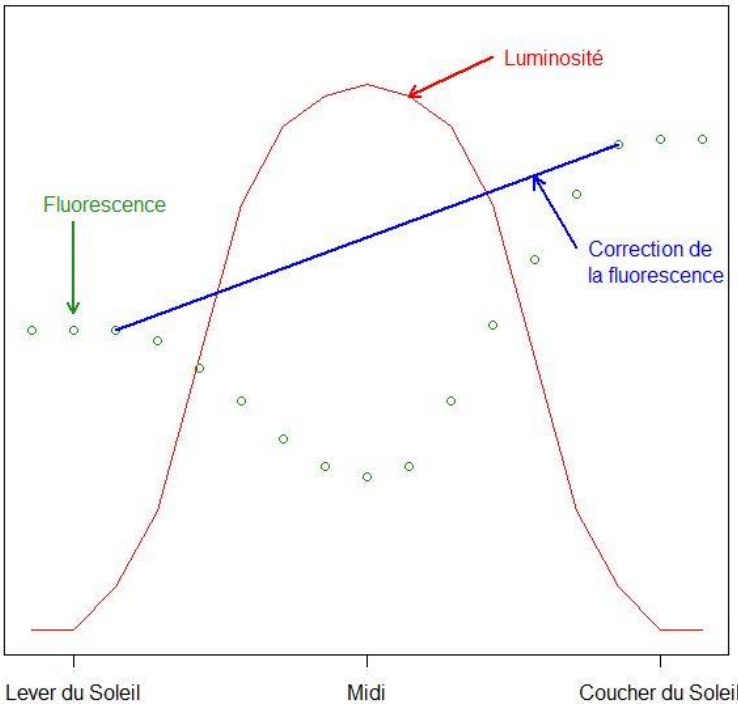
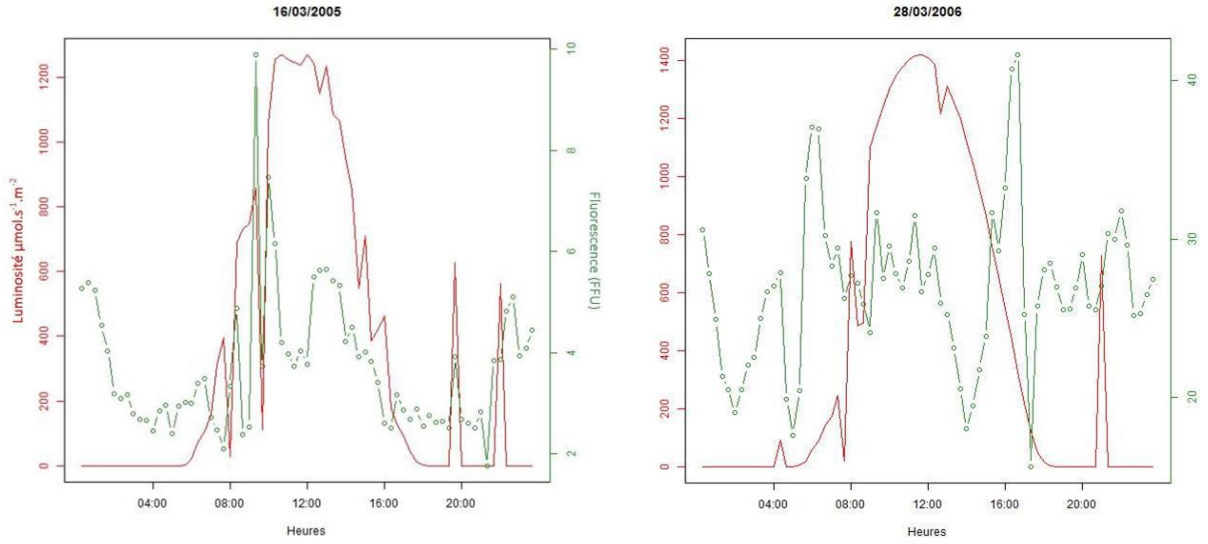


Figure 4.1. Représentation schématique de l'effet de quenching : diminution erronée de la biomasse phytoplanctonique représentée par la fluorescence lors de l'une saturation de la luminosité absorbée par les cellules phytoplanctoniques.

La station MAREL-Carnot effectue des mesures en continue (jour et nuit), nous pouvons donc nous demander si la fluorescence mesurée par la station reflète un effet quenching. Les données de la fluorescence, au moment des développements phytoplanctoniques de plus grande ampleur, sur la période 2005-2009, sont analysées pour plusieurs jours et sur plusieurs semaines. La variabilité des données étant importante, les résultats ne seront illustrés que sur

deux jours pour faciliter la visualisation du phénomène sur les données *in-situ* : le 16/03/2005 et le 27/03/2006 (figure 4.2). L'évolution de la fluorescence sur ces périodes ne présente pas de diminution générale lorsque la luminosité est à son maximum.



(a) (b)

Figure 4.2. Étude de l'effet quenching sur les données de la station MAREL-Carnot en analysant la relation entre l'évolution de la luminosité (µmol de photons .s^{-1}.m^{-2}, en rouge) et de la fluorescence (FFU, en vert) sur deux jours représentatifs d'un développement de biomasse phytoplanctonique : le 16/03/2005 et le 27/03/2006.

Nous pouvons donc conclure que la fluorescence mesurée par la station MAREL-Carnot pour les années 2005 à 2009 ne traduit pas d'effet quenching et par conséquent, aucune correction n'est nécessaire sur ces données.

4.2.2. Alignement temporel

La fréquence d'acquisition des données MAREL-Carnot est de 20 minutes. Cependant, les mesures des différents capteurs ne se font pas simultanément, il existe donc un décalage temporel pouvant aller de quelques secondes à quelques minutes. Les caractéristiques de l'eau de mer ne se modifient pas radicalement toutes les minutes, ainsi afin de synchroniser les mesures, après suppression de l'information sur les secondes, un alignement temporel par intervalle de 20 minutes est effectué. Pour chaque heure notée hh, on obtient l'alignement suivant :

- [hh:00, hh:20 [= hh:10
- [hh:20 , hh:40 [= hh:30
- [hh:40 , hh:59] = hh:50

Ainsi l'instant 01:21 sera étiqueté comme étant l'instant 01:30. Notre base de données est par conséquent constituée de $N_p = 131\,472$ instants débutant le 01/01/2005 à 00:10 et finissant le 31/12/2009 à 23:50 avec un pas d'échantillonnage de 20 minutes. Pour les nutriments dont

la fréquence est moins importante (12 heures), l'approche est différente et est présentée dans la partie suivante.

4.2.3. Complétion des données manquantes par moyenne mobile

Différentes méthodes d'imputation ont été décrites dans le chapitre 2. Nous avons retenu l'imputation par moyenne mobile pour compléter les données de la station MAREL-Carnot.
La moyenne mobile est appliquée uniquement sur les données manquantes afin de ne pas lisser les données et notamment pour ne pas perdre l'information haute fréquence. Cette complétion est opérée en deux étapes successives :

1) Régularisation des données associées aux nutriments afin de s'aligner sur la fréquence des autres paramètres.
2) Imputation des valeurs manquantes.

La première étape a pour but d'aligner les données associées aux nutriments à la fréquence de 20 minutes, puisque les analyseurs de nutriments fournissent une à deux valeurs (réplicat) toutes les 12 heures. Une moyenne mobile sur une fenêtre de 40 points équivalent à 12 heures est utilisée (Chapitre 2). A partir de cette étape, la base de données de 2005 à 2008 contient 105 192 points dans $\mathbb{R}^{16}$. La moitié des instants de cette base (48 157 points) possède au moins une donnée manquante sur l'un de ses paramètres (due à un problème de capteur). Notre modélisation étant basée sur des vecteurs d'observations complets, il paraît essentiel de conserver la richesse de notre base et d'imputer les données manquantes ponctuelles.
La seconde étape consiste à appliquer sur chaque paramètre de la base, une imputation des données manquantes via une moyenne mobile calculée sur une fenêtre de 500 points équivalent à une semaine. Cette taille de fenêtre a été retenue puisqu'elle correspond à l'échelle temporelle moyenne d'une efflorescence phytoplanctonique.
Suite à ces deux étapes de complétion, cela conduit à obtenir une base de données de $N_p = 84\,614$ points sans valeurs manquantes.

4.2.4. Analyse en Composantes Principales

Une Analyse en Composantes Principales (ACP) est réalisée afin d'étudier les corrélations entre les paramètres dans l'optique de réduire notre base de données. Cela permettra également de diminuer les temps de calcul. L'ACP est une méthode mathématique permettant de décrire un tableau de données de type individu / variables (Legendre et Legendre, 1998). Elle consiste à chercher un sous-espace F_k de dimension k inférieur à celui de l'espace de départ, tel que le nuage, une fois projeté dans ce sous-espace, soit au minimum déformé. Ainsi elle propose une représentation des données dans un espace réduit où les directions de ce nouvel espace représentent aux mieux les corrélations entre n variables et mettent en évidence la structure des données.

4.2.4.1. Analyse des résultats de l'ACP

La base de données contenant $N_p = 84\,614$ points dans $\mathbb{R}^{16}$ est analysée. La variance expliquée cumulée atteint les 54 % sur les trois premières composantes principales et 99 % lorsque l'on atteint la onzième composante (figure 4.3).

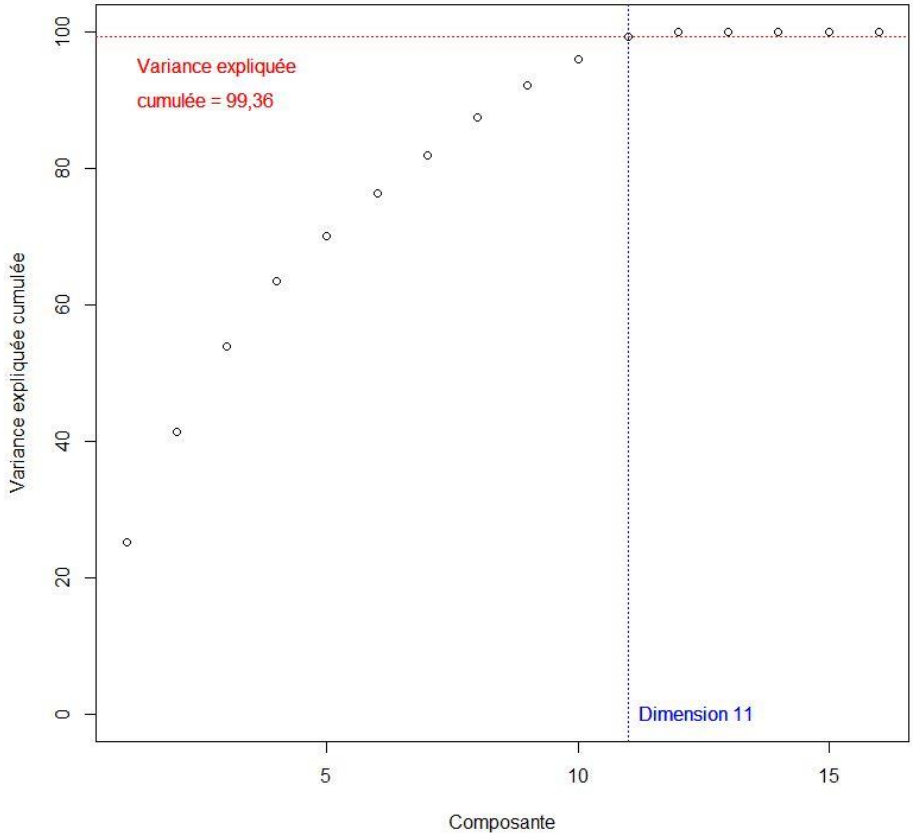


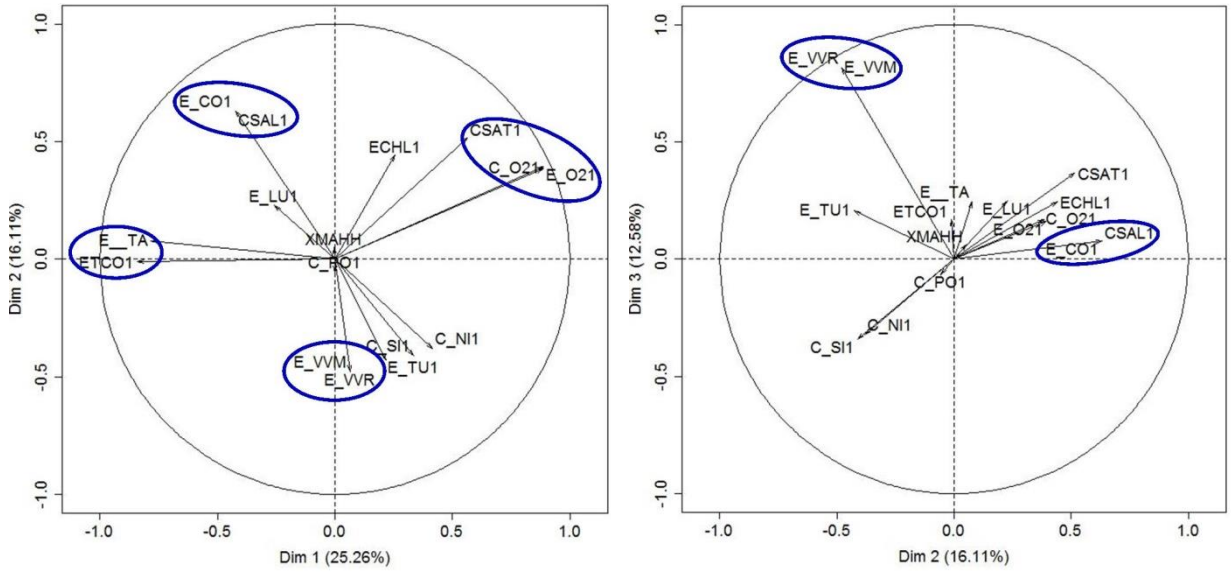
Figure 4.3. ACP sur la base $N_p = 84\,614$ points dans $\mathbb{R}^{16}$. Évolution du pourcentage de variance expliquée cumulée en fonction de chaque composante principale.

Les cercles de corrélations permettent de visualiser rapidement les différents paramètres corrélés entre eux, ainsi que ceux qui sont structurants pour chaque composante obtenue (figure 4.4).

Les couples de paramètres corrélés entre eux avec leur coefficient de corrélation de Pearson ρ associé avec sa p-value (significativité) sont listés ci-après :

- La concentration en oxygène dissous corrigée (C_O21) et la concentration en oxygène dissous non corrigée (E_O21) : $\rho = 1^{***}$;
- La concentration en oxygène dissous corrigée (C_O21) et la saturation en oxygène (CSAT1) : $\rho = 0{,}84^{***}$;
- La concentration en oxygène dissous non corrigée (E_O21) et la saturation en oxygène (CSAT1) : $\rho = 0{,}82^{***}$;
- La salinité (CSAL1) et la conductivité (E_CO1) : $\rho = 1^{***}$;
- La température de l'eau (ETCO1) et la température de l'air (E__TA) : $\rho = 0{,}89^{***}$;

- La vitesse du vent en moyenne (E_VVM) et la vitesse du vent en rafale (E_VVR) : $\rho = 1^{***}$.



(a) (b)

Figure 4.4. Cercles de corrélations issues de l'analyse en composantes principales des données issues de la station MAREL-Carnot sur la période 2005-2008 ($N_p = 84\,614$ points dans $\mathbb{R}^{16}$) sur les dimensions 1 et 2 (a), ainsi que les dimensions 2 et 3 (b). Les cercles bleus rassemblent les paramètres corrélés entre eux.

Ces résultats sont à remettre en relation avec les équations données au chapitre 1. En effet, l'oxygène dissous corrigé est calculé à partir des paramètres de l'oxygène dissous non corrigé, la salinité elle-même calculée à partir de la conductivité. Ces dernières sont donc des informations redondantes.

La saturation en oxygène est calculée à partir de la salinité, la température de l'eau et l'oxygène dissous non corrigé.

Nous décidons de ne garder qu'un seul paramètre par couple, c'est-à-dire la concentration en oxygène dissous corrigée, la salinité, la température de l'eau, la vitesse du vent en rafale. Les résultats obtenus ici étaient attendus mais permet d'illustrer la démarche de sélection des paramètres.

Après l'analyse de corrélation, $D_p = 10$ paramètres physico-chimiques, non corrélés noté NC, sont retenus et détaillés dans la colonne NC du tableau 4.1. Il faut noter que le signal de fluorescence a été volontairement écarté des paramètres conservés pour construire notre modélisation. Étant un indicateur de biomasse phytoplanctonique, celui-ci sera utilisé pour valider le résultat de classification puisque nous n'avons pas d'expertise sur ce que devrait être la dynamique des efflorescences du phytoplancton. Par conséquent, il nous permettra de valider nos résultats sans biaiser le système.

Tableau 4.1. Liste des signaux mesurés par la station MAREL-Carnot avec leur acronyme associé. Les paramètres retenus pour les analyses et le développement du modèle sont marqués en vert dans la colonne NC (Non Corrélé).

Paramètre	Acronyme	NC (Non Corrélé)
Oxygène dissous corrigé	C_O21	
Oxygène dissous non corrigé	E_O21	
Saturation en oxygène	CSAT1	
Fluorescence	ECHL1	
pH	E_PH1	
Salinité	CSAL1	
Conductivité	E_CO1	
Température de l'eau	E__TA	
Température de l'air	ETCO1	
Hauteur d'eau	XMAHH	
Vitesse du vent en moyenne	E_VVM	
Vitesse du vent en rafale	E_VVR	
Direction du vent	E_VDM	
P.A.R.	E_LU1	
Turbidité	E_TU1	
Concentration en Nitrate	C_NI1	
Concentration en Phosphate	C_PO1	
Concentration en Silicate	C_SI1	

4.2.4.2. Limite de la classification à partir de l'Analyse en Composantes Principales

L'analyse en composantes principales permet d'étudier les relations entre les variables. De ce fait, est-il possible de réaliser une classification en se basant sur les projections des paramètres dans l'espace des composantes principales (dimension 1 à i) afin d'obtenir une information sur la dynamique de l'efflorescence mais également sur des évènements rares (ouverture du barrage) ? L'ACP est réalisée sur la base de données contenant $N_p = 84\,614$ points dans $\mathbb{R}^{10}$. Les mois, les années et la fluorescence sont projetés en paramètres supplémentaires. Un découpage peut être réalisé en fonction de la proximité des directions des paramètres selon les dimensions (cosinus au carré). Le résultat de classification obtenu découpe la base de données selon les saisons de l'année (figure 4.5) : printemps, été, automne-hiver.

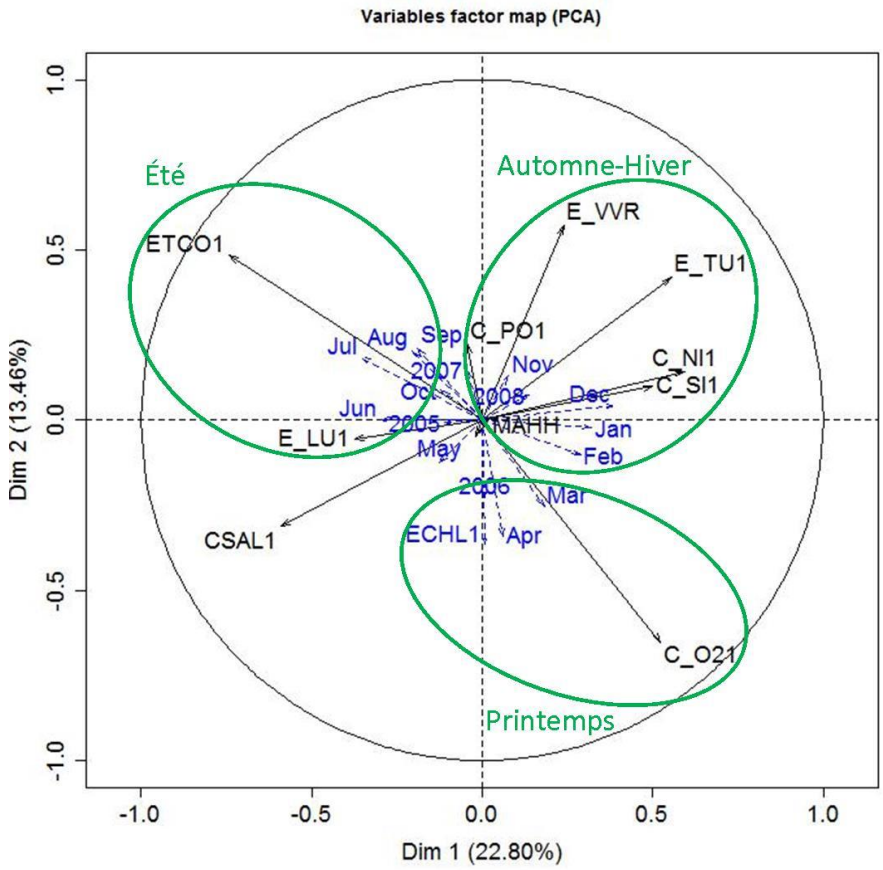


Figure 4.5. ACP sur $N_p = 84\,614$ points dans $\mathbb{R}^{\mathbf{10}}$. Projection sur les deux premières composantes (Dim 1 et Dim 2) de l'ACP des paramètres originaux en noir et des paramètres supplémentaires en bleu. Les cercles verts représentent la classification obtenue lorsque les directions des paramètres sont utilisées.

Ce découpage (ou classification) basé sur des données acquises à haute fréquence n'apporte aucune information supplémentaire par rapport à ce qui est mis en évidence à partir d'une approche conventionnelle, basse fréquence, à savoir une dynamique saisonnière du phytoplancton : efflorescence au printemps, période non productive en automne-hiver. Un autre découpage possible serait de ne pas classer selon la projection des paramètres (cosinus au carré) mais d'appliquer un algorithme de classification non supervisée sur les composantes de l'ACP (exemple du K-means). Or les données ne seront pas obligatoirement linéairement séparables pour chaque plan (projection des dimensions deux à deux) : les données peuvent donc avoir une forme non-convexe. Un algorithme tel que celui du K-means ne permettra pas d'obtenir un partitionnement optimal vis-à-vis de la structure géométrique des données. Pour lever l'hypothèse sur la forme des données, nous choisissons d'utiliser la classification spectrale (chapitre 3) sur la base de données NC dans l'espace initial ($N_p = 84\,614$ points dans $\mathbb{R}^{10}$) par la suite.

4.3. Validation d'un Modèle de Markov Caché non supervisé à 2-états fixés

Dans le but de tester notre système hybridé, il était nécessaire d'obtenir une expertise sur les états environnementaux recherchés. L'exemple le plus simple a été de considérer la stratégie de la Directive Cadre sur l'Eau (DCE - 2000/60/CE) qui pour définir un bon état écologique doit mettre en place un programme de surveillance. Pour les paramètres environnementaux qui nous intéressent, ce programme définit un découpage en deux périodes : la période productive et la période non productive noté par la suite e_i : $E = \{e_1, e_2\}$. Par conséquent, les données de mars à octobre sont labellisées e_1, correspondant à la période productive (en terme de capacité de production de biomasse du phytoplancton), et les autres e_2 pour la période non productive. Cette labellisation permettra de comparer les performances de notre système hybridé (développé au chapitre 3) avec d'autres algorithmes d'apprentissage automatique (classification hiérarchique, Expectation-Maximization).

Deux critères sont considérés pour évaluer les résultats de classifications :

- Le taux de reconnaissance (RR pour Recognition Rate) estimé à partir de la matrice de confusion, somme des éléments classés dans les mêmes groupes sur le nombre total d'éléments ;

$$RR = \frac{quand(s_i = E_i)}{card(E_i)} \qquad (4.1)$$

- Le chevauchement mensuel (Overlap) entre les états s_1 et s_2 définis comme suit :

$$Overlap = \frac{\sum_i [|s_1(i)| + |s_2(i)| - max(|s_1(i)|, |s_2(i)|)]}{|s_1 \cup s_2|} \qquad (4.2)$$

$|.|$ est l'opérateur cardinal et $|s_j(i)|$ définit le nombre de points labélisé s_j durant le i^{ème} mois, $j = 1$ ou 2. Les périodes productive et non productive du phytoplancton sont censées n'avoir aucun recouvrement.

La base de données 2005 à 2008 a été découpée en deux sets, l'un pour la construction du modèle et sa validation, le second pour tester le modèle. Les données de 2005 à 2008 sont utilisées pour construire les paramètres du MMC-NS (apprentissage) et les données de l'année 2009 sont utilisées pour tester la modélisation temporelle.

4.3.1. Validation de la génération des symboles

Le nombre de symboles $\boldsymbol{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M\}$ requis pour caractériser un état est tout d'abord analysé, ainsi que la méthode de sélections des symboles maximisant le taux de reconnaissance RR et minimisant le chevauchement Overlap.
Deux approches sont testées pour construire les M symboles par état :

- La sélection aléatoire de M symboles par état à partir des données est testée 100 fois ;

- La sélection par identification de structure via l'algorithme du K-means en fixant $K = M$.

L'algorithme du plus proche voisin (1-ppv) est utilisé pour estimer l'état de chacune des données à partir de la base étiquetée des M symboles. Pour montrer la pertinence de notre approche multi-paramètre, le 1-ppv est tout d'abord appliqué pour chaque paramètre (approche monodimensionnelle) et ensuite sur la matrice multidimensionnelle (pour 2005-2008) à partir des symboles sélectionnés par sélection aléatoire.

Pour l'analyse monodimensionnelle, sur les 100 tests réalisés, la température de l'eau (ETCO1) est le paramètre le plus discriminant, avec un taux de reconnaissance de 75,1 % (±3,5) pour un symbole par état à 77,8 % (±0,4) pour 1000 symboles par état. L'approche monodimensionnelle ne permet pas d'obtenir une partition biclasse satisfaisante.
Pour l'analyse multidimensionnelle, le tableau 4.2 résume la moyenne et l'écart-type des deux critères, RR et Overlap, pour différentes valeurs de *M*. La distribution approximative des données avec un unique représentant aléatoire donne un faible taux de reconnaissance (68,1 %) et souvent un chevauchement important (18,4 %) des deux états environnementaux désirés. Ce chevauchement est aux alentours de 10 % lorsque 100 symboles sont sélectionnés aléatoirement par état.
L'algorithme du K-means est une approche géométrique adaptée pour les jeux de données linéairement séparables. Cet algorithme nécessite de connaître le nombre de symboles (centres) M désirés. Ici, avec 10 symboles par état, le chevauchement est d'environ 10 %. Cent symboles par état sont nécessaires pour obtenir un chevauchement inférieur à 5 %.

Nous pouvons donc conclure que la sélection des symboles par l'algorithme K-means est plus performante que la sélection aléatoire. Les résultats de classification sont cependant corrects lorsqu'une sélection aléatoire de $M = 1000$ symboles par état est réalisée mais génère un chevauchement plus important que l'algorithme K-means.

Tableau 4.2. Scores de RR et Overlap du 1-ppv (moyenne et écart type en pourcentage) pour M symboles par état en accord avec la labélisation de la DCE : base de données 2005-2008

Valeur de M	Sélection aléatoire		K-means	
	RR	Overlap	RR	Overlap
1	68,1 (8,9)	18,4 (6,8)	82,7 (0,1)	11,3 (0,1)
10	79,4 (4,0)	18,5 (3,7)	89,8 (0,8)	10,0 (0,7)
100	87,6 (0,9)	12,2 (0,9)	95,1 (0,2)	4,9 (0,2)
1000	94,7 (0,2)	5,2 (0,2)	98,4 (0,1)	1,6 (0,1)

Au regard de la pertinence du K-means, l'algorithme STFKM (Self Tuning Fast K-means) proposé au chapitre 3 est utilisé pour rechercher automatiquement le nombre de symboles qui

décrit la structure des données et permettre d'optimiser les temps de calcul si nécessaire. L'impact de la sélection par STFKM est mesuré avec un classifieur supervisé, un séparateur à vaste marge (SVM pour Support Vector Machine). Le classifieur SVM a pour but de classer une donnée vis-à-vis de points supports qui caractérisent l'hyperplan séparateur des classes avec une large marge associée (Cortes et Vapnik, 1995). Ce classifieur a été choisi puisqu'il permet d'obtenir une solution optimale globale. Un modèle SVM (noyau à base radiale, coût de contrainte à 1 et $\gamma = 0{,}1$) a été entraîné sur les données 2005-2008 et a été testé sur les données 2009 avec dix validations croisées. Trois expériences pour l'apprentissage du SVM ont été menées :

- Sur l'ensemble des données ;
- Sur 1000 symboles sélectionnés aléatoirement par état ($N = 2$ et $M = 2000$) ;
- Sur les symboles issus de la quantification vectorielle par STFKM.

La génération aléatoire de 1000 symboles par état ayant donné un taux de reconnaissance fort, celle-ci fut par conséquent confrontée aux sorties de l'algorithme STFKM.
Les capacités d'apprentissage (reconnaître une donnée déjà apprise pour construire le SVM) et de généralisation (test des données non utilisées pour construire le modèle) du SVM pour ces trois études avec les critères RR et Overlap sont résumés dans le tableau 4.3. En généralisation, l'algorithme STFKM permet de garder un taux de reconnaissance (92,6 %) et un chevauchement (7,4 %) similaire au SVM sans échantillonnage. Le STFKM-SVM donne une meilleure capacité d'apprentissage que la sélection aléatoire des M symboles.

Dans ce contexte supervisé, nous pouvons conclure que la quantification vectorielle obtenue par STFKM permet une réduction pertinente des données.

Tableau 4.3. Capacité d'apprentissage et de généralisation (Test) du SVM pour les 3 expériences (sans échantillonnage, par échantillonnage aléatoire et par quantification vectorielle (algorithme Self Tuning Fast K-means)) mesurée à partir du taux de reconnaissance RR et du chevauchement Overlap (%).

Expérience	Apprentissage		Test	
	RR	Overlap	RR	Overlap
Pas d'échantillonnage – SVM	97,4	2,1	92,9	7,0
1 000 aléatoires – SVM	93,3	6,7	92,2	6,7
STFKM – SVM	95,4	4,6	92,6	7,4

L'algorithme Self Tuning Fast K-means possède une initialisation aléatoire des centres de gravité. De ce fait, la stabilité de cet algorithme est évaluée à partir de l'indice de Rand (RI acronyme de Rand Index (Milligan et Cooper, 1986)) des 10 générations de symboles obtenues. L'indice de Rand est une mesure de similarité entre deux classifications de données P_1 et P_2 d'un ensemble donné de n éléments $P = \{P_1, \dots, P_n\}$. Il faut noter que le nombre de groupes dans chaque partition peut être différent. RI est calculé comme suit :

$$RI(P_1, P_2) = \frac{a+b}{\binom{n}{2}} \qquad (4.3)$$

où a (respectivement, b) est défini par le nombre de paires d'éléments dans E qui sont du même ensemble dans P_1 (respectivement, dans des ensembles différents) et dans le même ensemble dans P_2 (respectivement dans des ensembles différents). $RI(P_1, P_2) = 1$ signifie que les partitions sont identiques ; à noter, que cet indice permet de comparer des partitions avec un nombre de groupes différents. Toutefois, il diminue lorsque cet écart augmente.

Dans un contexte non supervisé, l'approche STFKM permet de respecter l'information haute fréquence sans perdre la structure des données. Les RI des 10 partitions obtenues et le nombre de symboles retenu sont quasi-similaires (tableau 4.4). Les RI, proche de 1, démontre que l'algorithme STFKM donne une quantification vectorielle robuste.

Tableau 4.4. Indicateurs de tendance centrale et de dispersion de l'Indice de Rand (RI) et de la valeur de M (nombre de symboles) pour 10 générations de symboles, avec Q1 le premier quantile et Q3 le troisième quantile.

Statistiques	Minimum	Q1	Médiane	Moyenne	Q3	Maximum
R1	0,99	0,99	0,99	0,99	0,99	0,99
M	2744	2749	2754	2759	2763	2790

4.3.2. Validation de la génération des états

La génération des états est réalisée par Classification Spectrale (SC). Celle-ci est comparée avec les approches non supervisées usuelles : Expectation-Maximization (EM), Classification Hiérarchique (HC) pour un nombre d'états $N = 2$ avec les symboles issus de l'algorithme STFKM.

L'algorithme EM est constitué de deux étapes (Moon, 1996) :

- L'étape E consiste à calculer le critère de vraisemblance des paramètres selon la forme et le volume des groupes recherchés.
- L'étape M consiste à redéfinir les paramètres du modèle.

La classification hiérarchique (Borcard *et al.*, 2011) regroupe les données en recherchant la plus petite distance (ou la plus forte similarité) pour chaque paire de points. Il existe deux types de classification hiérarchique :

- Ascendante : au départ, chaque donnée correspond à un groupe. Une donnée fusionne avec une autre, puis une troisième donnée fusionne avec une quatrième donnée ou avec le groupe obtenu de la fusion précédente, et ainsi de suite jusqu'à ce qu'il ne reste qu'un seul groupe constitué de l'ensemble des données.
- Descendant : au départ, il n'y a qu'un seul groupe qui est ensuite divisé en deux, puis chacun d'entre eux sera divisé en deux et ainsi de suite jusqu'à ce que chaque groupe ne soit constitué que d'une seule donnée.

Pour ces comparaisons, par convention, un voisinage à 7 points est considéré pour l'initialisation de la matrice de similarité de la classification spectrale. Nous choisissons d'utiliser les algorithmes permettant d'obtenir les meilleurs résultats, c'est-à-dire l'algorithme EM avec des groupes de forme sphérique et de volume égal, et la classification hiérarchique ascendante (distance euclidienne et méthode de Ward). EM est aussi performant sur chacun des 10 paramètres, seuls les résultats du paramètre le plus discriminant (salinité CSAL1) sont présentés. L'algorithme 1-ppv est utilisé pour labelliser les données à partir du modèle construit.

Pour rappel, la base de données 2005-2009 est découpée comme suit :

- Base de données d'apprentissage : 2005-2008
- Base de données test : 2009

Le tableau 4.5 présente les scores de RR et d'Overlap pour l'analyse conjointe. Malgré son très faible chevauchement aux alentours de 0,5 %, le résultat obtenu par classification hiérarchique (coupe obtenue à partir d'un arbre à 2 groupes) ne représente pas une séparation des périodes productive et non productive. L'état 1 représente 84 118 des 84 614 points, soit 99 % de la base construite, et l'état 2 ne concerne que quelques points en août, septembre et novembre pour la base construite (février, avril et juin pour la base de test) sans aucune interprétation biologique (évènement rare) ou capteur (saut ou dérive). L'approche EM offre le meilleur résultat RR sur les deux bases de données. L'approche STFKM donne un RR inférieur par rapport à EM seul, mais son Overlap pour les grandes bases de données (base de données construite) est réduit. L'approche STFKM-SC renvoie des résultats considérés comme un bon compromis pour la labellisation de la DCE au sens du taux de reconnaissance et de chevauchement, et sera pertinent pour un nombre d'états supérieur à deux. En effet, nous nous attendons à ce que notre système MMC hybride soit capable de détecter plus de deux états.

Tableau 4.5. Taux de reconnaissance (RR) et de chevauchement (Overlap) en pourcentage pour chaque classification non supervisée réalisée sur la base de données construite et la base de données test.

Méthode	Base de données d'apprentissage		Base de données test	
	RR	Overlap	RR	Overlap
S-EM (CSAL1)	70,1	14,7	77,0	11,9
STFKM-EM (EII)	83,6	13,4	91,2	4,3
STFKM-HC	66,9	0,6	66,9	0,5
STFKM-SC	79,1	11,9	84,1	5,7

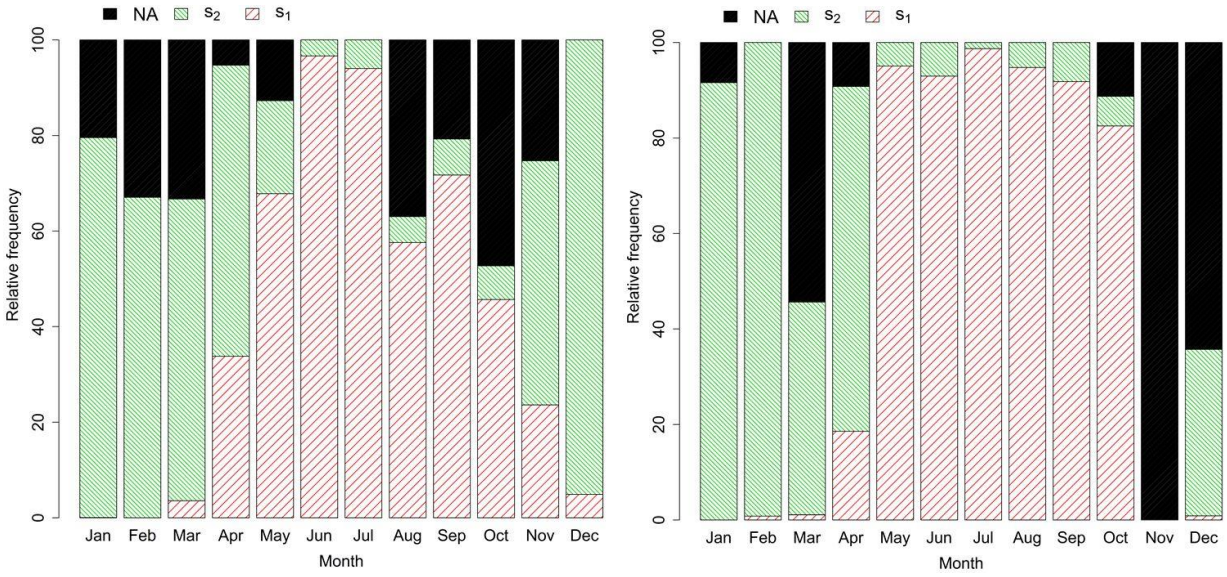
4.3.3. Validation de la modélisation temporelle

A partir de tests, la fiabilité de notre modèle hybridé est évaluée : la procédure entière pour la construction d'un MMC-NS à partir d'une classification non supervisée est répétée 10 fois. Les 10 partitions ont un indice de Rand moyen de 0,95. Nous pouvons donc conclure que l'ensemble de l'étape STFKM-SC (génération des symboles et des états) est robuste. Les partitions de symboles et d'états minimisant notre critère de coupe $Ncut$ défini au chapitre 3 sont conservées pour construire notre MMC-NS. Pour ces tests, le nombre d'états est fixé à $N = 2$ et les autres paramètres sont réglés automatiquement : la variance expliquée est fixée à 95 %. En accord avec le critère de coupe, le MMC construit possède $M = 2\,794$ symboles.

$$\lambda(N = 2, M = 2\,794, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}) \qquad (4.4)$$

En accord avec la labellisation de la DCE, 79,3 % de la base de données construite est bien reconnue avec un Overlap de 11,7 %. 82,1 % de la base de données test 2009 est bien reconnu avec un Overlap de 6,7 %. Les données du mois de novembre de la base de données test étant manquantes, cela entraîne une augmentation du taux de reconnaissance et une diminution du chevauchement, ce qui explique que les résultats soient plus performants pour la base de données test que pour la base de données construite. La figure 4.6 illustre la distribution des états labellisés par la prédiction du MMC-NS pour la base de données construite et la base de données test 2009. En 2005-2008, l'état s_1 en couleur rouge est lié avec la période débutant en mars et se terminant en décembre avec une dominance sur la période avril-novembre, tandis que l'état s_2 de couleur verte est dominant sur la période de novembre à avril. Au cours de l'année 2009, l'état s_1 est lié avec la période avril à octobre, tandis que l'état s_2 est dominant de décembre à avril. Plusieurs données en mars puis entre août et novembre n'ont pas d'état estimé (noté NA en couleur noir sur la figure 4.6) dû à au moins une valeur manquante dans $\mathbb{R}^{10}$, cela engendre des confusions du système dans les prédictions. En effet, les données sont présentées au système de manière concaténée quels que soient les trous dans la séquence, et donc peuvent fausser le calcul basé sur des probabilités de transitions.
Les états s_1 et s_2 correspondent bien aux deux principaux états environnementaux sur lesquels se base la DCE pour définir la stratégie d'échantillonnage pour le phytoplancton (périodes productive et non productive).

Pour valider le modèle construit sur la période 2005-2008 et la quantification des symboles dans un cadre totalement non supervisé, les signaux originaux de la station MAREL-Carnot pour l'année 2009 : $\boldsymbol{X}(T)$, sont comparés à des signaux reconstruits à partir du MMC-NS : $\widehat{\boldsymbol{X}}(T)$. Pour une observation $x(t)$, le système estime son état s_i ayant la plus grande vraisemblance. Puis les symboles $\boldsymbol{v}_k$ les plus présents dans cet état sont retenus.
La similarité Sim, définie au chapitre 2, est comprise entre 0 et 1 : la valeur 1 signifie que les signaux reconstruits sont exactement les mêmes que les originaux.



(a) (b)

Figure 4.6. Distribution des états $\boldsymbol{s_1}$ *(en rouge) et* $\boldsymbol{s_2}$ *(en vert) par mois sur (a) la période 2005-2008, base de données à partir de laquelle le MMC-NS est construit et (b) l'année 2009, base de données test. La couleur noire, nommée NA, concerne les mesures dont l'état n'a pu être estimé (au moins une donnée manquante par instant).*

Les données des années 2005 à 2008 participent à la construction du MMC-NS, et les signaux reconstruits sont comparés aux originaux de manière à pouvoir apprécier la puissance du modèle. Les similarités entre les signaux originaux et reconstruits pour chaque année sur la période 2005-2008 et la moyenne associée à la période entière sont résumées dans le tableau 4.6. La dernière colonne de ce tableau correspond à l'année 2009. Les similarités et la moyenne de celles-ci sont supérieures à 0,83 pour 2005-2008. Le signal reconstruit étant constitué de symboles et la prédiction pouvant être faussée par les données manquantes, les similarités ne sont pas égales à 1. Ainsi, nous concluons que la reconstruction des signaux est très proche des données d'origine et que l'algorithme de quantification vectorielle pour les états du MMC-NS est efficace. Par conséquent, le système proposé possède une forte puissance de généralisation, pour l'année 2009 ($Sim = 0{,}79$), qui n'a pas participé à la construction du MMC. Cette étape de comparaison du signal original de la concentration en oxygène dissous corrigé (C_O21) en fonction du signal reconstruit de ce paramètre pour l'année 2009 à partir du modèle développé sur la base de données 2005-2008 est illustrée sur la figure 4.7. Le coefficient de détermination entre ces deux signaux est de 0,795***.

Tableau 4.6. Scores de similarité pour la reconstruction des signaux basée sur la classification des états à partir du MMC-NS.

Année	2005	2006	2007	2008	Moyenne	2009
Sim	0,87	0,84	0,83	0,86	0,85	0,79

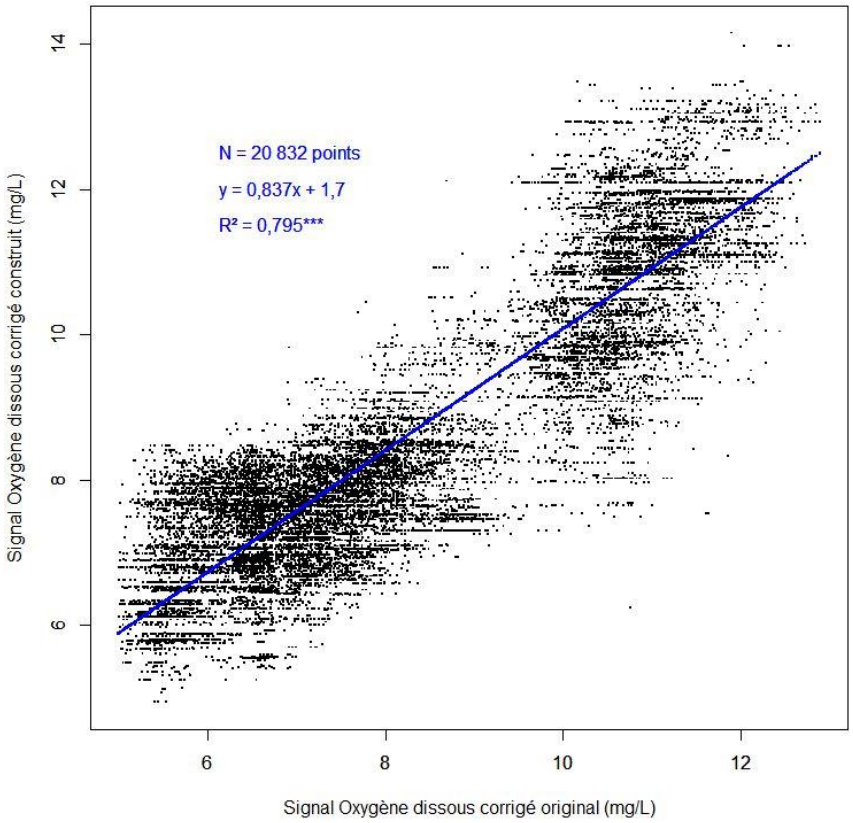


Figure 4.7. Relation entre le signal original de la concentration en oxygène dissous corrigé (C_O21) et le signal reconstruit de ce paramètre pour l'année 2009 à partir du modèle développé sur la base de données 2005-2008. La droite de régression linéaire entre ces deux paramètres est représentée en bleu.

Nous considérons que le modèle de Markov caché est maintenant validé par une dynamique biologique à 2-états fixés, la prochaine étape est d'augmenter le nombre d'états pour subdiviser la segmentation et essayer d'avoir une meilleure compréhension du déterminisme et de la dynamique d'une efflorescence.

4.4. Généralisation du Modèle de Markov Caché à N-états

La construction du modèle MMC-NS a défini automatiquement un nombre d'états N égal à 7 et un nombre de symboles $M = 2\,884$ sur la base de données 2005-2008 :

$$\lambda(N = 7,\ M = 2\,884,\ \boldsymbol{\pi},\ \boldsymbol{A},\ \boldsymbol{B}) \qquad (4.5)$$

Cette base est constituée de $D_p = 10$ paramètres non corrélé (NC) et $N_p = 84\,614$ instants. La volonté d'augmenter le nombre d'état est liée au besoin de déterminer la dynamique du phytoplancton de manière plus précise, en exploitant au maximum l'information contenue dans la base de données haute fréquence. Il s'agit, par exemple, de pouvoir identifier au minimum, la période d'efflorescence, de pré-efflorescence, de post-efflorescence, mais également de détecter des événements rares tels que l'ouverture de barrage qui permet l'arrivée d'eau douce dans la rade de Boulogne-sur-Mer ou des problèmes de capteurs.

4.4.1. Classification et interprétation d'un modèle à 7-états

La validation de la modélisation temporelle est effectuée à partir du même protocole que pour le MMC à 2-états. La reconstruction des signaux de 2009 est telle que l'on calcule une similarité proche de 0,8 avec les données originales.

Les états estimés sur la période 2005-2008 sont projetés sur le signal de fluorescence qui n'a pas été utilisé pour la construction du MMC-NS afin de mettre en évidence la pertinence du séquencement des états lorsque l'on considère la dynamique de développement de la biomasse phytoplanctonique (figure 4.8) (la couleur noire représente les données non labellisées à cause d'au moins un paramètre possédant une donnée manquante à cet instant). Le système MMC-NS permet de caractériser la distribution et le séquencement des états pour une année type (période 2005-2008) afin d'aborder les variations saisonnières des états, ou pour la période considérée dans son ensemble pour mieux comprendre la variabilité inter-annuelle (figure 4.9).

Il apparaît ainsi des états caractéristiques de certaines périodes de l'année :

- s_1 de janvier à mai
- s_2 d'avril à octobre
- s_3 de novembre à mars
- s_4 de mai à novembre

Les 3 autres états n'apparaissent pas à la même période d'une année sur l'autre.

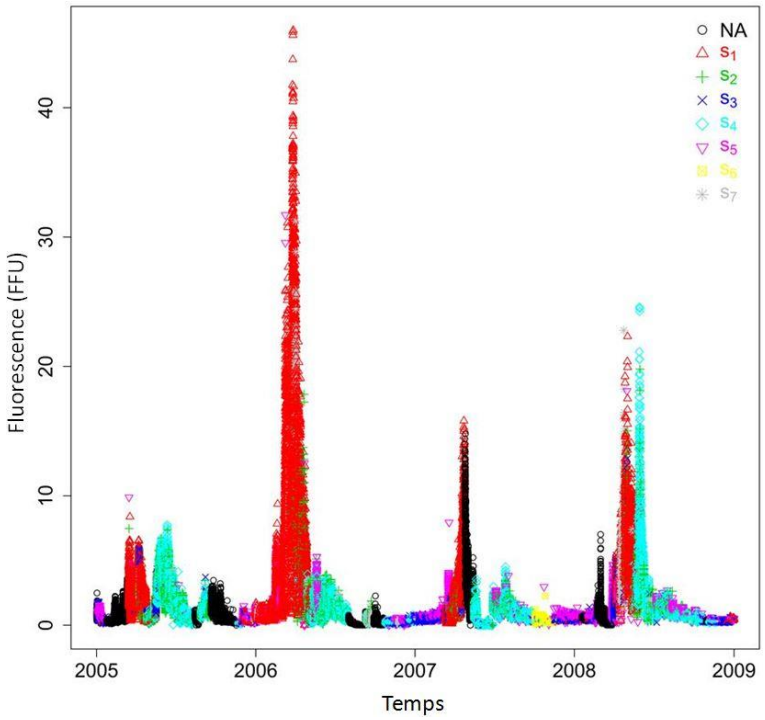


Figure 4.8. Résultats de la classification spectrale : projection des états ($\boldsymbol{s_i}$) sur le signal de fluorescence (FFU) mesuré par la station MAREL-Carnot sur la période 2005-2008. Le noir concerne les mesures dont les états n'ont pas été estimés (au moins une donnée manquante NA pour un paramètre à cet instant).

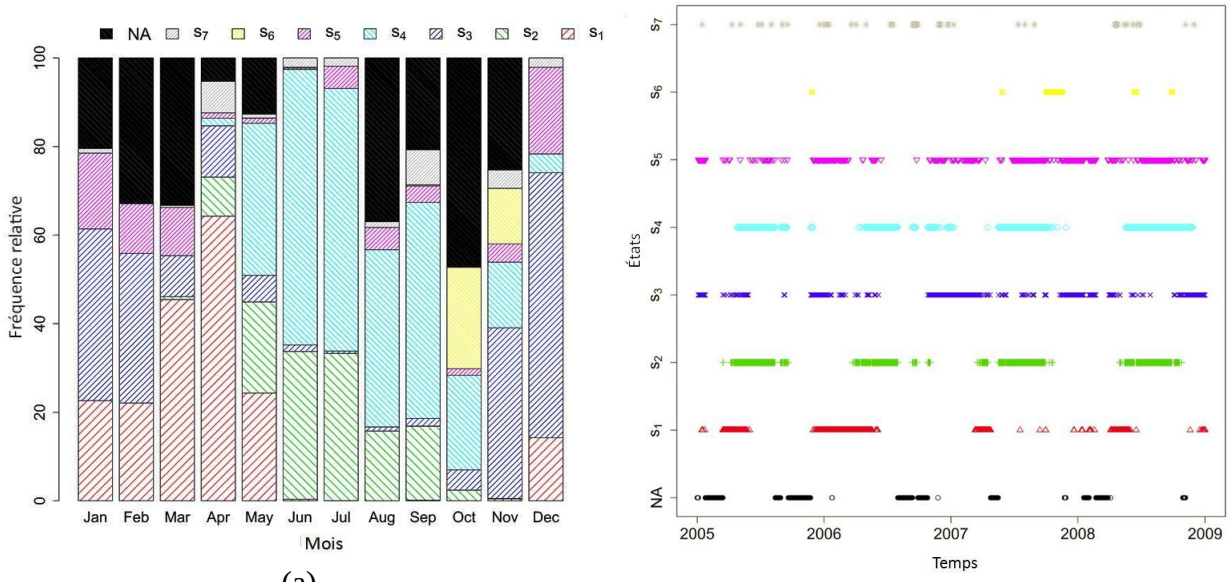


Figure 4.9. Résultats de classification spectrale : l'état NA concerne les mesures dont les états ne sont pas estimés (au moins une donnée manquante à un instant t), (a) distribution des états par mois pour un cycle saisonnier typique sur la période 2005-2008 et (b) séquencement des états sur la période 2005-2008.

Pour interpréter les résultats du modèle et les relier à notre problématique écologique, une analyse de corrélation (méthode de Pearson) entre les paramètres et les états obtenus est réalisée (tableau 4.7). Les corrélations les plus proches de 1 ou -1 permettent de connaître les paramètres structurants chacun de ces états. Par exemple, l'état s_1 est structuré par l'oxygène dissous et la température (coefficients de corrélation de 0,64 et de -0,50, respectivement).

Tableau 4.7. Coefficients de corrélations entre les paramètres et les états déterminés à partir d'une classification non supervisée sur les paramètres non corrélés (NC) issus de la station MAREL-Carnot sur la période 2005-2008 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 par état : seuils arbitraires).

État	s_1	s_2	s_3	s_4	s_5	s_6	s_7
Couleur	rouge	vert	bleu	cyan	rose	jaune	gris
C_NI1	-0,14	-0,24	**0,54**	-0,24	0,12	0,02	-0,03
C_O21	**0,64**	-0,16	-0,03	**-0,38**	0,05	-0,17	0,01
C_PO1	-0,08	-0,03	-0,04	-0,07	-0,02	-0,03	**0,57**
C_SI1	-0,11	-0,16	0,20	-0,25	0,15	0,04	**0,47**
CSAL1	0,12	0,13	**-0,36**	0,10	-0,25	**0,42**	0,00
E_LU1	-0,08	**0,73**	-0,21	-0,24	-0,06	-0,06	0,00
E_TU1	-0,11	-0,14	-0,01	-0,22	**0,76**	-0,03	-0,05
E_VVR	-0,25	-0,03	0,16	-0,05	**0,31**	-0,06	-0,03
ETCO1	**-0,50**	**0,32**	**-0,37**	**0,56**	-0,14	0,07	0,05
XMAHH	0,00	0,03	0,02	-0,03	0,01	0,01	0,00

L'initiation de l'efflorescence phytoplanctonique principale et l'étape de croissance phytoplanctonique sont caractérisées par l'état s_1 (rouge). Celui-ci est principalement présent entre février et mai avec une température faible (8,11 °C en moyenne ($\pm$ 2,52 °C) ; figure 4.11 (a)). Les valeurs de la fluorescence dans cet état ont une moyenne de 4,19 FFU $\pm$ 5,52 FFU), le troisième quantile est de 5,07 FFU et le maximum des valeurs est à 45,99 FFU (niveau de fluorescence maximum durant l'efflorescence printanière de l'année 2006) (figure 4.10).

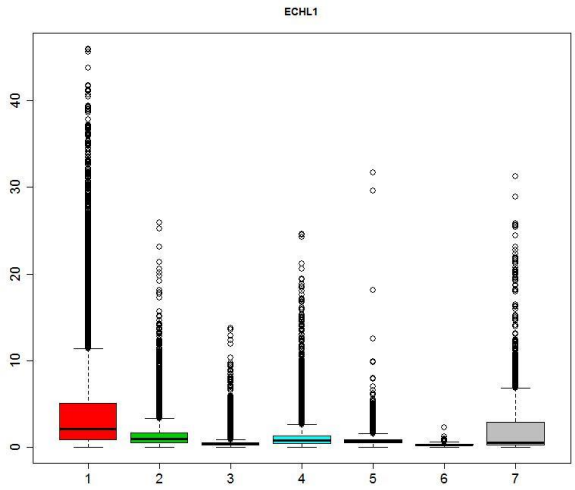
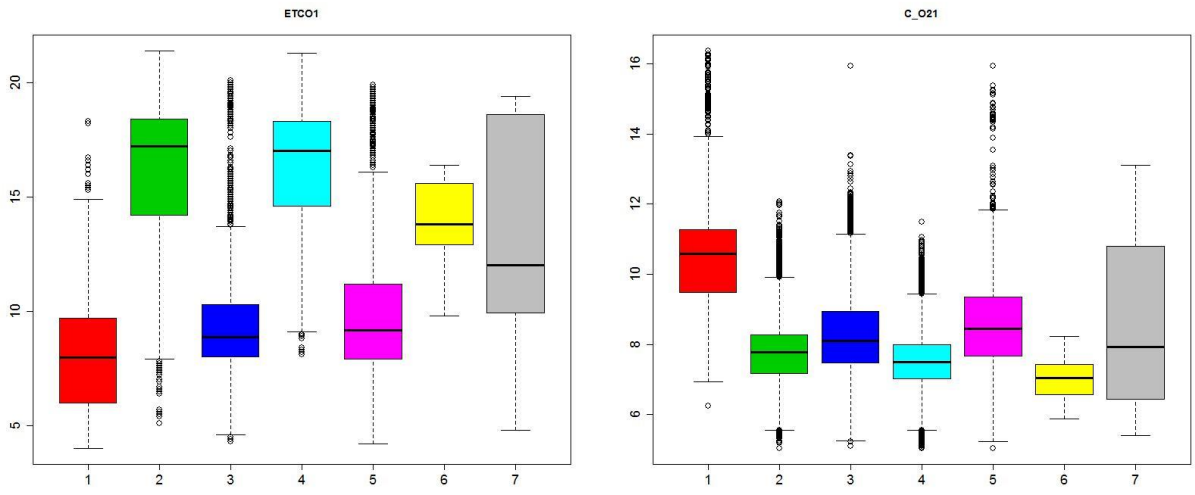


Figure 4.10. Boîtes de dispersion de la fluorescence (FFU) mesurée par la station MAREL-Carnot sur la période 2005-2008 pour chacun des 7 états obtenus après classification spectrale de la base de données NC (paramètres non corrélés entre eux).

De fortes concentrations en oxygène (valeur moyenne de l'oxygène dans cet état s_1 : 10,40 mg/L ($\pm$ 1,32 mg/L)), expliquées par une forte production phytoplanctonique (photosynthèse), sont observées dans cet état (tableau 4.7 et figure 4.11 (b)). Durant l'état s_1, le phytoplancton utilise principalement le stock hivernal de nutriments, et par conséquent, cet état correspond à la période de production nouvelle (par opposition à la période de production régénérée).

Les états s_2 (vert) et s_4 (cyan) suivent l'état s_1, et sont identifiés comme la période de production régénérée quand la production phytoplanctonique est basée sur la transformation de la matière particulaire constituée à partir de l'efflorescence précédente – état s_1 – en des formes de nutriments disponibles de nouveau (figures 4.8 et 4.9), c'est-à-dire de mai à octobre, d'où des températures plus élevées (figure 4.11 (a)).



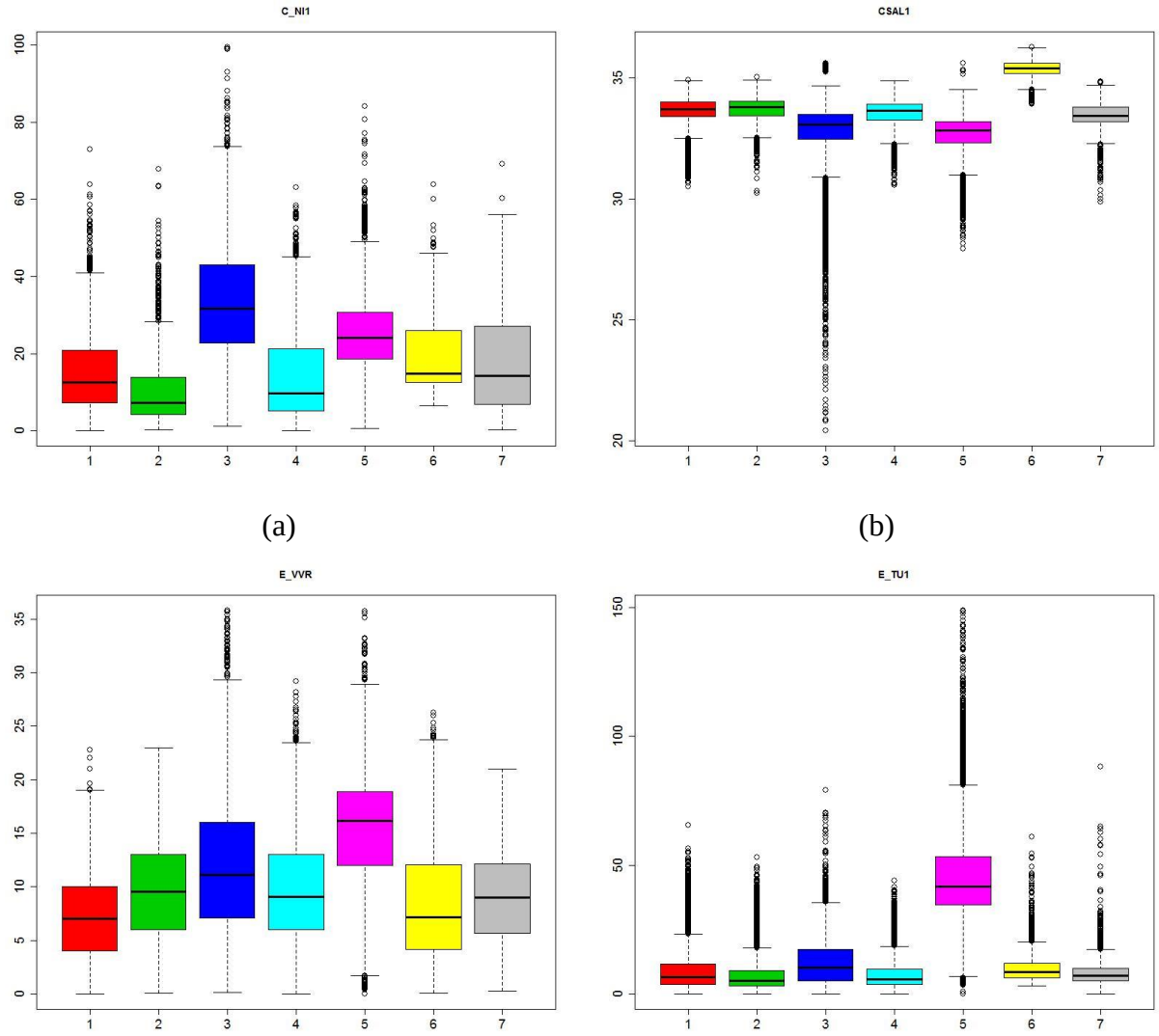
(a) (b)

Figure 4.11. Boîtes de dispersion de (a) la température de l'eau (°C) et (b) la concentration en oxygène dissous corrigé (mg.L^{-1}) mesurée par la station MAREL-Carnot sur la période 2005-2008 pour chacun des 7 états obtenus après classification spectrale de la base de données NC (paramètres non corrélés entre eux).

L'état s_3 correspond à la période non productive (principalement présent de novembre à février (figure 4.9)) et est principalement structuré par une concentration élevée en nitrate ainsi qu'une salinité et une température de l'eau plus faible (tableau 4.7). La concentration en nutriments (notamment en nitrate (figure 4.12 (a))) est élevée puisqu'ils ne sont pas consommés durant cette période de l'année (lumière insuffisante, turbidité trop importante). Nous notons également des dessalures importantes liées à des apports d'eau douce en raison de précipitations accrues en cette période (salinité minimum égale à 20,41) (figure 4.12 (b)). Les vents dont la vitesse moyenne est de 11,60 m.s^{-1} ($\pm$ 5,74 m.s^{-1}) (42 km.h^{-1}), peuvent atteindre une vitesse de 36,11 m.s^{-1} (130 km.h^{-1}) (figure 4.12 (c)), ce qui va contribuer à augmenter la turbidité par brassage des masses d'eaux : turbidité de 11,67 NTU en moyenne ($\pm$ 9,13 NTU), pouvant atteindre une valeur maximale de 79,10 NTU (figure 4.12 (d)).

Les états s_5 (rose) et s_7 (gris) correspondent à des évènements rares, extrêmes ou de courtes durées. Il peut s'agir, par exemple, de périodes avec une forte turbidité durant des tempêtes (figure 4.12 (c) et (d)) et de périodes avec de fortes concentrations de phosphate et de silicate par effet de remise en suspension ou apports via le bassin versant, respectivement (figure 4.13 et tableau 4.7). L'état s_5 apparait de manière quasi-continue chaque année durant les mois de janvier, février et décembre. Le reste de l'année, il apparait rarement excepté de juillet à septembre où il apparait sur 5 – 6 instants (non séquentiel) par jours. L'état s_7 apparait par séquence d'une demi-journée mais de manière éparse durant l'année. Par exemple, en 2005, il apparait une demi-journée tous les mois d'avril à septembre mais l'écart entre ces apparitions ne sont jamais les mêmes alors qu'en 2007 il apparait une demi-journée en janvier, puis en juillet et août. Des recherches supplémentaires sont nécessaires pour mieux comprendre les

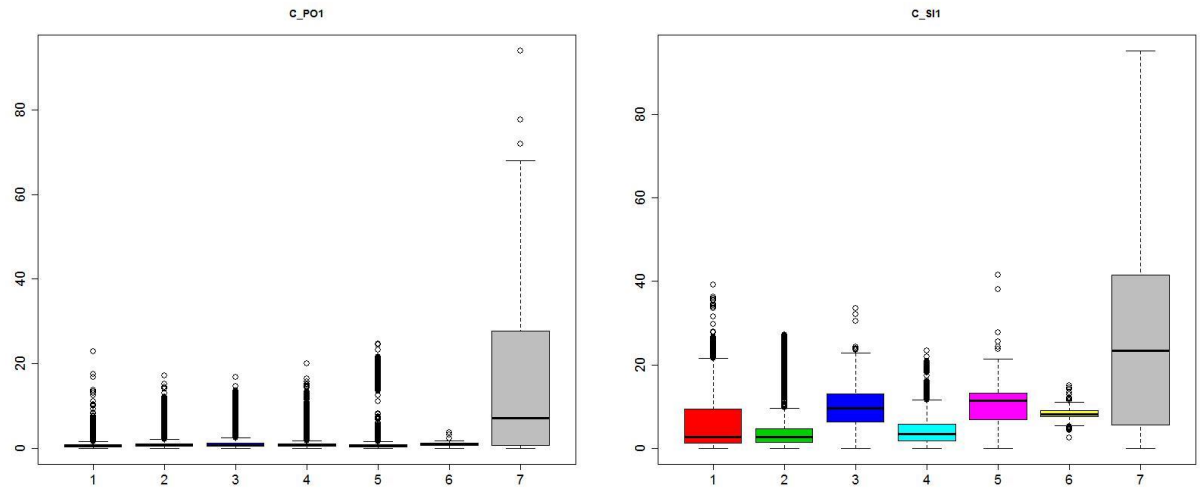
processus impliqués durant ces états. En effet, ce type d'état ne pourra être défini plus précisément qu'en intégrant de nouvelles sources de données comme, par exemple, les périodes de dragage du port, les opérations d'ouverture et de fermeture du barrage Marguet dans des situations diverses (gestion de base des flux d'eau à la mer, gestion de crue,…), la direction et la vitesse du vent et des courants,…



(a) (b)

(c) (d)

Figure 4.12. Boîtes de dispersion de (a) la concentration en nitrate (µmol.L^{-1}) (b) la salinité, (c) la vitesse du vent en rafale (m.s^{-1}) et (d) la turbidité (NTU) mesurées par la station MAREL-Carnot sur la période 2005-2008 pour chacun des 7 états obtenus après la classification spectrale de la base de données NC (paramètres non corrélés entre eux).



(a) (b)

Figure 4.13. Boîtes de dispersion des concentrations (a) en phosphate (µmol.L^{-1}) et (b) en silicate (µmol.L^{-1}) mesurées par la station MAREL-Carnot sur la période 2005-2008 pour chacun des 7 états obtenus après la classification spectrale de la base de données NC (paramètres non corrélés entre eux).

L'état s_6 (jaune) est caractérisé par une forte salinité (tableau 4.7) due à une défaillance du capteur de conductivité en 2007 visible sur les données de salinité (figures 4.12 (b) et 4.14). Les valeurs de la salinité sont comprises entre 33,91 et 36,28 dans cet état. Contrairement aux autres états dont la moyenne des valeurs est de l'ordre de 32-33, la moyenne de l'état s_6 est 35,36 ($\pm$ 0,37). De plus, lorsque les données de salinité à cette période sont comparées avec celles mesurées sur le point le plus côtier de la radiale de Boulogne-sur-Mer du réseau REPHY / SRN (Lefebvre et Mégret, 2014) située à 2 milles nautiques de la station MAREL-Carnot (chapitre 1), un problème évident de défaillance du capteur apparait. La salinité au point côtier de la radiale SRN est comprise entre 34 et 34,5 (Figure 9). Ce point côtier de la radiale SRN est considéré comme un prélèvement au large au regard de la position de la station MAREL-Carnot et doit par conséquent présenter une salinité plus importante si l'on considère le gradient croissant théorique pour ce paramètre dans un système côtier sous influence d'apports d'eaux douces (via la Liane). Les prélèvements du REPHY / SRN se faisant à Pleine Mer ± 2 heures, l'influence marine est encore plus forte. Les données de la station MAREL-Carnot ne peuvent ainsi en aucun cas être supérieures à celles de la station côtière du SRN (figure 4.14). L'état s_6 n'a donc pas un sens biologique mais il illustre la capacité du MMC-NS à mettre en évidence des événements inhabituels.

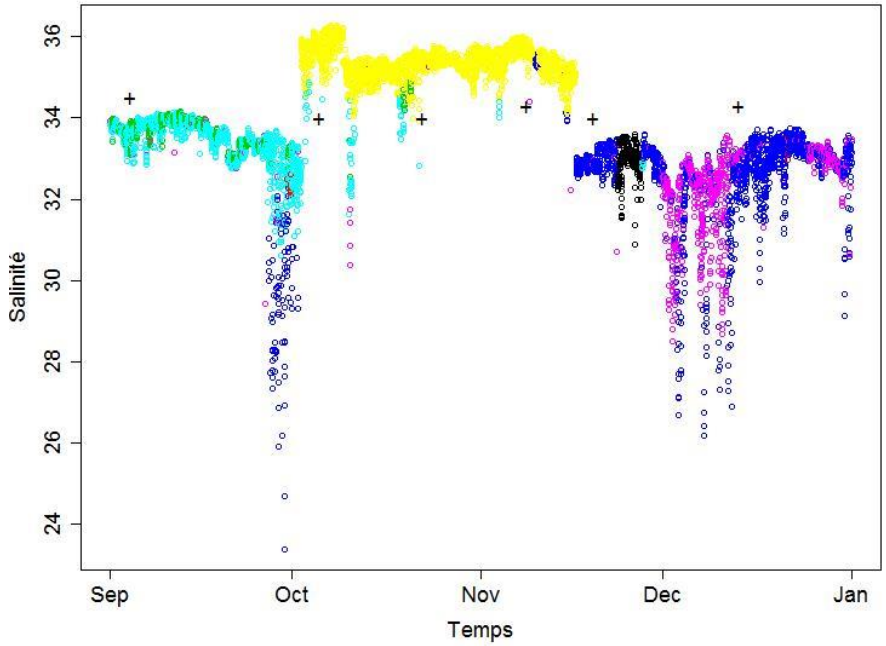


Figure 4.14. Résultats de la classification spectrale : projection des états sur le signal de salinité mesuré par la station MAREL-Carnot au cours de l'année 2007. Le saut de capteur en octobre et novembre a été classé dans un état $\boldsymbol{s_6}$ *(jaune). Les ronds noirs concerne les mesures dont les états n'ont pas été estimés (au moins une donnée manquante NA pour un paramètre à cet instant). Les données mesurées sur le point côtier de la radiale de Boulogne-sur-Mer du réseau REPHY / SRN sont représentées par les croix noires.*

4.4.2. Estimation des états de nouvelles données entrantes

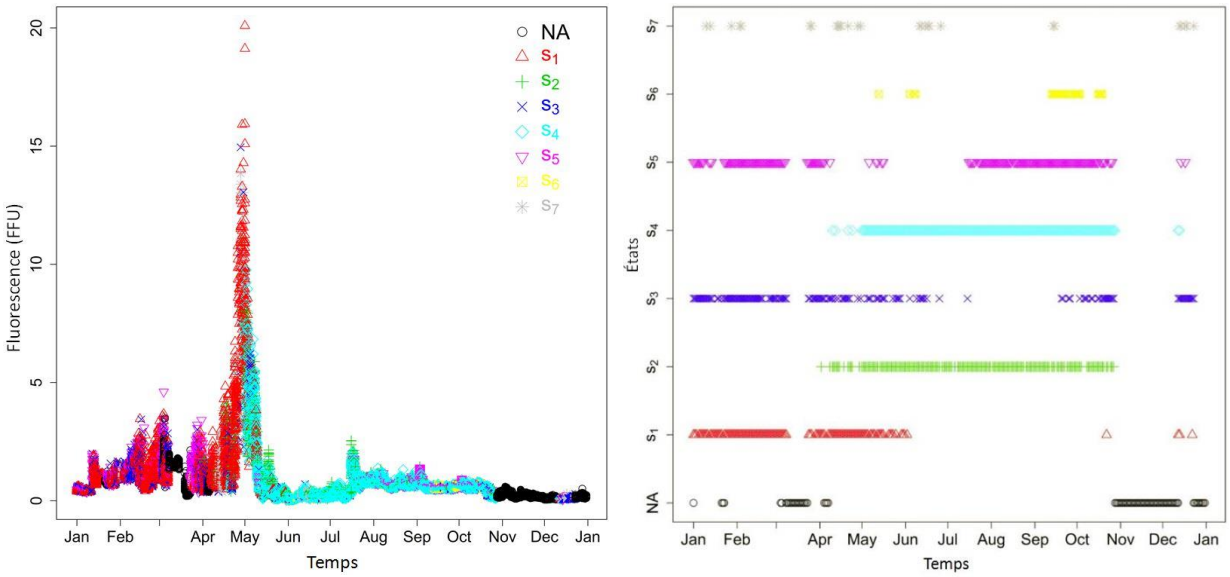
Les paramètres MMC-NS à 7-états sont calculés avec $N = 7$ états et $M = 2884$ symboles, et il est ainsi possible d'estimer les états de nouvelles données entrantes. Ces nouvelles données sont les mesures effectuées par la station MAREL-Carnot au cours de l'année 2009. A terme, ces nouvelles données seront acquises en temps réel et il s'agira alors de définir le pas de temps pour leur intégration dans le système de décision MMC-NS. L'algorithme de Viterbi, décrit dans le chapitre 3, permet d'estimer les états de ces données. Le chemin optimal entre les états pour les données de l'année 2009 est représenté sous la forme d'un séquencement (figure 4.14 (b)) permettant ainsi d'identifier des états caractéristiques de périodes clefs :

- s_1 : de janvier à mai ;
- s_2 : d'avril à octobre ;
- s_3 : de janvier à avril ;
- s_4 : d'avril à octobre.

L'état s_5 est présent de janvier à avril, mais pas en séquence complète (quelques données par jours) et l'état s_7 apparait par période d'une demi-journée pour les mois de janvier à avril ainsi qu'en juin, septembre et décembre.

La projection des états estimés sur la fluorescence (figure 4.15 (a)) et le séquencement d'états correspondent à notre hypothèse minimale de segmentation de la dynamique des efflorescences, avec une période de pré-bloom hivernale (état s_3 principalement) suivie de l'efflorescence phytoplanctonique principale basée sur une entrée externe et nouvelle de nutriments (état s_1), puis à un bloom basé sur une production régénérée (état s_2 et s_4).

Nous avons vu que l'état s_6 défini sur la base de données NC 2005-2008 correspond à une défaillance du capteur de conductivité impactant les données de salinité. Cependant, durant l'année 2009, ce genre de saut n'existe pas. L'estimation, en utilisant les matrices de transitions et d'émissions, attribue cet état aux données possédant une plus forte salinité (de l'ordre de 34,23 $\pm$ 0,12).



(a) (b)

Figure 4.15. Résultats de l'estimation des états par le modèle MMC-NS : (a) projection des états ($\boldsymbol{s_i}$) sur le signal de fluorescence (FFU) mesuré par la station MAREL-Carnot et (b) le séquencement des états sur l'année 2009. Le noir concerne les mesures dont les états n'ont pas été estimés (au moins une donnée manquante NA pour un paramètre à cet instant).

Ces résultats sont confirmés par les corrélations entre les états et les paramètres (tableau 4.8). Les états sont structurés de la même façon que sur la période 2005-2008 (tableau 4.7) sauf pour l'état s_6. Cette état estimé ne correspond plus à un saut de capteur, d'où la structuration de cet état par des paramètres différents.

Tableau 4.8. Coefficient de corrélations entre les états estimés à partir de l'algorithme de Viterbi et les paramètres mesurés par la station MAREL-Carnot sur l'année 2009 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 par état : seuils arbitraires).

État	s_1	s_2	s_3	s_4	s_5	s_6	s_7
Couleur	rouge	vert	bleu	cyan	rose	jaune	gris
C_NI1	0,15	-0,22	**0,51**	-0,36	-0,03	-0,03	-0,02
C_O21	**0,51**	-0,28	0,28	**-0,56**	0,00	**-0,10**	0,05
C_PO1	0,01	-0,09	0,07	-0,18	-0,01	-0,03	**0,45**
C_SI1	0,03	-0,16	0,16	-0,24	**0,08**	0,06	**0,43**
CSAL1	0,02	0,18	-0,20	0,29	-0,02	**0,10**	0,04
E_LU1	-0,09	**0,75**	-0,14	-0,21	-0,01	-0,03	0,02
E_TU1	-0,08	-0,08	-0,08	-0,12	**0,07**	0,00	-0,05
E_VVR	-0,21	-0,03	0,05	-0,09	**0,12**	-0,03	-0,03
ETCO1	**-0,44**	**0,33**	**-0,38**	**0,58**	-0,01	**0,09**	-0,03
XMAHH	0,01	0,03	-0,02	-0,03	0,00	0,01	0,00

4.5. Classification spectrale pour chaque année

La construction du modèle MMC-NS par classification spectrale a défini un nombre d'états N égal à 7. Cette classification spectrale a été réalisée sur un ensemble de quatre années consécutives. Il s'agit maintenant de vérifier la pertinence des résultats en considérant les années individuellement afin de mieux discerner l'effet de la variabilité saisonnière *vs* la variabilité interannuelle dans les résultats. Pour répondre à cette question, une classification spectrale est réalisée pour chacune des années de 2005 à 2008 avec $N = 7$ états. Pour la suite, nous introduisons la notation $c_i(200X)$, ainsi, par exemple, $c_1(2005)$ correspond à l'état 1 défini à partir de la classification de l'année 2005.

4.5.1. Variabilité saisonnière et / ou interannuelle

Afin de comparer les partitions obtenues entre les états c_i de chaque année et s_j du modèle MMC construite sur les 4 années, nous utilisons l'indice de Rand (cf. section 3.3.1) ainsi que l'indice de Rand ajusté d'Hubert et Arabie initialement compris dans l'intervalle sur [-1,1]. Cet indice permet de savoir si l'indice de Rand obtenu est pertinent, pour cela, il doit être supérieur à zéro (Milligan et Cooper, 1986). Nous avons redéfini cet indice sur [0,1], appelé indice de confiance par la suite, afin d'en faciliter l'interprétation. Par conséquent il est nécessaire d'avoir un indice de confiance supérieur à 0,5.

Chacun des indices obtenus est pertinent et est de l'ordre de 0,90 (tableau 4.9). Ces résultats permettent de conclure que la partition obtenue sur la base $\boldsymbol{X}$ définit principalement des variabilités saisonnières. Cependant, il n'est pas opportun de construire notre modèle de

Markov caché à partir d'une seule année. En effet, les indices ne sont pas égaux à 1, les différences observées sont dues à la fois aux données manquantes (efflorescence de l'année 2007) et à une variabilité interannuelle (différents niveaux des efflorescences, par exemple en 2006, la fluorescence atteint un maximum de 45,99 FFU).

Tableau 4.9. Indice de Rand et son indice de confiance entre la partition calculée à partir la classification spectrale de la base $\boldsymbol{X}$ *et la partition obtenue pour chaque année de façon indépendante avec* $N = 7$.

Année	2005	2006	2007	2008
Indice de Rand	0,92	0,88	0,91	0,87
Indice de confiance	0,88	0,79	0,84	0,76

Une partie de la variabilité est saisonnière, nous recherchons maintenant quelle année structure chacun des états s_i.

4.5.2. Identification des années structurantes

Pour chacune des années de 2005 à 2008, une matrice de confusion notée mc entre les partitions des états s_i et c_j est calculée. Le pourcentage de structuration de chaque état s_i noté PS_i par année sur l'ensemble des états c_j, développé spécifiquement pour l'interprétation des résultats de ce travail de thèse, est calculé comme suit :

$$PS_i(année) = \frac{card(c(année)|s_i)}{card\big(mc(année)\big)} \qquad (4.6)$$

Le tableau 4.10 regroupe les pourcentages de structuration de chaque état s_i par année sur l'ensemble des états c_j. Nous remarquons que trois états sont structurés par des années distinctes :

- s_1 est structuré par les années 2005 et 2006 ;
- s_3 est structuré par les années 2007 et 2008 ;
- s_6 est structuré par l'année 2007.

Les états s_2 et s_4 sont structurés par les quatre années. En ce qui concerne les états s_5 et s_7, les pourcentages sont trop faibles pour affirmer qu'ils sont structurés par une ou des années en particulier.

Tableau 4.10. Pourcentages de structuration de chaque état $\boldsymbol{s_i}$ *pour chaque année sur l'ensemble des états* $\boldsymbol{c_j}$*. Les pourcentages mis en évidence permettent de connaître les années structurantes pour un état donné.*

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
$PS_i(2005)$	**23**	**19**	16	**34**	6	0	2
$PS_i(2006)$	**39**	**13**	13	**25**	4	0	6
$PS_i(2007)$	10	**12**	**27**	**27**	11	**13**	1
$PS_i(2008)$	11	**12**	**27**	**34**	11	0	4

4.5.3. États dominants

Pour connaître les états dominants $c_j(année)$ dans l'état s_i, le ratio entre chaque cellule de la matrice de confusion mc_{ij} de l'année étudié et le nombre de données présentes dans l'état s_i de cette année est calculé. Ce ratio, compris entre 0 et 1, est appelé le coefficient de domination $Dom_i(année)$ et est calculé comme suit :

$$Dom_{ij}(année) = \frac{mc_{ij}(année)}{card(c(année)|s_i)} \qquad (4.7)$$

En associant ce coefficient de domination avec les coefficients de corrélations entre les états $c_j(année)$ et les paramètres (Annexe 3), il est possible de définir l'année et les paramètres structurants d'un état s_i (tableaux 4.11 à 4.14).

Tableau 4.11. Coefficient de domination pour chaque état de l'année 2005 et paramètres structurants associés.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	Paramètres structurants (moyenne $\pm$ écart-type)
$c_1(2005)$	-	-	-	-	-	-	-	
$c_2(2005)$	-	-	-	-	-	-	-	
$c_3(2005)$	-	-	-	0,65	-	-	-	Très forte température (17,62 $\pm$ 2,12 °C) Faible concentration en nitrate (6,43 $\pm$ 6,14 µmol.L^{-1}) et en oxygène dissous (7,13 $\pm$ 0,87 mg.L^{-1})
$c_4(2005)$	0,64	-	-	0,25	-	-	-	Concentration en oxygène dissous moyenne (9,50 $\pm$ 1,35 mg.L^{-1}) Concentration en silicate faible (1,42 $\pm$ 1,32 µmol.L^{-1})
$c_5(2005)$	-	-	-	-	-	-	-	
$c_6(2005)$	0,33	-	-	-	-	-	-	Concentration en nitrate moyenne (21,69 $\pm$ 8,80 µmol.L^{-1}) Température de l'eau faible (8,17 $\pm$ 1,14 °C) Concentration en oxygène dissous forte (9,88 $\pm$ 1,23 mg.L^{-1})
$c_7(2005)$	-	0,67	-	-	-	-	-	Forte luminosité (1 263 $\pm$ 353,89 µmol de photons .s^{-1}.m^{-2})

Tableau 4.12. Coefficient de domination pour chaque état de l'année 2006 et paramètres structurants associés.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	Paramètres structurants (moyenne $\pm$ écart-type)
$c_1(2006)$	-	-	-	0,52	-	-	-	Forte Température (17,48 $\pm$ 2,69 °C) Concentration en nitrate faible (4,52 $\pm$ 6,26 µmol.L^{-1}) Faible turbidité (2,42 $\pm$ 3,82 NTU)
$c_2(2006)$	0,64	-	-	-	-	-	-	Concentration en oxygène dissous forte (11,42 $\pm$ 0,54 mg.L^{-1}) Température faible (5,56 $\pm$ 0,69 °C)
$c_3(2006)$	0,33	-	-	-	-	-	-	Concentration en nitrate faible (5,60 $\pm$ 3,74 µmol.L^{-1}) Une température faible (8,74 $\pm$ 6,62 °C) Forte salinité (33,99 $\pm$ 0,39)
$c_4(2006)$	-	-	-	-	-	-	-	
$c_5(2006)$	-	-	-	-	-	-	-	
$c_6(2006)$	-	0,76	-	-	-	-	-	Forte luminosité (1 291 $\pm$ 364,72 µmol de photons .s^{-1}.m^{-2}) Très faible concentration en nitrate (4,65 $\pm$ 3,78 µmol.L^{-1})
$c_7(2006)$	-	-	-	0,33	-	-	-	Concentration en nitrate élevée (24,97 $\pm$ 10,09 µmol.L^{-1}) Vents puissants (12,1 $\pm$ 5,50 m.s^{-1})

Tableau 4.13. Coefficient de domination pour chaque état de l'année 2007 et paramètres structurants associés.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	Paramètres structurants (moyenne $\pm$ écart-type)
$c_1(2007)$	-	-	-	-	-	-	-	
$c_2(2007)$	-	-	-	-	-	0,97	-	Forte salinité (35,07 $\pm$ 0,64) : saut de capteur
$c_3(2007)$	-	0,82	-	-	-	-	-	Forte luminosité (1 214 $\pm$ 421,52 µmol de photons .s^{-1}.m^{-2}) Forte température (17,18 $\pm$ 1,60 °C) Concentration en silicate faible (3,28 $\pm$ 1,51 µmol.L^{-1})
$c_4(2007)$	-	-	0,23	-	-	-	-	Fortes concentrations en nitrate (45,94 $\pm$ 16,28 µmol.L^{-1}) et phosphate (8,15 $\pm$ 13,88 µmol.L^{-1})
$c_5(2007)$	-	-	0,69	-	-	-	-	Forte concentration en silicate (13,33 $\pm$ 2,14 µmol.L^{-1}) Très faible température (8,75 $\pm$ 0,92 °C)
$c_6(2007)$	-	-	-	0,81	-	-	-	Forte température (17,54 $\pm$ 1,17 °C) Faibles concentrations en silicate (3,27 $\pm$ 1,65 µmol.L^{-1}) et nitrate (12,15 $\pm$ 6,75 µmol.L^{-1})
$c_7(2007)$	-	-	-	-	-	-	-	

Tableau 4.14. Coefficient de domination pour chaque état de l'année 2008 et paramètres structurants associés.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	Paramètres structurants (moyenne $\pm$ écart-type)
$c_1(2008)$	-	-	0,66	-	-	-	-	Forte concentration en nitrate ($48{,}28 \pm 14{,}94$ µmol.L^{-1}) Très faible température ($8{,}07 \pm 1{,}10$ °C)
$c_2(2008)$	-	-	-	0,47	-	-	-	Forte température ($16{,}5 \pm 2{,}10$ °C) Faible concentration en oxygène dissous ($7{,}47 \pm 0{,}69$ mg.L^{-1})
$c_3(2008)$	-	-	-	-	-	-	-	
$c_4(2008)$	-	-	0,21	0,47	-	-	-	Faible concentration en oxygène dissous ($7{,}60 \pm 0{,}68$ mg.L^{-1})
$c_5(2008)$	-	-	-	-	-	-	-	
$c_6(2008)$	-	0,80	-	-	-	-	-	Fortes luminosité ($1\,244 \pm 398{,}48$ µmol de photons .s^{-1}.m^{-2}) et température ($16{,}86 \pm 2{,}10$ °C)
$c_7(2008)$	-	-	-	-	-	-	-	

L'état s_1 est structuré par les années 2005 et 2006. Dans la section 4.4.1, l'état s_1 a été défini comme la phase d'initiation de l'efflorescence phytoplanctonique principale et l'étape de croissance phytoplanctonique. Les caractérisations ci-dessus confirment que nous nous trouvons dans un tel état. Les années 2007 et 2008 ont très peu participé à la formation de cet état puisque des données manquantes se situent au moment de la période d'efflorescence.

L'état s_2 est structuré par les années 2005 à 2008 et correspond à l'état d'efflorescence régénérée qui se produit en été (confirmé par la description ci-dessus).

L'état s_4 correspond également à un état d'efflorescence régénérée et est structuré ici par les années 2005 à 2008.

L'état s_3 est structuré par les années 2007 et 2008. Cet état s_3 avait été étiqueté comme étant la période non productive. Ceci est confirmé par la structuration, à partir des classifications, des années prises indépendamment les unes des autres.

L'état s_6 correspondant à la défaillance du capteur de salinité durant l'année 2007 a bien été détecté lors de la classification de cette année uniquement.

Les états s_5 et s_7 correspondent aux évènements rares, extrêmes et de courtes durées, c'est pourquoi ceux-ci ne sont pas représentatifs d'un état dans la classification par année.

Lors de la construction du MMC-NS, les variabilités saisonnières et interannuelles sont prises en compte. Nous avons donc un maximum d'information permettant d'estimer au mieux les états d'une nouvelle base de données $\boldsymbol{X}(T)$.

4.6. Conclusions - Perspectives

Deux modèles de Markov cachés à N-états sont construits dans le but de mieux comprendre la dynamique des efflorescences à partir des signaux haute fréquence multi-capteurs de la station de mesure MAREL-Carnot (Ifremer, Boulogne-sur-Mer) sans aucune connaissance biologique *a priori* (notion d'étiquetage des différents instants).

L'analyse des résultats obtenus démontre l'intérêt et la stabilité de chaque algorithme (génération des états et des symboles) défini au chapitre 3 : l'information haute fréquence est préservée. La construction du MMC-NS à 2-états permet de détecter les périodes productive et non-productive, tel qu'établi par la DCE - 2000/60/CE pour évaluer le bon état environnemental. Un MMC-NS à 7-états a été proposé pour améliorer la connaissance sur la dynamique de l'efflorescence phytoplanctonique dans un écosystème tempéré, temporairement dominé par une algue nuisible : *Phaeocystis globosa*. Le séquencement d'état

obtenu coïncide avec la dynamique décrite à partir des mesures de systèmes basse fréquence proche de la station MAREL-Carnot (réseau REPHY / SRN).

Le système MMC-NS proposé permet de définir l'appartenance à un état pour de nouvelles données entrantes. En utilisant les principales caractéristiques statistiques des paramètres mis en avant pour un état donné, il est possible de mieux comprendre les facteurs de contrôle et les effets environnementaux directs et indirects des efflorescences. Par ailleurs, appliquer en temps quasi-réel, le système MMC-NS s'avère être un outil très prometteur en terme de détection de situations à risque ou inhabituelles. En effet, l'apparition de l'état qui caractérise une efflorescence peut être liée à la prolifération d'une espèce nuisible. Le scientifique pourra alors au regard de ce résultat décider d'effectuer des prélèvements et des analyses supplémentaires (mise en place d'une stratégie adaptative).

A terme, il serait intéressant d'aller au-delà d'une simple estimation de la biomasse globale du phytoplancton en déployant des systèmes comme les fluorimètres spectraux ou les cytomètres en flux, qui permettraient d'avoir accès à une information de type approche taxonomique préliminaire (classes d'algues définies par rapport à des empreintes de fluorescence) ou plus approfondie, parfois jusqu'à l'espèce, respectivement.

De même, afin d'améliorer l'interprétation des résultats, l'acquisition de données environnementales (vents, courants, …) ou de contexte (dragage, ouverture de barrage,…) complémentaires s'avère nécessaire.

La principale limite dans la construction du modèle est due aux données manquantes : cela a un effet sur l'estimation des états et leurs caractérisations (génération des symboles). Par exemple, l'efflorescence de l'année 2007 n'est pas prise en compte ce qui entraine une erreur dans la construction de la matrice de transitions.

Pour y remédier, deux possibilités sont envisageables :

- Un traitement des données en amont par complétion ;
- Une segmentation des états par bloc.

Ensuite, le modèle pourrait être amélioré notamment vis-à-vis de ses matrices probabilistes d'émissions et de distributions initiales en utilisant les approches développées dans (Schmitt et Huang, 2014) cherchant à caractériser les cycles et les fréquences des états par fonctions de densités de probabilités.

Chapitre 5 : Autres applications : systèmes instrumentés et autres environnements

5.1. Introduction

Les deux chapitres précédents ont permis d'expliquer comment définir un Modèle de Markov Caché régi par des observations sans labellisation (non supervisé) et l'intérêt d'adapter celui-ci à partir de mesures hautes fréquences issues de la station MAREL-Carnot pour décrire la dynamique de la biomasse phytoplanctonique.

Les observations MAREL-Carnot ont une résolution de 20 minutes et ont été acquises sur une longue période (2005-2009). Afin de tester la capacité du modèle proposé à répondre à d'autres problématiques environnementales, l'approche a été mise en œuvre sur des observations issues de systèmes de mesures différents et déployés afin de répondre à des objectifs sensiblement différents. Ainsi, alors que la station MAREL-Carnot permet d'aborder à haute résolution temporelle la dynamique de la qualité de l'environnement marin en point fixe (approche Eulérienne), il est ici question de s'intéresser à des données à :

1. Haute résolution temporelle ET spatiale (approche Lagrangienne) en milieu marin, acquise à l'échelle d'une semaine ;
2. Haute résolution temporelle issues d'une station instrumentée mis en service en eau douce, durant une période de 1 mois.

Les données recueillies par un système mobile, le Pocket FerryBox, seront décrites dans la première partie. Ces observations ont été collectées lors d'une campagne océanographique qui a eu lieu en 2012 en Manche orientale. L'objectif est d'avoir une vision synoptique à haute résolution et à l'échelle de la Manche orientale de la communauté phytoplanctonique, en pleine période d'efflorescence, en définissant des ensembles de masses d'eaux sur la base des différences de composition pigmentaire des groupes phytoplanctoniques qui y sont détectés par fluorimétrie spectrale. Il s'agit ainsi de tester la capacité du modèle à estimer cette typologie des masses d'eaux à partir de jeux de données nouveaux.

Dans la deuxième partie, seront traitées des données issues d'une station fixe implémentée temporairement sur la rivière Deûle en 2009. Cette étude a pour objectif de contribuer aux réflexions visant à la mise en place d'un programme pérenne d'observation et de surveillance du phytoplancton en eau douce en insistant sur la valeur ajoutée des mesures à haute résolution par rapport aux approches conventionnelles, basses résolutions. Parallèlement, il s'agira aussi de vérifier que le modèle est capable de détecter un état environnemental particulier, correspondant à des perturbations qui seraient engendrés par la navigation sur ce cours d'eau.

5.2. Haute résolution spatiale en milieu marin

L'étude ci-après a été menée sur des observations extraites d'un système mobile haute fréquence appelé Pocket FerryBox lors de campagnes menées lors du projet DYMAPHY en 2012 en Manche orientale.

5.2.1. Contexte scientifique

Le projet InterReg IVa « 2 Mers Seas Zeeën », intitulé DYMAPHY, Développement d'un système d'observation DYnamique pour la détermination de la qualité des eaux MArines, basé sur l'analyse du PHYtoplancton, a été cofinancé par le Fonds Européen de Développement Régional (FEDER) de 2010 à 2013 (www.dymaphy.eu). Ce projet regroupait plusieurs partenaires scientifiques de laboratoires de recherches français, anglais et hollandais en bordure des « 2 Mers » (la Manche et la Mer du Nord) dont les entités associées sont les suivantes :

- IFREMER, l'Institut Français de Recherche pour l'Exploitation de la MER, dont notamment le Laboratoire Environnement et Ressources de Boulogne-sur-Mer du Centre Manche Mer du Nord, France ;
- L'Université du Littoral Côte d'Opale (ULCO) au travers de 2 laboratoires : le Laboratoire d'Informatique, Signal et Image de la Côte d'Opale (LISIC) à Calais et le Laboratoire d'Océanologie et Géoscience (LOG) à Wimereux, France ;
- CEFAS, le Centre for Environment Fisheries & Aquaculture Science, Lowestoft, Angleterre ;
- Le Rijkswaterstaat, Middelbourg, Pays-Bas ;
- L'Université de Lille 1 (LOG), France ;
- Le Centre National de la Recherche Scientifique (CNRS-LOG), France.

Les objectifs principaux étaient :

- Le développement de procédures opérationnelles standards pour la surveillance de la structure de la communauté phytoplanctonique *in situ* et en temps réel ;
- La construction d'une banque d'observations multi-sites ;
- La mise en place d'outils d'évaluation de la qualité des eaux marines dans la région des « 2 Mers ».

Afin d'atteindre ces objectifs, plusieurs campagnes ont été réalisées à bord du Navire Océanographique « Côtes de la Manche » dont notamment celles de 2012 découpées en 3 tronçons notés Leg 1 à 3 dont les périodes d'échantillonnage sont les suivantes (figure 5.1) :

- Leg 1 du 20 au 21 avril 2012 ;
- Leg 2 du 27 au 30 avril 2012 ;
- Leg 3 du 31 mai au 4 juin 2012.

Seules les données collectées par le Pocket FerryBox lors des Leg 1 (1 567 prélèvements) et Leg 2 (2 593 prélèvements) seront étudiés ici.

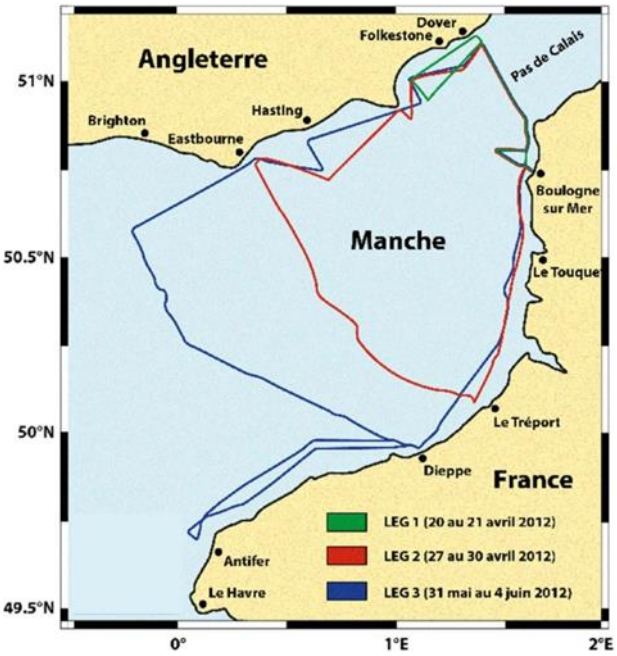


Figure 5.1. Représentation du trajet du Navire Océanographique « Côtes de la Manche » lors des campagnes en mer réalisé en 2012 : Leg 1 en vert, Leg 2 en rouge, Leg 3 en bleu.

5.2.2. Présentation des données et prétraitements

Le Pocket Ferry Box utilisé de marque 4H-JENA © (www.4h-jena.de) est un système de mesures de 27 kg autonome grâce à sa batterie portable (25 kg). Il est équipé de plusieurs capteurs (tableau 5.1). La particularité du système déployé dans DYMAPHY est d'être couplé à un fluorimètre spectral (Algae Online Analyser, AOA, bbe © (www.bbe-moldaenke.de/home/)) (figure 5.2).

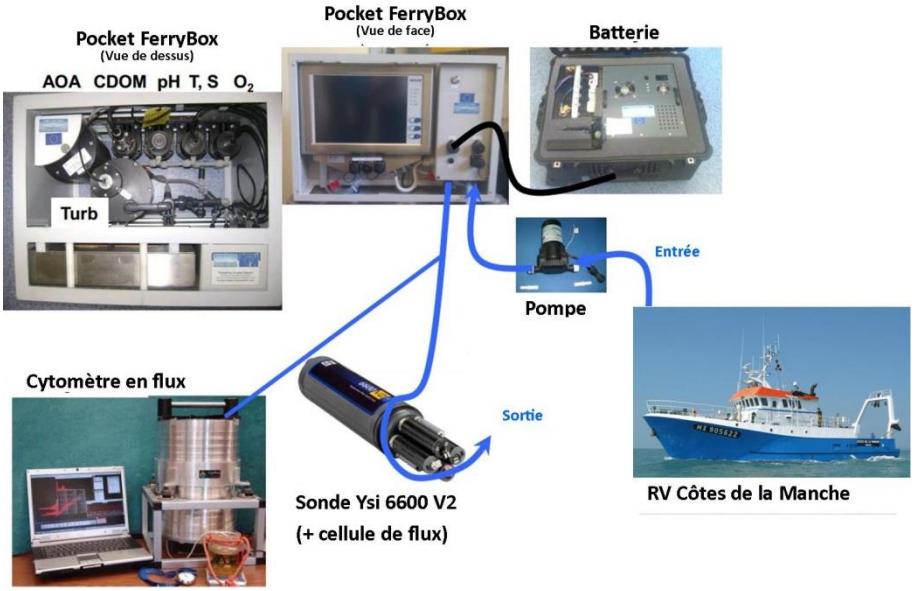


Figure 5.2. Le Pocket FerryBox, couplé à une sonde Ysi et un cytomètre en flux, tel que déployé lors des campagnes DYMAPHY en 2012.

Tableau 5.1. Paramètres mesurés par le Pocket FerryBox avec la gamme capteur et la précision de mesure.

Paramètre	Unité	Gamme capteur	Précision
Conductivité	mS/cm	0 – 70	0,003
Température	°C	-3 – 35	0,002
Salinité	PSU	2 – 42	0,005
Oxygène dissous	µmol/L^{-1}	0 – 500	8
Saturation en oxygène	%	0 – 120	0,4
Chlorophylle totale	µg.Chl-*a*.L^{-1}	0 – 200	0,01
Turbidité	NTU	0 – 750	0,2
pH	UpH	0 – 14	0,1
Matière organique dissoute colorée	ppb	0 – 2500	25ppb

Les eaux marines présentent une grande diversité phytoplanctonique qui peut être représentée par différents groupes d'algues. Les plus communs d'entre eux sont le groupe des algues bleu-vert (Cyanophycées), le groupe des algues vertes (Chlorophycées), le groupe des algues brunes (Bacillariophytes ou Diatomées, Dynophytes ou dinoflagellés) et un groupe dit mixé (Cryptophycées). Sur cette base, le fluorimètre spectral (AOA) permet la détermination du spectre de fluorescence et des cinétiques de fluorescence des algues. Les différentes classes d'algues peuvent être différenciées par leurs compositions pigmentaires et par conséquent par les différentes réponses de fluorescence à une lumière de différentes couleurs (figure 5.3). Chaque classe d'algue à son empreinte caractéristique, qui est un schéma spécial selon lequel elle répond à des excitations de différentes longueurs d'onde (470, 525, 570, 590 et 610 nm). Afin de s'affranchir de l'interférence provoquée par la présence de matières organiques dissoutes colorés, une mesure est également faite à 370 nm. Pour différencier les classes d'algues, les proportions de chacune des classes sont calculées à partir d'une méthode d'optimisation basée sur un modèle additif des spectres de chaque classe (Beutler *et al.*, 2002; Ruser *et al*., 1999). Le logiciel bbe fournit les concentrations de chaque classe en équivalent chlorophylle-*a* par litre (eq µg. Chla.L^{-1}) sachant que la gamme du capteur est comprise entre 0 et 200 eq µg. Chla.L^{-1} avec une précision de 0,01 eq µg. Chla.L^{-1}.

Les concentrations en équivalent chlorophylle-*a* des quatre classes d'algues formeront la base d'observations. Après la vérification que ces concentrations se situent dans la gamme du capteur, elles seront centrées-réduites avant de construire l'arbre de décision d'un classifieur hiérarchique et du Modèle de Markov Caché Non Supervisé (MMC-NS).

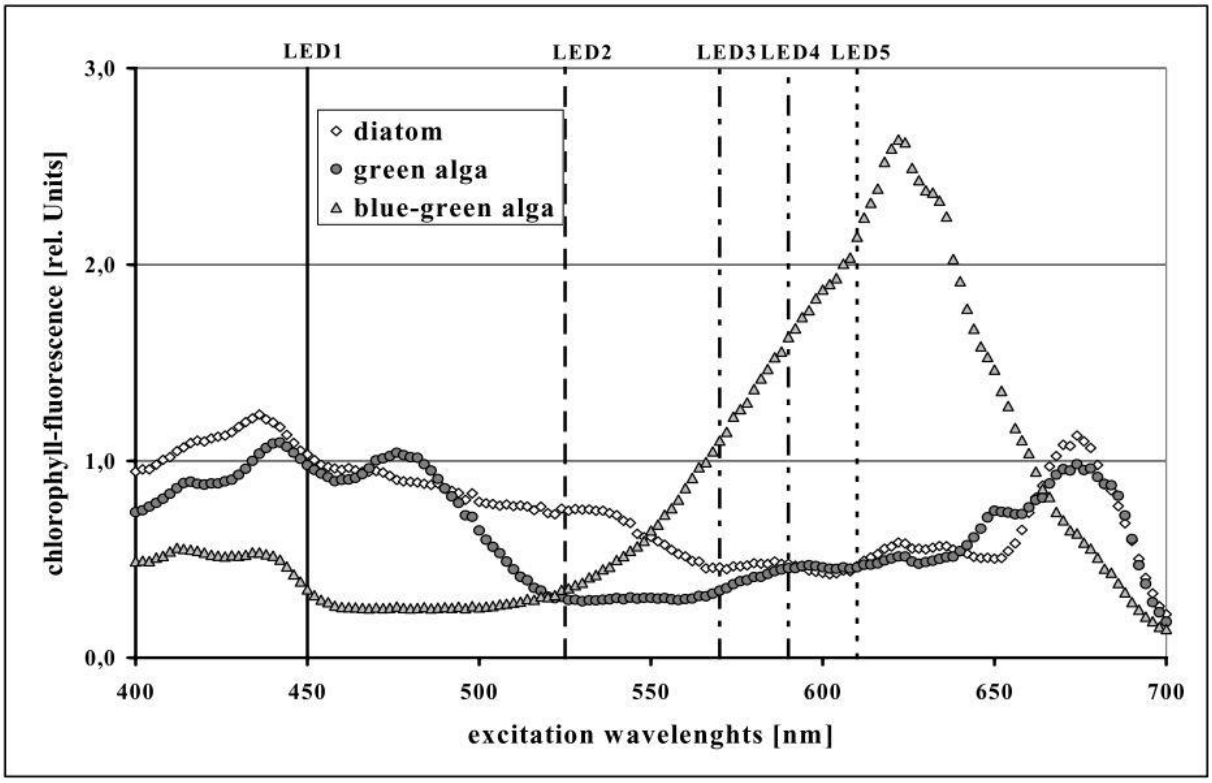


Figure 5.3. Spectre d'excitation de la fluorescence de certaines bacillariophycées, (diatomées), de chlorophycées (algues vertes) et de cyanophycées (algues bleues-vertes) à une longueur d'onde d'émission de 720 nm. Les cinq longueurs d'onde d'excitation de l'AOA sont représentées afin d'illustrer la notion d'empreinte caractéristiques de chaque classe d'algue (source : Ruser et al., 1999).

5.2.3. Classification des observations

Deux techniques de segmentation automatique sont comparées à un découpage à dire d'experts : la classification hiérarchique et la classification par MMC-NS. Afin de comparer les résultats obtenus, l'indice de Rand et son indice de confiance sont utilisés (Chapitre 4). Pour rappel, ces deux indices sont compris entre 0 et 1, sachant qu'un indice de Rand égal à 1 signifie que les segmentations sont identiques et un indice de confiance inférieur à 0,5 signifie que la partition n'est pas déterministe mais peut-être obtenue aléatoirement.

5.2.3.1. Découpage à dire d'experts

L'évolution de la proportion relative des classes algales et de la concentration totale en chlorophylle estimées par le fluorimètre spectrale (figure 5.4 (a)) permet de mettre en évidence certains schémas de distribution des classes algales et une variabilité qui n'aurait pu être observé via la mise en œuvre d'approches conventionnelles :

- Les zones caractérisées par une dominance des classes d'algues brunes et vertes (qui reflètent en fait la présence de la prymnésiophycée *Phaeocystis globosa*) se situent essentiellement du côté français. La concentration en chlorophylle totale chute considérablement entre le début et la fin du Leg2, ce qui traduit la phase de déclin du bloom de *Phaeocystis globosa.*

- Les cryptophycées et les algues bleu-vert ne sont observées qu'en Manche centrale et le long des côtes anglaises.
- La concentration en chlorophylle totale est particulièrement faible le long des côtes anglaises par rapport à ce qui est mesuré côté français.

Les changements observés se font à l'échelle de quelques minutes, et donc à l'échelle de quelques centaines de mètres, et confirment ainsi la distribution hétérogène du phytoplancton (notion de distribution par patch ou patchiness) en Manche orientale malgré un fort brassage des masses d'eaux. D'un point de vue biologique, (Seuront, 2005) indique que la prymnésiophycée *Phaeocystis globosa* aurait un rôle important dans ce mode de distribution en raison de ces propriétés cohésives (processus de coagulation entre les cellules phytoplanctoniques). Par ailleurs, d'un point de vue physique, la présence de la structure frontale en zone côtière française (le « fleuve côtier » ; (Brylinski et Lagadeuc, 1990)), les transports de masses d'eaux liés à la marée et / ou au vent, la dérive résiduelle vers le nord-est jouent un rôle important sur cette distribution.
Cette notion de patchiness est structurante pour la stabilité, la dynamique de l'écosystème, la diversité et la productivité régionale (Martin, 2003). A titre d'exemple, la distribution par patch du phytoplancton est importante pour la dynamique des relations proies-prédateurs et par conséquent elle est essentielle pour comprendre les relations trophiques au sein de l'écosystème marin. Des relations non-optimales entre phyto~ et zooplancton au niveau des premiers maillons peuvent ainsi avoir des conséquences négatives aux plus hauts niveaux trophiques et par conséquent impacter la disponibilité de la ressource halieutique. En termes de flux au sein de l'écosystème (bilan Carbone), la non prise en compte de ce patchiness par l'utilisation de résultats acquis à des échelles de temps et d'espace non adaptées entraine inévitablement une sur-estimation de la biomasse du phytoplancton et introduit par conséquent un biais dans toutes les estimations de flux entre compartiments biologiques.

L'interprétation des graphiques de proportions des classes algales (figure 5.4) à dire d'experts, c'est à dire, ici, sans recours aux statistiques, de manière pragmatique via une simple analyse exploratoire des résultats bruts, a conduit à une segmentation différente pour chaque Leg. Les protocoles de découpage des observations à dire d'experts pour le Leg1 (Leg2), noté Leg1Ex (respectivement, Leg2Ex) sont détaillés ci-dessous.

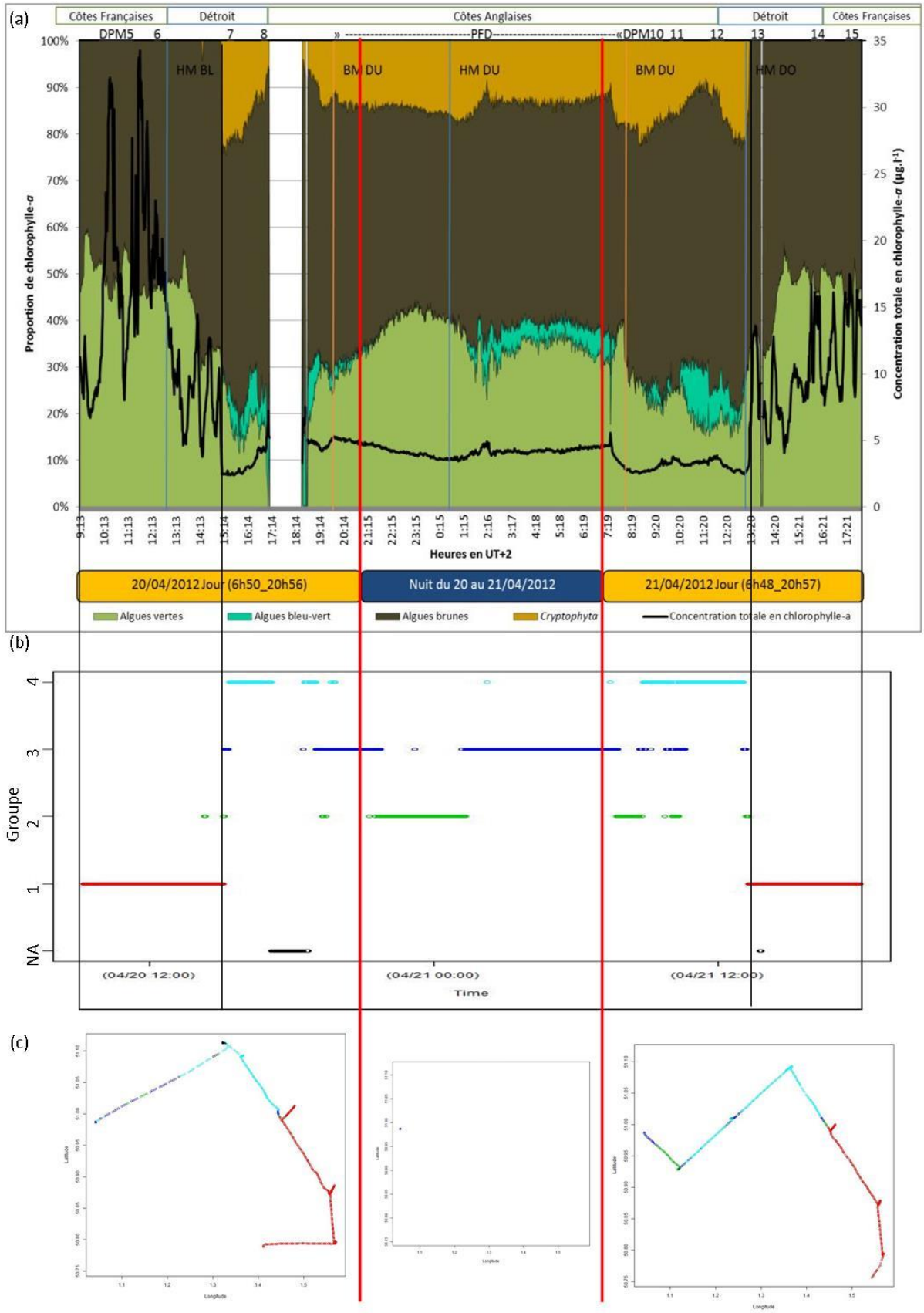


Figure 5.4. Segmentation à dire d'experts du Leg 1 : (a) répartition des différents groupes d'algues phytoplanctoniques avec les empreintes d'algues vertes, bleu-vert, brunes et les

cryptophycées, (b) le séquencement des groupes définis à dire d'experts et (c) la projection de cette segmentation sur le trajet effectué lors de ce Leg 1.

Données issues du Leg 1

Le premier découpage consiste à séparer les observations selon la présence ou l'absence d'algues autres que des algues de type brunes et vertes. On note s_1 le groupe contenant ces deux types d'algues. Le second groupe s_2 contient les deux groupes de s_1, mais également des cryptophycées (couleur orange sur la partie haute de la figure 5.4). Lorsque l'ensemble des algues sont présentes (brunes, vertes, cryptophycées, bleu-vert), elles sont regroupées dans le groupe s_3.

Puisque pour une composition spectrale donnée (nombre d'empreintes présentes), il apparaît des proportions différentes pour certains groupes, il a été décidé de subdiviser certains ensembles. Ainsi, chacun des groupes définis ci-dessus sera divisé en deux selon la proportion des algues vertes présente : supérieure ou inférieure à 25 %, valeur arbitraire définie par l'expert. Le groupe s_3 est divisé en s_3 (proportion d'algues vertes supérieure à 25 %) et s_4 (proportion d'algues vertes inférieure à 25 %). Le groupe s_2 est divisé en s_2 (proportion d'algues vertes supérieure à 25 %) et s_5 (proportion d'algues vertes inférieure à 25 %). Le groupe s_1 est divisé en s_1 (proportion d'algues vertes supérieure à 25 %) et s_6 (proportion d'algues vertes inférieure à 25 %). Un découpage en 6 groupes est donc obtenu.

Le nombre d'instants contenus dans les groupes s_5 et s_6 étant inférieur à 10 pour un nombre total de 1 567 instants, le découpage en fonction des algues vertes peut ne pas être considéré (tableau 5.2) comme porteur d'une grande information. Dans ce cas, un découpage en 4 groupes sera retenu.

Tableau 5.2. Nombre d'instants contenus dans chaque groupe (NA correspond aux données manquantes) obtenu lors du découpage de la base de données du Leg1 à dire d'experts.

	NA	s_1	s_2	s_3	s_4	s_5	s_6
$N = 6$	71	500	242	458	287	7	2
$N = 4$	71	502	249	458	287	-	-

Données issues du Leg 2

Un premier découpage en 3 groupes est effectué tel que :

- s_1 est le groupe contenant uniquement des algues brunes et vertes ;
- s_2 est le groupe où toutes les classes d'algues (brunes, vertes, cryptophycées, bleu-vert) sont présentes.
- s_3 est le groupe où des algues brunes et vertes mais également des cryptophycées sont présentes à chaque instant.

Pour les mêmes raisons que précédemment, le groupe s_3 peut être découpé selon la proportion en algues brunes. Le groupe s_3 rassemble les instants dont la proportion d'algues brunes est inférieure à 60 % et s_4 ceux pour laquelle elle est supérieure à 60 %. Le groupe s_2 peut être

découpé selon la proportion en cryptophycées. Le groupe s_2 rassemble les instants dont la proportion d'algues brunes est supérieure à 60 % et s_4 ceux pour laquelle elle est inférieure à 25 %. Une segmentation en cinq groupes est ainsi obtenue (tableau 5.3).

Tableau 5. 3. Nombre d'instants contenus dans chaque groupe (NA correspond aux données manquantes) obtenu lors du découpage de la base de données du Leg2 à dire d'experts.

	NA	s_1	s_2	s_3	s_4	s_5
$N = 3$	221	1 597	391	384	-	-
$N = 5$	221	1 597	368	204	23	180

5.2.3.2. Approche conventionnelle : classification hiérarchique

En écologie numérique, il est d'usage d'utiliser une classification hiérarchique (CH) ascendante (Borcard *et al.*, 2011; Legendre et Legendre, 1998) lorsqu'il s'agit de créer des groupes d'éléments (espèces, stations de prélèvements, …) présentant un certain niveau de ressemblance (ou de dissemblance). Cette méthode de classification est donc appliquée sur les données du Leg 1 avec le paramétrage suivant : distance euclidienne et méthode de Ward (agrégation des classes selon un critère d'inertie interclasse maximum). Ceci permettra de comparer les résultats avec le système MMC-NS proposé (section suivante). Le nombre de groupes a été fixé à $N = 4$ afin de comparer la partition obtenue avec le découpage effectué à dire d'experts. La projection des états sur le trajet effectué durant cette campagne ainsi que le séquencement des états permettent d'apprécier temporellement le découpage (figures 5.5 (b) et 5.5 (c)).

Avec un nombre de groupes fixé identique ($N = 4$), l'indice de Rand entre la partition obtenue par classification hiérarchique et celle experte est non biaisé et a un score de 0,76 avec un indice de confiance de 0,69. Il est donc possible de conclure que 70 % du découpage est similaire. La table de confusion (tableau 5.4) permet de montrer les différences de partitionnement. Seul l'état $s_2(CH)$ (composition en algues vertes et algues brunes uniquement) a une correspondance identique entre les deux partitions.

Tableau 5.4. Table de confusion entre les deux partitions obtenues par classification hiérarchique (CH) et classification experte.

CH vs expert	s_1	s_2	s_3	s_4
$s_1(CH)$	262	240	0	0
$s_2(CH)$	6	0	243	0
$s_3(CH)$	0	0	173	285
$s_4(CH)$	0	0	107	180

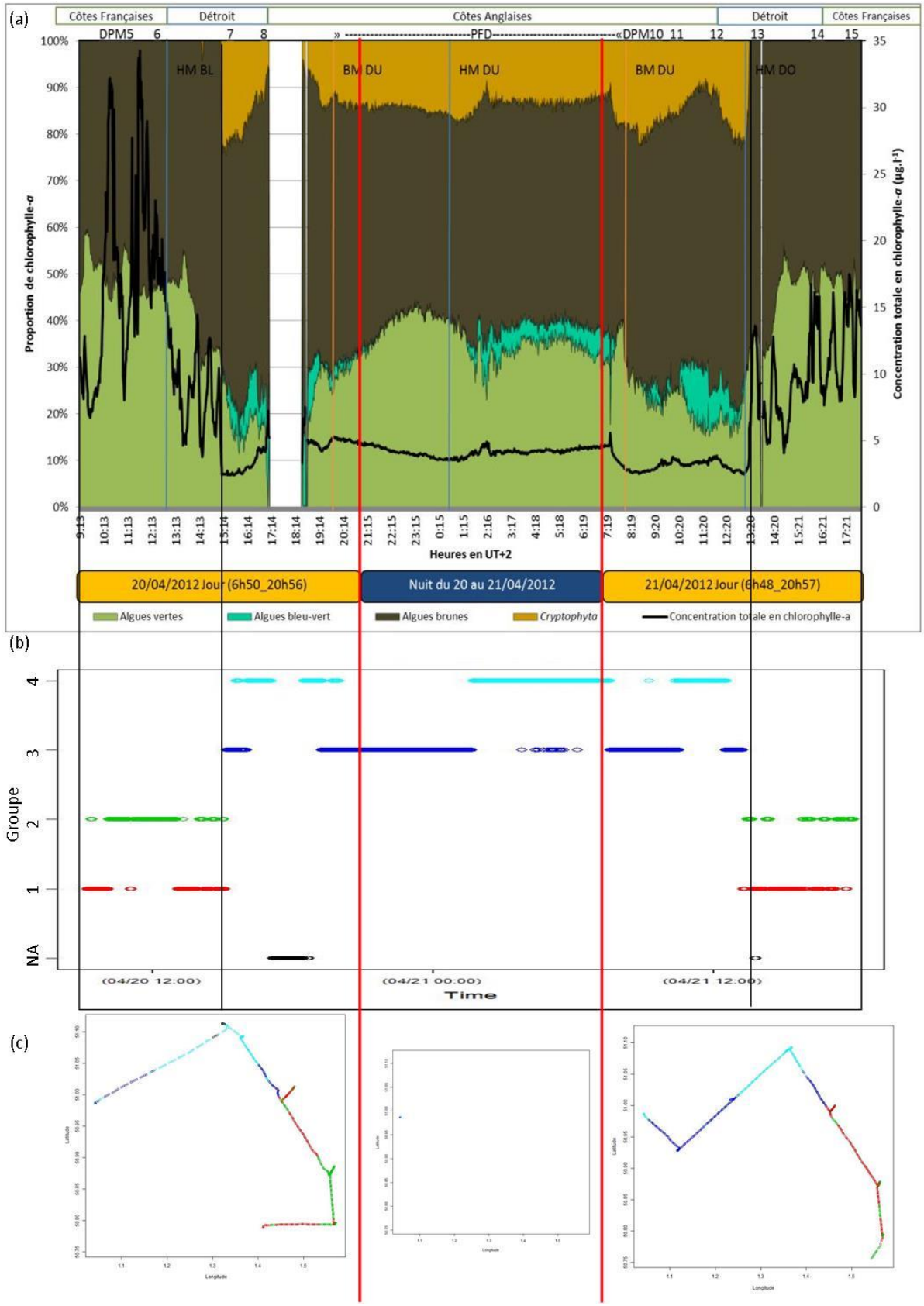


Figure 5.5. Segmentation classification hiérarchique ascendante du Leg 1 : (a) répartition des différents groupes d'algues phytoplanctoniques avec les empreintes d'algues vertes, bleu-vert, brunes et les cryptophycées, (b) le séquencement des groupes définis la classification hiérarchique et (c) la projection de cette segmentation sur le trajet effectué lors de ce Leg 1.

La différence de découpage résiduelle peut s'expliquer :

- D'une part, par le fait que l'expert utilise l'information de proportion, alors que l'arbre CH est construit à partir des concentrations brutes.
- D'autre part, si on analyse l'arbre de décision, associé à la classification hiérarchique obtenu à celui de la classification expert, on remarque que la tête de l'arbre est différente avec un seuil non nul de cryptophycées. En effet, celui-ci a été construit par agrégation de classes de lien minimum (inertie intra-classe minimum) contrairement à la découpe descendante de l'expert (figures 5.6 et 5.7).

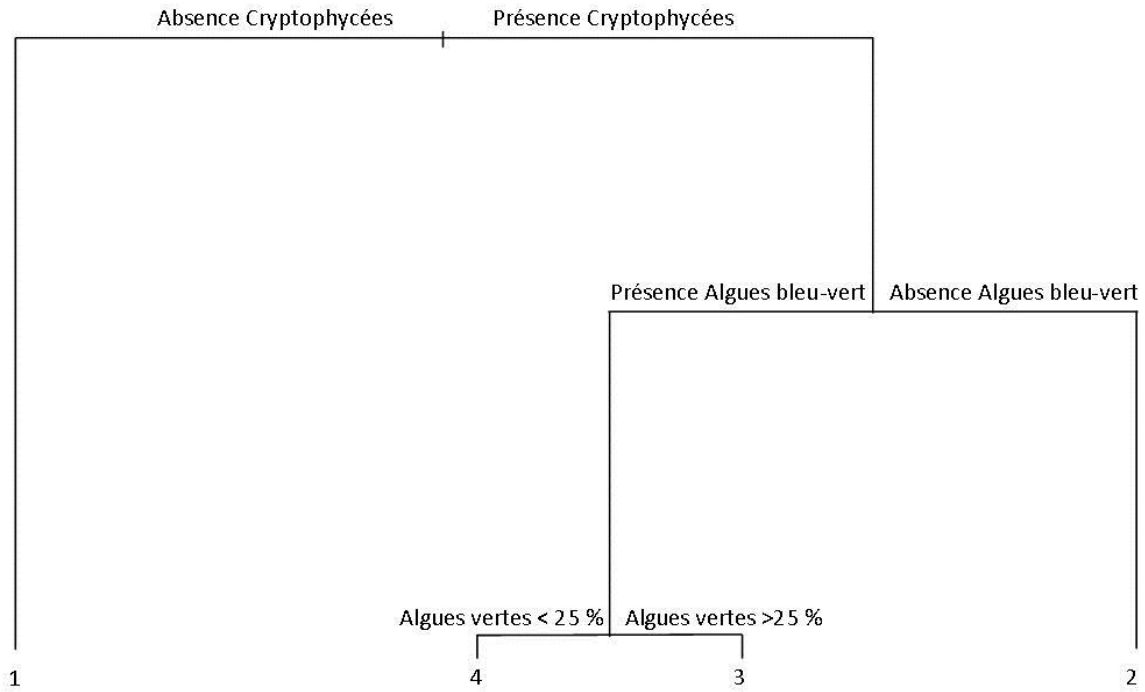


Figure 5.6. Arbre de décision issue de la classification à dire d'experts où les séparations sont effectuées selon les proportions de classes algales.

En se basant uniquement sur les proportions des classes algales, on note que les groupes s_1 et s_2 définis par la classification hiérarchique sont caractérisés par de fortes proportions en algues vertes et brunes (ainsi qu'une très faible proportion de cryptophycées dans le groupe s_1 : 0,13 %). Les groupes s_3 et s_4 sont principalement structurés par les algues bleu-vert (tableau 5.5).

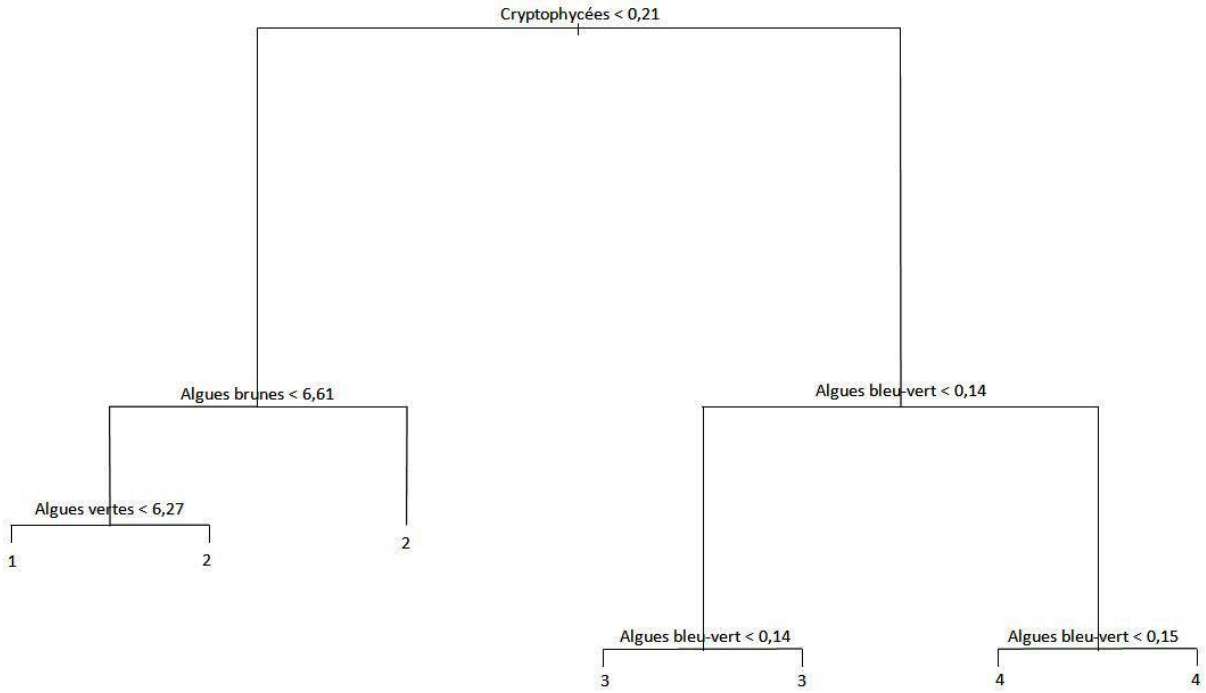


Figure 5.7. Arbre de décision issu de la classification hiérarchique sur le Leg 1 où les séparations sont effectuées selon les concentrations de classes algales.

Tableau 5.5. Proportions relatives (%) des quatre classes algales réparties dans les groupes définis lors de la classification hiérarchique ascendante (distance euclidienne et méthode de Ward).

	Vert	Bleu-vert	Brunes	Cryptophycées
s_1	46,29	0,00	53,58	0,13
s_2	46,12	0,00	53,88	0,00
s_3	32,14	1,17	50,36	16,33
s_4	27,09	6,55	53,92	12,45

Trois expériences sont ensuite menées pour tester les capacités d'estimation d'une classification hiérarchique vis-à-vis de la classification à dire d'experts. L'arbre de décision construit à partir des données du Leg 1 et de leurs labels par classification hiérarchique est réutilisé afin d'estimer de nouvelles données.

Une première comparaison entre la segmentation estimée du Leg 1 en utilisant l'arbre de décision et celle associée à la génération de l'arbre à $N = 4$ donne un indice de Rand et son indice de confiance à 0,98, ce qui permet de valider la construction de l'arbre.

Afin de tester les capacités d'apprentissage de cet arbre de décision, la segmentation estimée du Leg 1 est comparée à la labellisation experte. L'indice de Rand entre la segmentation estimée et celle à dire d'experts est égale à 0,76 avec un indice de confiance de 0,69. La table de confusion (tableau 5.6) permet de conclure que l'approche par classification hiérarchique permet d'identifier clairement l'état s_1 (présence des algues vertes et brunes uniquement) en

deux groupes. Cependant, elle ne permet pas de discriminer les états s_2, s_3, s_4 du dire d'experts, les groupes obtenus s_3 et s_4 étant un mélange de ces états.

Tableau 5. 6. Table de confusion entre les partitions du Leg 1 estimés par l'arbre issu de la classification hiérarchique et la classification experte.

	s_1	s_2	s_3	s_4
$s_1(CH)$	261	4	0	0
$s_2(CH)$	241	0	0	0
$s_3(CH)$	0	245	177	100
$s_4(CH)$	0	0	281	187

Le Leg 2 est ensuite estimé à partir de l'arbre de décision issu du Leg 1 et comparer à la labellisation du Leg 2 par l'expert afin de tester le pouvoir de généralisation. L'analyse des tables de confusion (tableau 5.7) pour $N = 3$ ou 5 selon le découpage expert montre que les groupes $s_1(CH)$ et $s_2(CH)$ correspondent globalement à l'état s_1 de l'expert (présence seulement d'algues brunes et vertes) et l'état $s_4(CH)$ à l'état s_2 de l'expert (présence des quatre types d'algues). Bien que le groupe $s_3(CH)$ définit par vote majoritaire semble correspondre à l'état s_3 (pas d'algues bleu-vert mais présence des 3 autres classes algales), une confusion non négligeable est à relever : 48 données appartiennent au label s_2 (faux positifs).

Tableau 5.7. Table de confusion entre les partitions du Leg 2 estimés par l'arbre issu de la classification hiérarchique et la classification experte.

	s_1	s_2	s_3
$s_1(CH)$	837	1	8
$s_2(CH)$	760	0	0
$s_3(CH)$	0	48	383
$s_4(CH)$	0	335	0

Les mêmes comparaisons sont réalisées entre la partition obtenue à partir du même arbre coupé à $N = 6$ et la labellisation experte en 6 groupes. Les résultats sont identiques avec un indice de Rand égal à 0,76 (l'indice de confiance est de 0,70). La tête de l'arbre restant identique pour 4 ou 6 groupes demandés, l'interprétation est identique : la classification hiérarchique ne répond pas pleinement au découpage expert. Une confusion potentielle existe si on cherche à estimer les états contenant des algues bleu-vert.

Même si l'approche par classification hiérarchique semble naturelle puisqu'elle offre, comme l'expert, un système de décision basé sur un arbre, ces deux systèmes ne répondent pas à la même problématique. La classification hiérarchique par agrégation de Ward est fondée sur un

critère de lien maximum entre groupes équivalent à un critère de lien minimum au sein d'un groupe (théorème de Huygens). Cependant, elle n'offre pas un critère conjoint de bonne séparation (coupe) inter~ et intra-classe contrairement à une technique par classification spectrale.

5.2.3.3. Approche utilisant le système MMC-NS

Les observations centrées-réduites sont segmentées par classification spectrale (figure 5.8). Pour estimer l'état d'une nouvelle observation, un système MMC-NS a été construit à partir des états obtenus par cette classification spectrale (sans réduction utilisant le STFKM puisque le volume des données n'est pas un obstacle calculatoire) et les symboles par quantification de l'ensemble des données.

Notations :

- Leg1A : la segmentation du Leg 1 obtenue par classification spectrale ($N = 6$) ;
- Leg1E1 : la segmentation du Leg 1 estimée à partir du MMC-NS dont les paramètres sont calculés à partir du Leg 1 ($\lambda(N = 6, M = 40)$) ;
- Leg2E1 : la segmentation du Leg 2 estimée à partir du MMC-NS dont les paramètres sont calculés à partir du Leg 1 ($\lambda(N = 6, M = 40)$) ;
- Leg2A : la segmentation du Leg 2 obtenue par classification spectrale ($N = 11$).
- Leg2E2 : la segmentation du Leg 2 estimée à partir du MMC-NS dont les paramètres sont calculés à partir du Leg2 ($\lambda(N = 11, M = 58)$) ;
- Leg1E2 : la segmentation du Leg 2 estimée à partir du MMC-NS dont les paramètres sont calculés à partir du Leg2 ($\lambda(N = 11, M = 58)$).

Résultats obtenus à partir du Leg 1

Comparaison

L'indice de Rand obtenu lors de la comparaison des partitions entre la segmentation estimée par MMC-NS (Leg1E1) et celle par classification spectrale (Leg1A) est de 0,98 et son indice de confiance est de 0,975. Le système MMC-NS est fiable : la seule différence observée est due aux valeurs manquantes (le 20/04/12 entre 17h10 et 18h20) ; les probabilités de transitions lors de l'estimation sont donc perturbées : le système ne prend pas en compte les trous dans la série de données car les observations sont concaténées dans l'approche actuelle.

Cinq expériences de comparaison (tableau 5.8) entre partition sont réalisées pour identifier les capacités du système MMC-NS vis-à-vis de la segmentation à dire d'experts et par classification spectrale uniquement, c'est-à-dire sans prise en compte de la dimension temporelle.

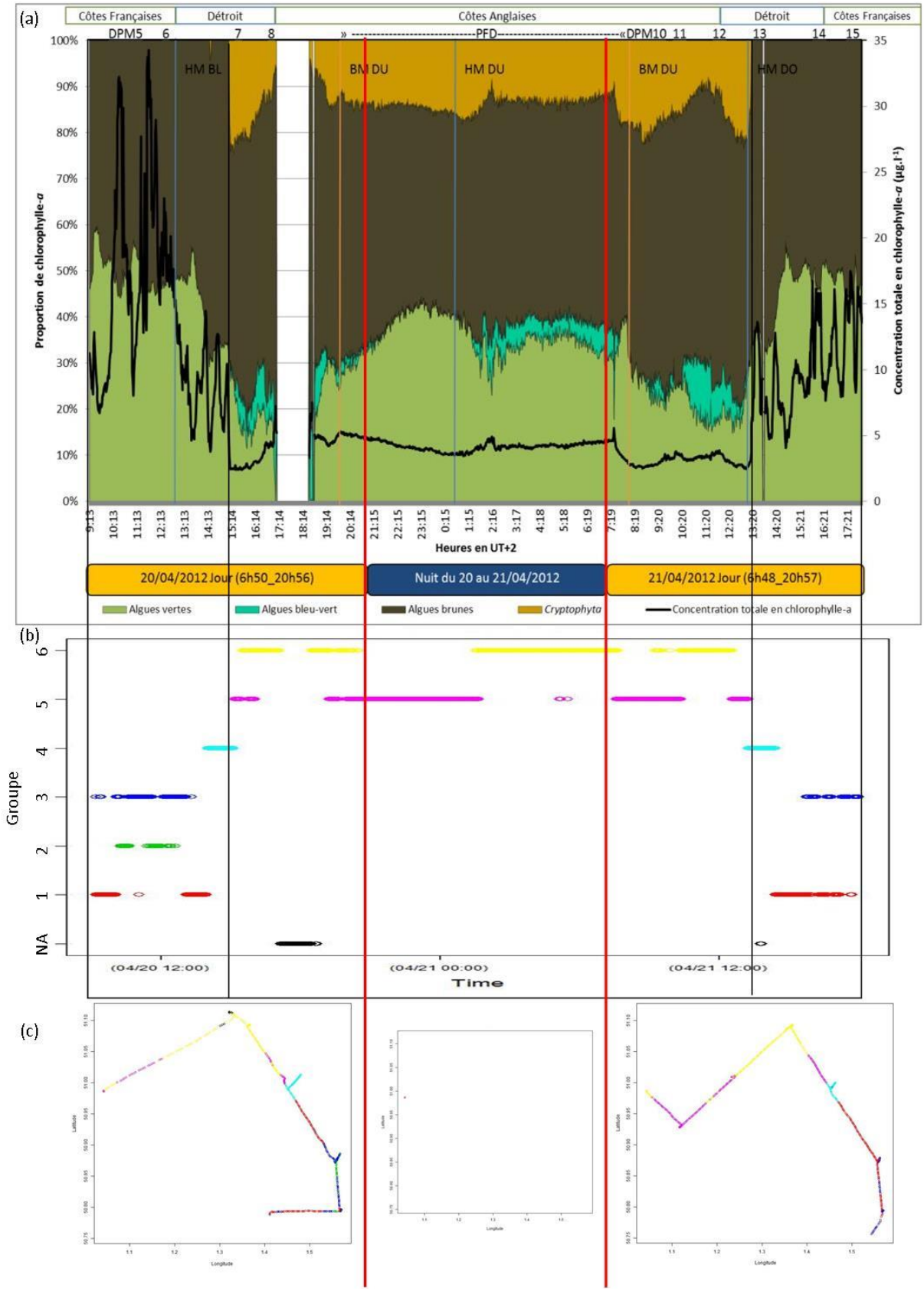


Figure 5.8. Segmentation par classification spectrale du Leg 1 : (a) répartition des différents groupes d'algues phytoplanctoniques avec les empreintes d'algues vertes, bleu-vert, brunes et les cryptophycées, (b) le séquencement des groupes définis par la classification spectrale et (c) la projection de cette segmentation sur le trajet effectué lors de ce Leg 1.

Une comparaison entre les résultats de notre modèle (Leg1A et Leg1E2) et le découpage expert sur les deux nombres d'états possibles ($N = 4$ et $N = 6$) est réalisée afin de tester les capacités du système à estimer une donnée n'ayant pas participé à la construction du modèle. A noter, le nombre d'états déterminé automatiquement par le système sur la base de données issue du Leg2 est de 11. Lors de l'estimation du Leg1, seuls 7 de ces 11 états sont présents dans l'estimation du Leg 1, cela signifie que certains états du Leg2 n'apparaissent pas dans le Leg1. Lorsque la partition Leg1A est comparée à la partition Leg1E2, l'indice de Rand est de 0,91 avec un indice de confiance de 0,87. Ces fortes valeurs confirment qu'une majorité des états apparus lors du Leg1 sont présents lors du Leg2.

Les indices de Rand entre les différentes segmentations du Leg 1 calculées et celles identifiées à dire d'experts sont de l'ordre de 0,76 avec un indice de confiance de l'ordre de 0,7 (tableau 5.8). De même que, pour la classification hiérarchique, les différences peuvent s'expliquer par la nature différente des données, le classifieur partant de l'information brute et non d'une proportion relative calculée à partir des concentrations par classe algale. Ensuite, il est opportun d'analyser conjointement :

- les confusions potentielles de chacun des groupes.
- Les éléments structurants chacun des groupes obtenus.

Tableau 5.8. Comparaison des résultats de classification (Leg1A) et d'estimation (Leg1E2) du système MMC-NS par rapport au découpage à dire d'experts (Leg1Ex) en utilisant l'indice de Rand et son indice de confiance. Une comparaison entre les sorties du système MMC-NS est disponible sur la dernière ligne du tableau.

Segmentations		Indice de Rand	Indice de confiance
Leg1Ex $N = 4$	Leg1A $N = 6$	0,75	0,67
Leg1Ex $N = 4$	Leg1E2 $N = 7$	0,77	0,70
Leg1Ex $N = 6$	Leg1A $N = 6$	0,75	0,67
Leg1Ex $N = 6$	Leg1E2 $N = 7$	0,77	0,70
Leg1E2 $N = 7$	Leg1A $N = 6$	0,91	0,87

Caractérisation des groupes obtenus

Apprentissage du Leg1

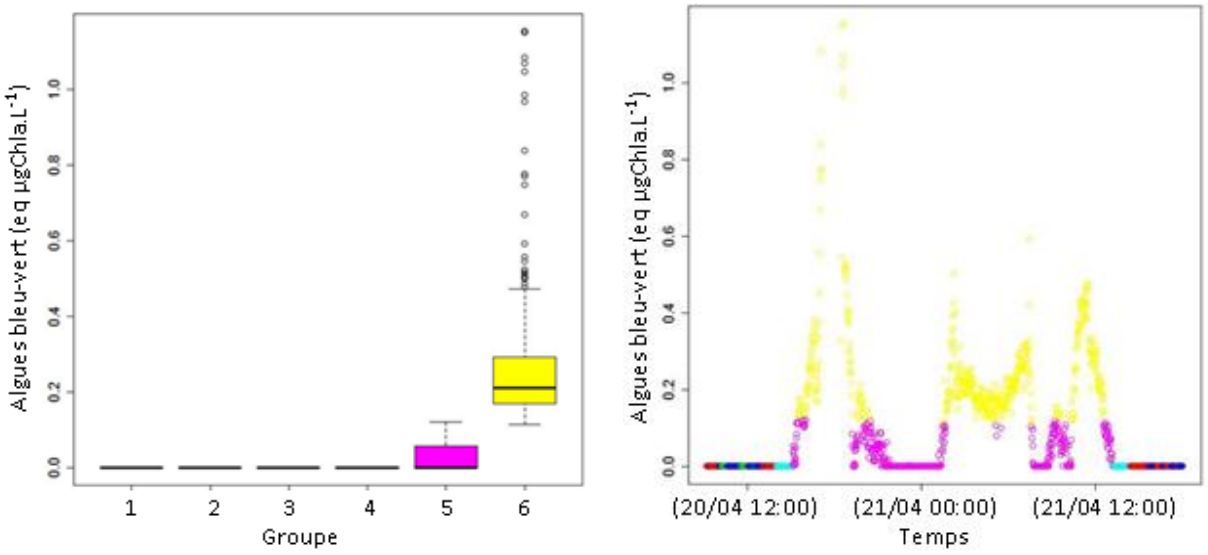
Le nombre de groupes, déterminé par la méthode du gap entre valeurs propres, est égal 6 ($N = 6$). En ne s'intéressant qu'aux proportions des groupes algaux spectraux, il résulte la correspondance suivante (tableau 5.9) :

- $s_5(Leg1A)$ et $s_6(Leg1A)$: présence de l'ensemble des classes algales (en bleu dans le tableau 5.9), structurés selon le niveau de concentrations des algues bleu-vert (seuil net sur la figure 5.9, $s_6(Leg1A)$ marqué par un niveau supérieur à 0,1 eq µgChla.L^{-1}) ;
- $s_4(Leg1A)$: présence de cryptophycées, d'algues brunes et vertes (en vert dans le tableau 5.9), ce groupe correspond à une période de transition, c'est-à-dire à l'apparition ou la disparition de certaines algues (couleur cyan sur la figure 5.10) ;
- $s_1(Leg1A)$, $s_2(Leg1A)$ et $s_3(Leg1A)$: présence d'algues brunes et vertes uniquement (en orange dans le tableau 5.9).

Tableau 5.9. Proportions relatives (%) des quatre groupes algaux réparties dans les N=6 groupes définis lors de la classification spectrale des données du Leg 1. Trois groupes de type expert sont mis en valeur grâce à un jeu de couleur : l'orange, le vert et le bleu.

	Vert	Bleu-vert	Brunes	Cryptophycées
$s_1(Leg1A)$	50,49	0,00	49,51	0,00
$s_2(Leg1A)$	46,27	0,00	53,73	0,00
$s_3(Leg1A)$	48,66	0,00	51,34	0,00
$s_4(Leg1A)$	34,67	0,00	65,01	0,33
$s_5(Leg1A)$	32,85	0,84	49,93	16,38
$s_6(Leg1A)$	26,99	6,26	53,92	12,83

L'origine du découpage de la classe s_1 (présence d'algues brunes et vertes uniquement) en trois sous-groupes $s_1(Leg1A)$, $s_2(Leg1A)$ et $s_3(Leg1A)$ s'explique par les niveaux des concentrations de chacune de ces algues. Ainsi, en comparant leur moyenne respective, les plus fortes concentrations sont présentes dans le groupe s_2 (12,08 µgChla.L^{-1} pour les algues vertes et 14,12 µgChla.L^{-1} pour les algues brunes), vient ensuite le groupe s_3 (7,37 µgChla.L^{-1} pour les algues vertes et 7,81 µgChla.L^{-1} pour les algues brunes), et pour finir les plus faibles concentrations appartiennent au groupe s_1 (4,25 µgChla.L^{-1} pour les algues vertes et 4,21 µgChla.L^{-1} pour les algues brunes) (figure 5.11).



(a) (b)

Figure 5.9. (a) Boîte de dispersion des algues bleu-vert pour chaque groupe obtenu par classification spectral pour le Leg 1 et (b) la projection du résultat de classification sur l'évolution temporelle de la concentration des algues bleu-vert.

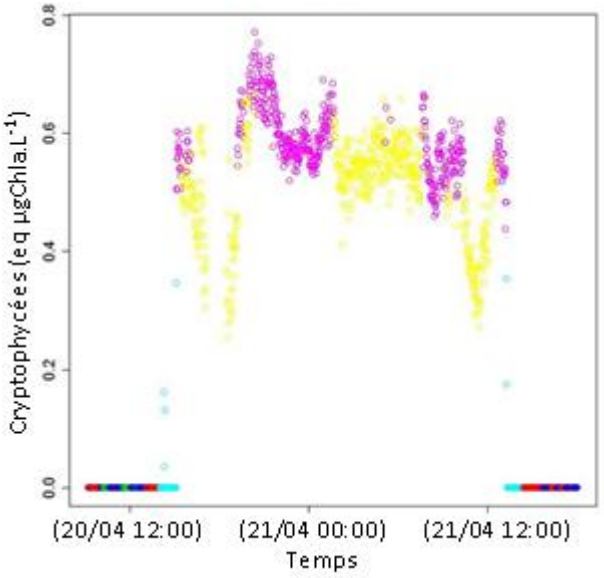


Figure 5.10. Projection du résultat de classification spectrale sur l'évolution temporelle de la concentration des cryptophycées.

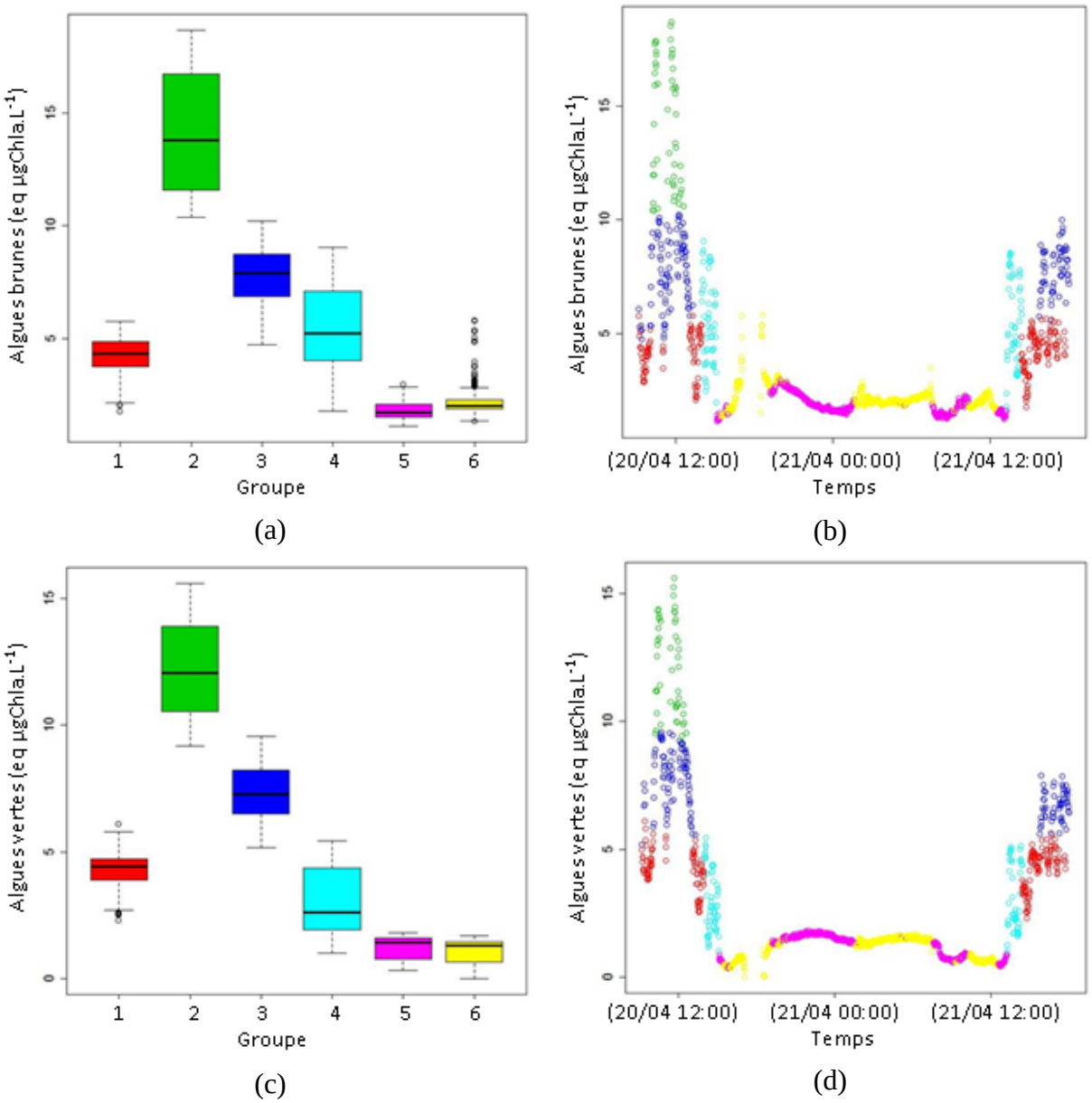


Figure 5.11. Boîtes de dispersion de (a) la concentration en algues brunes (eq µgChla.L^{-1}) et (c) en algues vertes (eq µgChla.L^{-1}) pour chacun des 6-états obtenus après classification spectrale des données du Leg 1, ainsi que la projection de ces groupes sur l'évolution temporelle de ces deux paramètres (b) et (d).

Estimation des états des données du Leg 2 à partir du Leg 1 appris

Les données du Leg 2 sont estimés par programmation dynamique en utilisant le système MMC-NS construit à partir du Leg 1. L'algorithme de Viterbi a permis de déterminer le chemin optimal entre les états et a estimé que ces données appartiennent à 5 des 6 états déterminés sur les données du Leg 1. Il faut prendre en compte que lors de la traversée de la Manche (le 28/04/12 entre 14h50 et 19h30), les données n'ont pas été enregistrées ce qui peut provoquer quelques erreurs d'estimations.

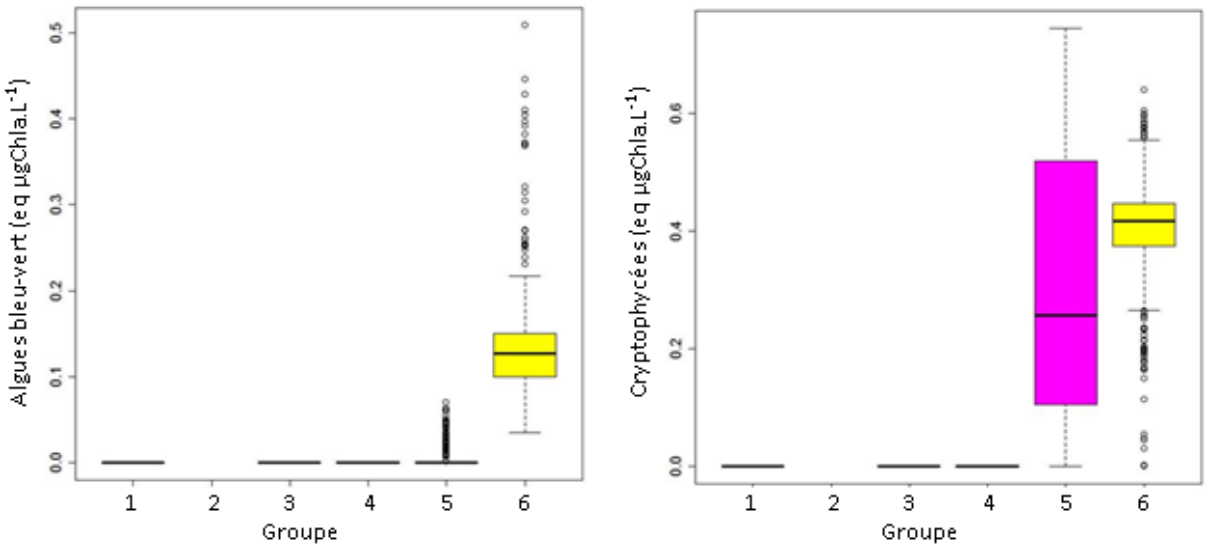
Tableau 5.10. Proportions relatives (%) des quatre groupes algaux réparties dans les N=5 groupes définis sur les données du Leg 2 lors de leur estimation à partir du système MMC-NS. Trois groupes de type expert sont mis en valeur grâce à un jeu de couleur : l'orange, le vert et le bleu.

Groupe	Algues vertes	Algues bleu-vert	Algues brunes	Cryptophycées
s_1	38,40	0,00	61,60	0,00
s_2	-	-	-	-
s_3	43,20	0,00	56,80	0,00
s_4	34,59	0,00	65,41	0,00
s_5	42,86	0,12	45,97	11,05
s_6	35,18	8,15	32,42	24,25

Le groupe s_2, qui apparait de manière fugace lors du Leg 1, n'est pas identifié par le modèle pour le Leg2. Comme précédemment, lorsque l'ensemble des quatre classes algales sont présentes (couleur bleu dans le tableau 5.10), deux groupes se distinguent où la différence est visible au niveau des concentrations en algues bleu-vert (valeur négligeable pour le groupe s_5, et 0,14 µgChla.L^{-1} pour le groupe s_6) et en cryptophycées (0,31 µgChla.L^{-1} pour le groupe s_5, et 0,40 µgChla.L^{-1} pour le groupe s_6) (figure 5.12).

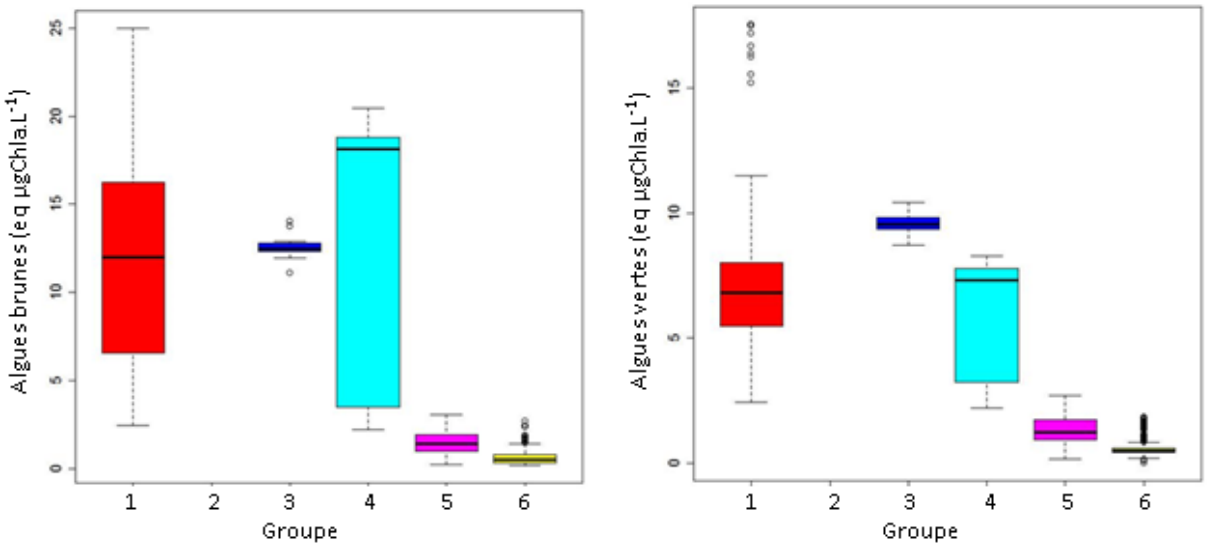
Les groupes $s_1$et s_3 se différencient principalement par la différence de concentration des algues vertes (6,74 µgChla.L^{-1} pour le groupe s_1, et 9,57 µgChla.L^{-1} pour le groupe s_3)

Le groupe s_4, contrairement à la classification des données du Leg 1, ne contient pas de cryptophycées. Cela n'est pas une erreur d'estimation, puisque celui-ci représentait une étape de transition possédant une certaine concentration en algues brunes et vertes (figure 5.13).



(a) (b)

Figure 5.12. Boîtes de dispersion des algues bleu-vert (a) et des cryptophycées (b) pour chaque groupe obtenu après leur estimation par le système MMC-NS.



(a) (b)

Figure 5.13. Boîtes de dispersion des algues brunes (a) et des algues vertes (b) pour chaque groupe obtenu après leur estimation par le système MMC-NS.

Résultats obtenus à partir des données du Leg2

Comparaison

Les données du Leg 2 sont maintenant classées par classification spectrale (LEg2A) et estimées via le MMC-NS construit à partir du Leg 1. Le nombre de groupes a été déterminé automatiquement à $N = 11$. Lorsque ces données sont estimées à partir de la classification du Leg 1, le nombre de groupes est de $N = 5$ et la segmentation est notée Leg2E1 (voir paragraphe ci-dessus). Le découpage expert sur les données du Leg 2 est noté Leg2Ex avec $N = 3$ et $N = 5$. Lorsque ces différentes partitions sont comparées entre elles (tableau 5.11), celles possédant un nombre N proche sont similaires, l'indice de Rand étant sensible au nombre de groupes. C'est pourquoi, le Leg2A n'obtient pas d'indice de Rand élevé avec une autre partition.

Caractérisation des groupes obtenus.

Apprentissage du Leg 2

Le même jeu de couleur que précédemment permet de simplifier les 11 groupes obtenus afin de ressembler au dire d'experts (tableau 5.12).

- Présence uniquement d'algues brunes et vertes (couleur orange dans le tableau 5.12) : la différenciation entre ces groupes (s_1, s_4 à s_6 et s_8 à s_{11}) se fait selon les niveaux de concentrations de ces deux groupes spectraux ;
- Présence de cryptophycées, d'algues brunes et vertes (couleur verte dans le tableau 5.12) : la présence de ces trois groupes simultanément n'apparait qu'ici, c'est pourquoi un seul groupe y est dédié : s_2 ;

- Présence de l'intégralité des groupes spectraux (couleur bleue dans le tableau 5.12) : la concentration des algues bleu-vert permet d'obtenir la différence entre les groupes s_3 et s_7.

Tableau 5.11. Comparaison en utilisant l'indice de Rand et son indice de confiance des résultats de classification (Leg2A) et d'estimation (Leg1E2) du système hybridé par rapport au découpage expert (Leg2Ex). Une comparaison entre les sorties de l'hybride est disponible sur la dernière ligne du tableau.

Segmentations		Indice de Rand	Indice de confiance
Leg2Ex $N = 3$	Leg2A $N = 11$	0,58	0,585
Leg2Ex $N = 3$	Leg2E1 $N = 5$	0,82	0,815
Leg2Ex $N = 5$	Leg2A $N = 11$	0,58	0,575
Leg2Ex $N = 5$	Leg2E1 $N = 5$	0,81	0,805
Leg2E1 $N = 5$	Leg2A $N = 11$	0,81	0,63

Tableau 5.12. Proportions relatives (%) des quatre groupes algaux répartis dans les N=11 groupes définis lors de la classification spectrale des données du Leg 2. Trois groupes de type expert sont mis en valeur grâce à un jeu de couleur : l'orange, le vert et le bleu.

Groupe	Algues vertes	Algues bleu-vert	Algues brunes	Cryptophycées
s_1	41,49	0,00	58,51	0,00
s_2	53,76	0,00	40,50	5,74
s_3	35,67	8,48	31,39	24,47
s_4	29,15	0,00	70,85	0,00
s_5	29,97	0,00	70,03	0,00
s_6	37,58	0,00	62,42	0,00
s_7	29,67	0,42	50,61	19,30
s_8	28,12	0,00	71,88	0,00
s_9	45,51	0,00	54,49	0,00
s_{10}	47,82	0,00	52,18	0,00
s_{11}	31,53	0,00	68,47	0,00

Estimation des états des données du Leg 1 à partir du Leg 2 appris

Les données du Leg 1 sont estimées en utilisant le MMC-NS construit à partir du Leg 2. Le nombre de groupe estimé est de $N = 7$, cela signifie que certains états n'apparaissent pas lors du Leg 1. La simplification en découpage à dire d'experts est retrouvée (tableau 5.13) :

- Présence uniquement d'algues brunes et vertes (couleur orange dans le tableau 5.13) : s_1, s_4 à s_6 et s_8 à s_{11} ;
- Présence de cryptophycées, d'algues brunes et vertes (couleur verte dans le tableau 5.13) : s_2 ;
- Présence de l'intégralité des groupes spectraux (couleur bleue dans le tableau 5.13) : s_3 et s_7.

Tableau 5.13. Proportions relatives (%) des quatre groupes algaux répartis dans les N=7 groupes définis sur les données du Leg 1 lors de leur estimation à partir du système MMC-NS. Trois groupes de type expert sont mis en valeur grâce à un jeu de couleur : l'orange, le vert et le bleu.

Groupe	Algues vertes	Algues bleu-vert	Algues brunes	Cryptophycées
s_1	48,34	0,00	51,66	0,00
s_2	31,77	0,00	65,64	2,59
s_3	26,61	5,81	54,10	13,47
s_4	NA	NA	NA	NA
s_5	NA	NA	NA	NA
s_6	45,25	0,00	54,75	0,00
s_7	34,63	0,41	48,88	16,08
s_8	NA	NA	NA	NA
s_9	47,55	0,00	52,45	0,00
s_{10}	42,27	0,00	57,73	0,00
s_{11}	NA	NA	NA	NA

5.2.3.4. Conclusion

La classification spectrale est capable d'identifier des états qui représentent des communautés phytoplanctoniques différentes puisque caractérisées par des signatures différentes des quatre classes algales considéré par le fluorimètre spectral. Au sein d'un même état, il est alors possible de différencier, à petite échelle de temps et donc d'espace, des ruptures liées à des changements de concentration pour une classe algale donnée (algues brunes, bleu-vert, vertes, cryptophycées). Cette structuration de l'écosystème à de petites échelles de temps et d'espace est un élément important à considérer lorsqu'il s'agit d'envisager d'améliorer les connaissances sur le déterminisme des efflorescences phytoplanctoniques. Comme cela a été présenté dans le chapitre 1, la dynamique d'une efflorescence ne peut s'expliquer que par des interactions et / ou via le contrôle de processus qui se produisent à différentes échelles de temps et d'espace. Alors que les approches couramment utilisées à ce jour pour mieux

comprendre l'écologie du phytoplancton intègrent régulièrement les phénomènes à grandes échelles, via par exemple, les indices climatiques (Breton *et al.*, 2006 ; Goberville *et al.*, 2010 ; Lefebvre *et al.*, 2011), les phénomènes à plus petites échelles ne sont que trop rarement appréhendés du fait du manque d'outils numériques disponibles pour optimiser le traitement et l'interprétation des données issues des systèmes de mesures à haute résolution.
L'approche d'identification d'états particuliers au sens de la composition phytoplanctonique des masses d'eau par MMC-NS semble plus cohérente que celle basée sur un arbre de décision issue d'une classification hiérarchique. En effet, en agrégeant les N groupes distingués par le MMC-NS pour obtenir un nombre d'états identiques fixés par l'expert, les partitions proposées sont cohérentes à celles du dire d'experts. De part sa capacité à prendre en compte le paramètre temporel des données via ses matrices probabilistes, elle a montré sa capacité à estimer l'état environnemental de nouvelles données collectées sur des zones géographiques différentes.
Par ailleurs cette modélisation permet de raffiner le découpage initial de l'expert en un nombre plus important que celui-ci et notamment de permettre l'exploitation réelle de la dimension haute fréquence des données.

5.3. Haute résolution temporelle en milieu continental

La recherche d'états particuliers en milieu marin par MMC-NS a été validée sur des données acquises en point fixe ou à haute résolution spatiale (< 200 m) et à haute résolution temporelle (de 1 à 20 minutes en fonction de l'étude considérée). Il s'agit maintenant d'étudier comment réagit ce type de modèle pour des données acquises sur une période de 1 mois avec une fréquence de 10 minutes, en milieu continental lors d'une étude mise en œuvre par l'Agence de l'Eau Artois Picardie (AEAP). Après validation des performances de la classification spectrale et / ou du système MMC-NS pour ces différents instruments et écosystèmes, nous pourrons ainsi conclure quant à l'aspect opérationnel du système hybridé proposé.

5.3.1. Objectif de l'étude

L'objectif général de cette étude s'inscrit dans les réflexions visant à la mise en place d'un programme pérenne d'observation et de surveillance du phytoplancton en eau douce en insistant sur la valeur ajoutée des mesures à haute résolution. Un objectif secondaire est de vérifier que le modèle proposé est capable de détecter un état environnemental particulier, correspondant à des perturbations qui seraient engendrées par la navigation sur ce cours d'eau, comme la remise en suspension d'algues toxiques (cyanophycées, responsables de la dégradation de la qualité des eaux de baignades et de boissons), mais également des nutriments favorisant ainsi la production de phytoplancton à petite échelle (modification de la biomasse et la nature des algues). La question sous-jacente est : à terme, la navigation étant sensée augmenter, la qualité des eaux va-t-elle se dégrader ou s'améliorer?

5.3.2. Le jeu de données

5.3.2.1. Source des données et paramètres étudiés

Le jeu de données a été fourni par l'Agence de l'Eau Artois Picardie. Les mesures ont été effectuées sur la Deûle du 15 avril 2009 au 15 mai 2009 sur le site AGORA SITA à Courcelles-les-Lens dans le Pas-de Calais à partir d'un laboratoire mobile de l'Agence de l'Eau et d'un fluoroprobe bbe ©. La fréquence d'échantillonnage est de 10 minutes et les paramètres mesurés sont : la concentration en chlorophycées (µg.L^{-1}), en diatomées (µg.L^{-1}), en cyanophycées (µg.L^{-1}), en cryptophycées (µg.L^{-1}), en chlorophylle totale (µg.L^{-1}), en Carbone Organique Total (mg.L^{-1}), en phosphate (PO_4^{3-} en µmol.L^{-1}), en nitrate (NO_3^- en µmol.L^{-1}), en azote ammoniacal (NH_4^+ en µmol.L^{-1}), en oxygène dissous (O_2 en mg.L^{-1}), la conductivité (mS.cm^{-1}), le pH (UpH), la température (°C), la turbidité (NTU), l'irradiance (lux) et la pluviométrie (mm).

Il faut noter que les prélèvements ont été réalisés au moment de la fin d'une efflorescence de diatomées.

5.3.2.2. Prétraitement des données

La base de données transmise par l'AEAP a été considérée comme validée. Aucun ajustement des données n'a été fait au regard d'une gamme de valeurs définie à dire d'experts. Aucune complétion n'a été mise en œuvre. Afin de s'affranchir des différences d'unités entre les variables, les données ont été centrées-réduites.

Il est à noter qu'il n'y a pas d'effet quenching (Chapitre 4), c'est-à-dire qu'il n'y a pas de diminution de la fluorescence causée par une régulation au niveau cellulaire de la photosynthèse provoquée par une exposition à des intensités lumineuses trop importantes.

Par ailleurs, il faut noter que l'irradiance n'est enregistrée qu'à partir de 2000 lux pour limiter la composante « bruit » du signal en raison de la présence de sources de lumières parasites (phares, lampadaire).

5.3.3. Méthodologie

Une classification spectrale est réalisée sur les données centrées-réduites et le nombre de groupes N est calculé automatiquement avec la méthode du gap (Chapitre 3).

Le paramètre « pluviométrie » n'a pas été intégré à l'étude. En effet, l'auget du pluviomètre bascule pour un volume précipité équivalent à 0,5 mm de pluie. Le fichier de mesures, au pas de temps de 10 minutes, comptabilise le nombre de basculements sur les 10 minutes et par conséquent le nombre de millimètres précipités sur cet intervalle. Les valeurs de ce paramètre n'excèdent jamais 1,5 mm, et sa valeur principale est égale à zéro. Ce paramètre est trop structurant et biaise la classification.

Le nombre de groupes N est calculé automatiquement et est égal à 4. Dans les figures ci-après et pour aider l'interprétation, un groupe nommé « groupe NA » a été ajouté. Celui-ci regroupe

l'ensemble des données non classées à cause d'une valeur manquante sur un des paramètres à un instant t.

5.3.4. Structuration des groupes identifiés par la classification spectrale

Afin de définir les paramètres structurants les groupes mis en évidence par la classification spectrale, une analyse en composante principale (ACP) est réalisée sur les données (matrice des corrélations). Afin de situer les groupes par rapport aux quinze variables actives, sans que ceux-ci interférent avec l'analyse, ces groupes (appelés également clusters sur les figures de l'ACP) déterminés à partir de la classification spectrale sont ajoutés en variables supplémentaires.

Le pourcentage de variation expliquées par les composantes principales 1 et 2 est, respectivement, de 33,89 et 20,06 % (53,95 % au total pour ces deux axes). L'ajout de la dimension 3 permet d'atteindre une contribution à la variation expliquée de 64,05 %.

Du point de vue de la structure des variables, le cercle de corrélation permet de mettre en évidence une bonne représentation des variables de concentration en ammonium (notée NH4), en phosphate (PO4), en oxygène dissous (O2), en diatomées, en chlorophycées, en cryptophycées et en chlorophylle totale (Total.chloro), conductivité (COND), concentration en carbone organique totale (COT) et température (TEMP) dans le plan principal (Dimensions 1 et 2) (figure 5.14). La variable turbidité est bien représentée dans le plan secondaire (Dimensions 1 et 3) (figure 5.15). L'axe 1 représente ainsi le développement de la biomasse végétale et une production d'oxygène en réponse à une consommation de nutriments, alors que l'axe 2 représente plutôt des apports de matière organique d'origine allochtone* et/ou autochtone*. L'axe 3 est fortement structuré par la turbidité (figure 5.15).

Pour définir les variables structurantes pour chaque groupe identifié grâce à la classification spectrale, une ACP a été effectuée sur chacun des groupes et pour le base de données à 15 variables (sans la pluviométrie) (tableaux 5.14 et 5.15). L'analyse des coefficients de corrélation au carré permet de hiérarchiser la contribution des variables à un axe.

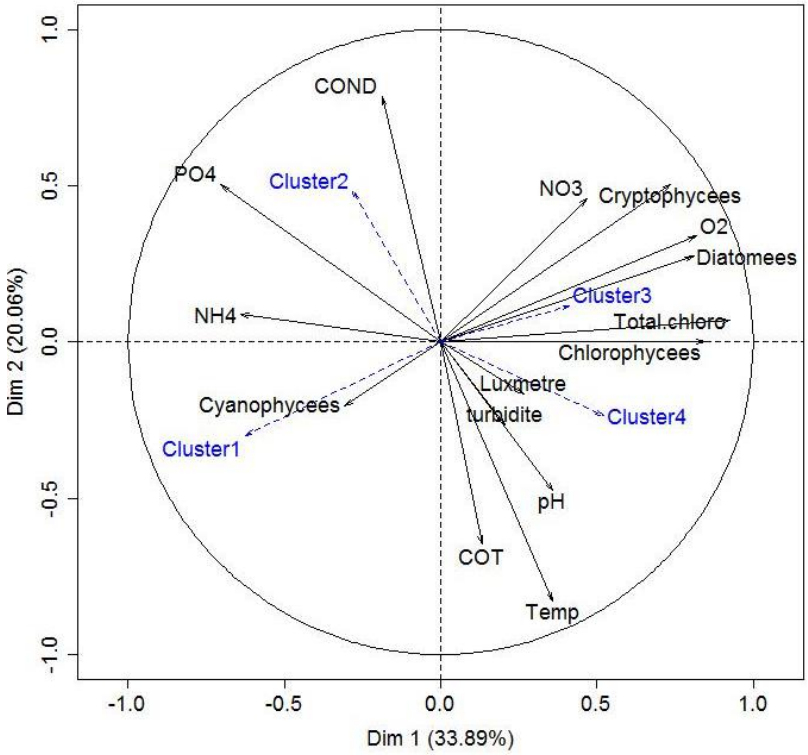


Figure 5.14. Projection sur les deux premières dimensions de l'ACP des paramètres mesurés par la station instrumentée sur la Deûle au printemps 2009.

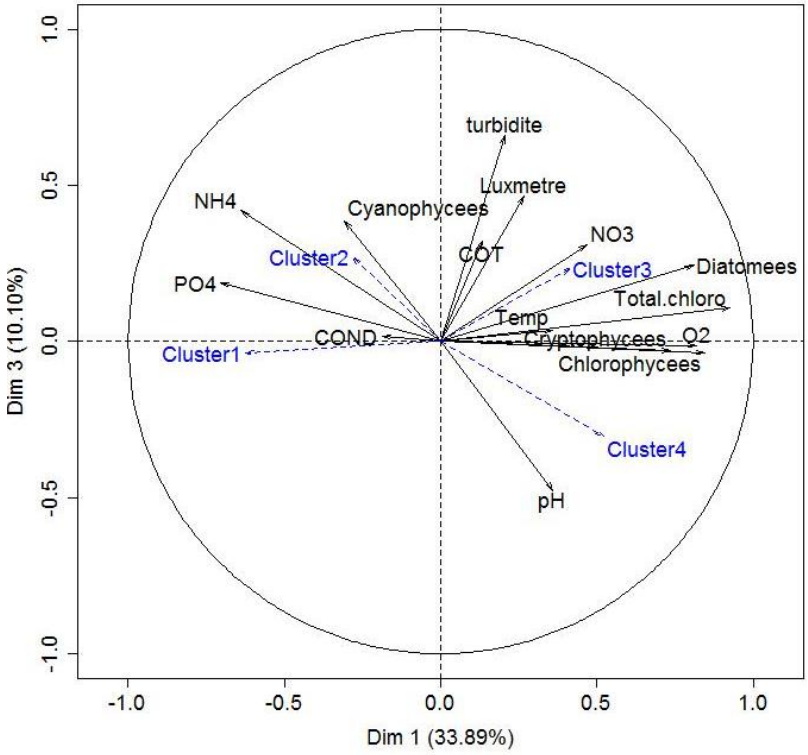


Figure 5.15. Projection sur la première et la troisième dimension de l'ACP des paramètres mesurés par la station instrumentée sur la Deûle au printemps 2009.

Tableau 5.14. Résultats des ACP par groupe : Contribution sur la dimension 1 de chaque paramètre mesuré sur la Deûle au printemps 2009 pour chaque groupe déterminé par classification spectrale.

Dimension 1	groupe 1	groupe 2	groupe 3	groupe 4
Chlorophycées	0,63	0,83	0,29	0,51
Cyanophycées	0,06	0,31	0,16	0,32
Diatomées	0,30	0,16	0,51	0,72
Cryptophycées	0,65	0,25	0,03	0,67
Chlorophylle totale	0,64	0,50	0,65	0,73
Carbone organique totale	0,11	0,28	0,57	0,04
Phosphate	0,24	0,03	0,77	0,00
Azote Ammoniacal	0,40	0,29	0,47	0,00
Nitrate	0,01	0,32	0,57	0,15
Conductivité	0,19	0,01	0,33	0,00
Oxygène dissous	0,82	0,54	0,84	0,62
pH	0,63	0,08	0,30	0,06
Température	0,18	0,45	0,90	0,00
Turbidité	0,00	0,06	0,00	0,00
Irradiance	0,01	0,22	0,20	0,02

Tableau 5.15. Résultats des ACP par groupe : Contribution sur la dimension 1 de chaque paramètre mesuré sur la Deûle au printemps 2009 pour chaque groupe déterminé par classification spectrale.

Dimension 2	groupe 1	groupe 2	groupe 3	groupe 4
Chlorophycées	0,11	0,00	0,04	0,23
Cyanophycées	0,16	0,15	0,50	0,01
Diatomées	0,50	0,00	0,37	0,04
Cryptophycées	0,03	0,30	0,66	0,05
Chlorophylle totale	0,31	0,01	0,19	0,05
Carbone organique totale	0,17	0,45	0,07	0,46
Phosphate	0,01	0,70	0,07	0,69
Azote Ammoniacal	0,12	0,01	0,25	0,07
Nitrate	0,00	0,08	0,01	0,09
Conductivité	0,05	0,57	0,07	0,66
Oxygène dissous	0,01	0,14	0,09	0,01
pH	0,00	0,00	0,31	0,16
Température	0,30	0,41	0,02	0,84
Turbidité	0,57	0,01	0,21	0,05
Irradiance	0,43	0,01	0,21	0,10

Par ailleurs, après avoir identifiées les variables structurantes pour chaque groupe, il est utile de définir les paramètres statistiques de base afin de pouvoir identifier plus clairement des états environnementaux caractéristiques de pressions (forte concentration en nutriments, par exemple), d'état (niveau de turbidité ou d'irradiance) ou d'impact (effets directs et indirects du développement de biomasse comme, par exemple, une modification de la concentration en oxygène). Ces états environnementaux vont inévitablement se succéder dans le temps (figure 5.16) et l'analyse du séquencement de ces groupes permettra de définir la dynamique du système afin d'en comprendre le déterminisme et d'envisager des actions de gestion si l'un des groupes identifiés s'avérait être incompatible avec une bonne qualité de l'environnement étudié.

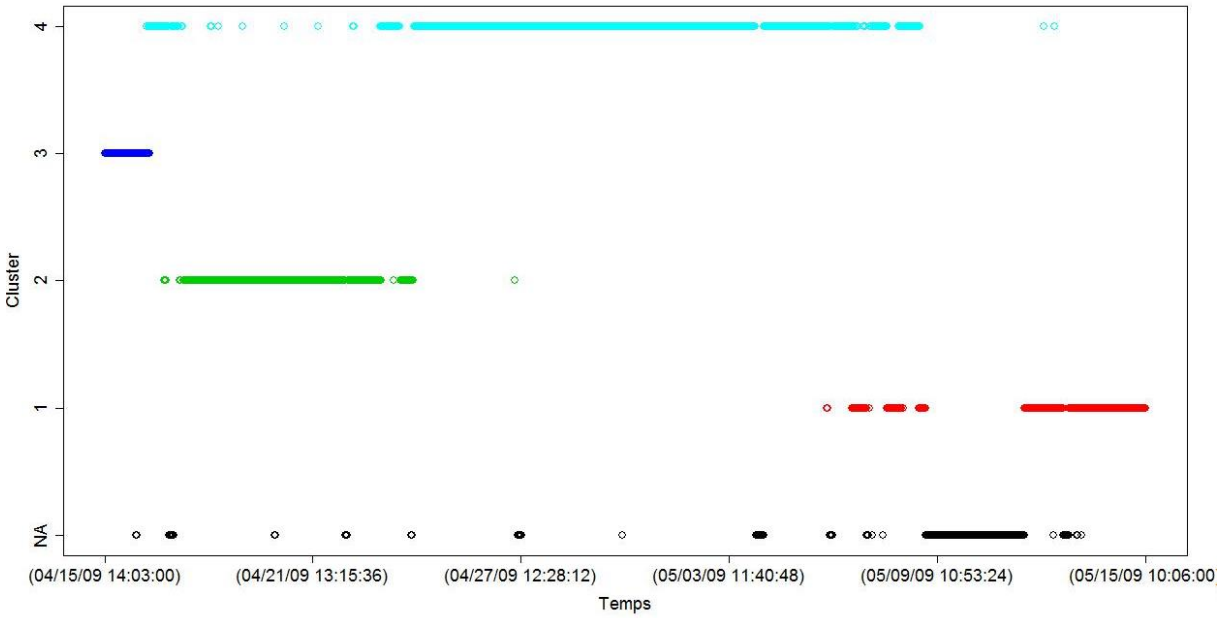


Figure 5.16. Séquencement des groupes définis par la classification spectrale des données mesurée sur la Deûle au printemps 2009. Le cluster NA concerne les mesures dont les états ne sont pas estimés (au moins une donnée manquante à un instant t).

Groupe NA

Lorsqu'un paramètre possède une valeur manquante, l'instant où se trouve cette valeur n'est pas classé pour l'ensemble des paramètres.

Groupe 1

Ce groupe présente les plus faibles concentrations en chlorophylle totale, en chlorophycées, en cryptophycées, en diatomées, en oxygène dissous et en nitrate (figures 5.17 (a) à (g)). Ce groupe se situe uniquement en fin de série (figure 5.16). Il correspond à une phase de fin d'efflorescence après épuisement du stock de nutriments, une disponibilité de lumière moindre (figure 5.17 (h)), mais sans contrôle majeur par la turbidité qui reste d'un niveau comparable à celle observée pour les autres groupes (figure 5.17 (i)). Le niveau de biomasse plus faible est tel que la production d'oxygène est réduite, sans toutefois atteindre des niveaux

susceptibles d'impacter le fonctionnement de l'écosystème. La forte variabilité de la concentration en ammonium dans ce groupe peut s'expliquer par des processus de transformation de la matière particulaire issue de l'efflorescence (groupe 3) en matière dissoute.

Groupe 2

Ce groupe présente des valeurs de température et de pH particulièrement faibles (figures 5.17). Temporellement, il fait suite au groupe 3 (figure 5.16), qui caractérise une efflorescence de diatomées. Les concentrations en nutriments ainsi que les concentrations en cryptophycées, en cyanophycées et en chlorophycées sont intermédiaires (figures 5.17 (c), (l) et (b)) ; l'efflorescence précédente de diatomées n'a pas conduit à un épuisement du stock de nutriments (figures (f), (g) et (o)). D'autres taxons peuvent ainsi se développer.

Groupe 3

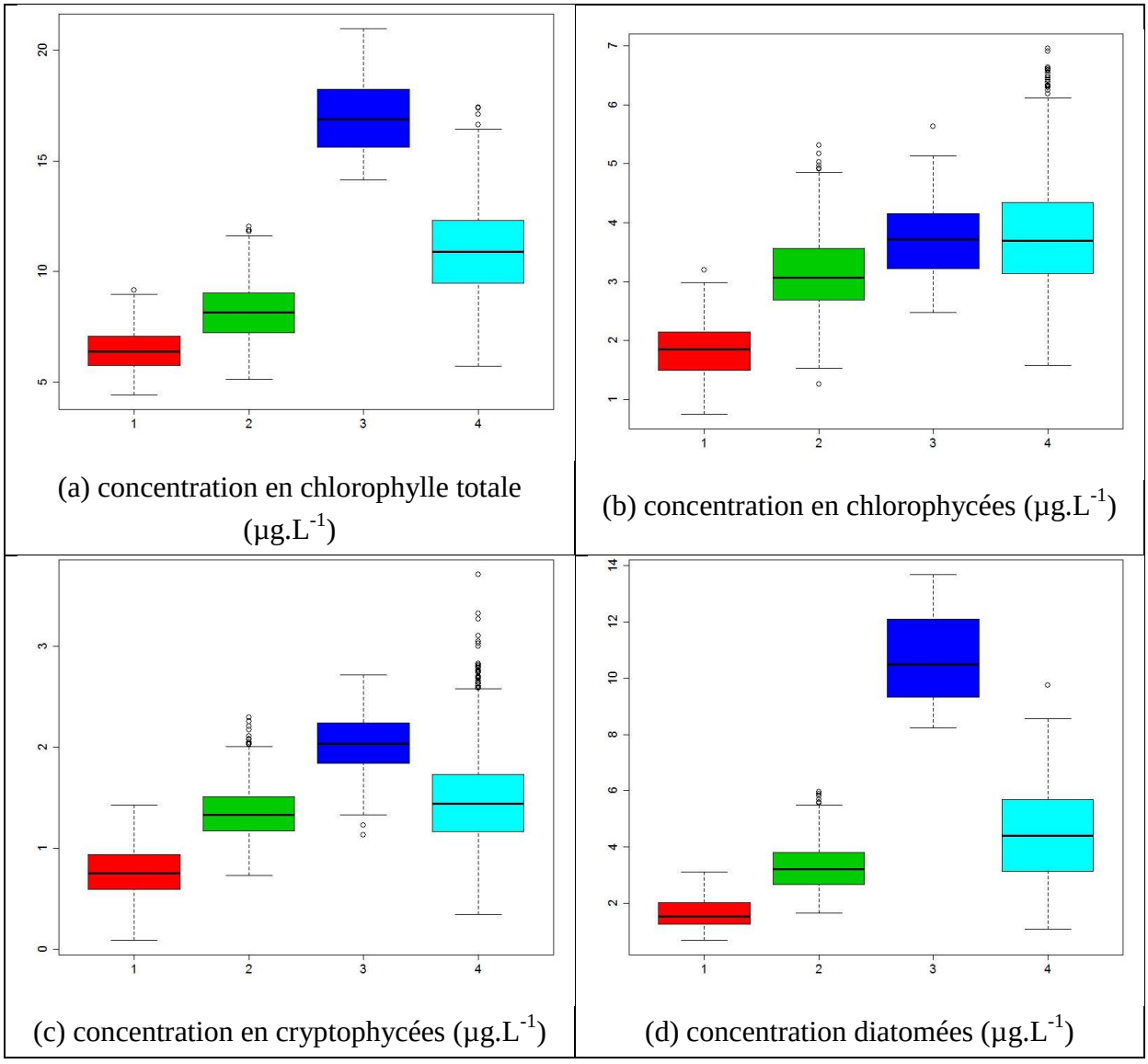
Les éléments de ce groupe sont caractérisés par de fortes concentrations en chlorophylle totale, en diatomées, en NO_3^- et en oxygène (figures 5.17 (a), (d), (g) et (e)). La production de biomasse végétale est dominée par les diatomées, ce qui induit une forte production d'oxygène. Ce développement est favorisé par des apports en azote non limitant. En effet, on aurait pu s'attendre à trouver des concentrations en NO_3^- plus importante dans un autre groupe (qui en terme de séquencement apparaîtrait avant ce groupe 3) en raison d'un décalage temporel entre la consommation du stock de nutriments présent lors d'une période P et la production de biomasse qui en résulte à la période suivant P + x (x pouvant être de l'ordre de quelques heures à quelques jours).

Le groupe 3 se situe uniquement en début de série (figure 5.16) et la projection de la classification sur les données ainsi que l'expertise des chercheurs de l'Agence de l'Eau valide le fait que ce groupe correspond à un évènement de type efflorescence (maximum de chlorophylle totale égale à 20,99 $\mu g.L^{-1}$) (figure 5.17 (a)).

Groupe 4

Ce groupe est marqué par un certain nombre de valeurs extrêmes hautes pour les paramètres de concentration en chlorophycées, en cryptophycées, en cyanophycées, en turbidité, en COT, en température, ainsi que par des valeurs particulièrement faibles (outliers) pour la conductivité, le pH (figures 5.17 (a), (b), (l), (i), (m), (j), (n) et (k)). Ce groupe semble intégrer les épisodes de remises en suspension des sédiments lors du panache des péniches. Il apparaît lors de ces épisodes, des augmentations de la concentration de certains groupes phytoplanctoniques, avec notamment les cyanophycées (figure 5.17 (l)), qui peuvent impacter négativement la qualité de l'eau. Il est à noter que si les remises en suspension dues aux passages des péniches étaient uniquement détectées, un groupe devrait apparaitre le jour et disparaître la nuit.

Figure 5.17. Boîte de dispersion de (a) la concentration en chlorophylle totale (µg.L^{-1}), (b) en chlorophycées (µg.L^{-1}), (c) en cryptophycées (µg.L^{-1}), (d) en diatomées (µg.L^{-1}), (e) en oxygène dissous (O_2 en mg.L^{-1}), (f) en azote ammoniacal (NH_4^+ en µmol.L^{-1}), (g) en nitrate (NO_3^- en µmol.L^{-1}), (h) l'irradiance (lux), (i) la turbidité (NTU), (j) la température (°C), (k) le pH (UpH), (l) la concentration en cyanophycées (µg.L^{-1}), (m) en Carbone Organique Total (mg.L^{-1}), (n) la conductivité (mS.cm^{-1}) ainsi que (o) la concentration en phosphate (PO_4^{3-} en µmol.L^{-1}) mesurée sur la Deûle au printemps 2009 pour chacun des 4-états obtenus après classification spectrale des données.



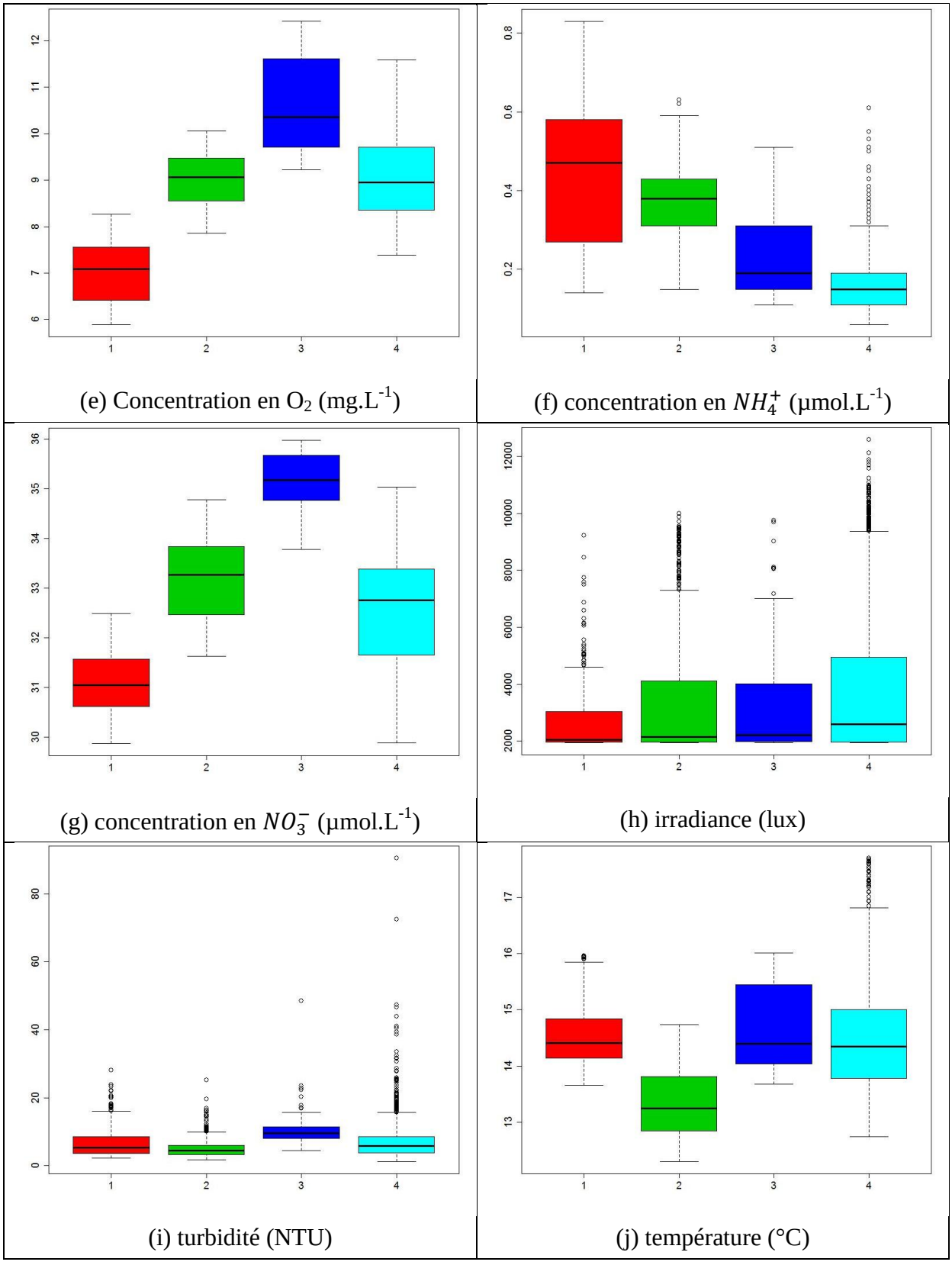
(a) concentration en chlorophylle totale (µg.L^{-1})

(b) concentration en chlorophycées (µg.L^{-1})

(c) concentration en cryptophycées (µg.L^{-1})

(d) concentration diatomées (µg.L^{-1})

Figure 5.17 suite :



(e) Concentration en O_2 (mg.L^{-1})

(f) concentration en NH_4^+ (µmol.L^{-1})

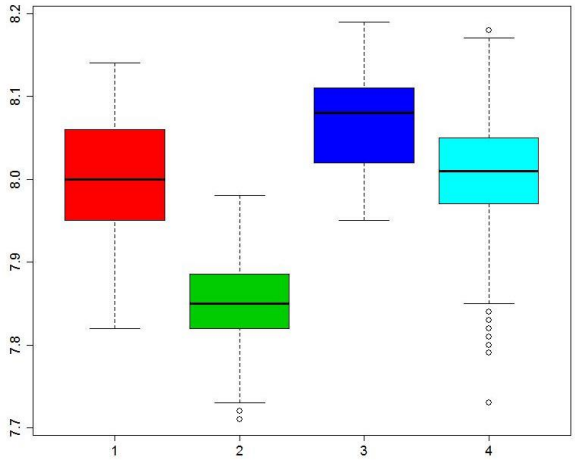
(g) concentration en NO_3^- (µmol.L^{-1})

(h) irradiance (lux)

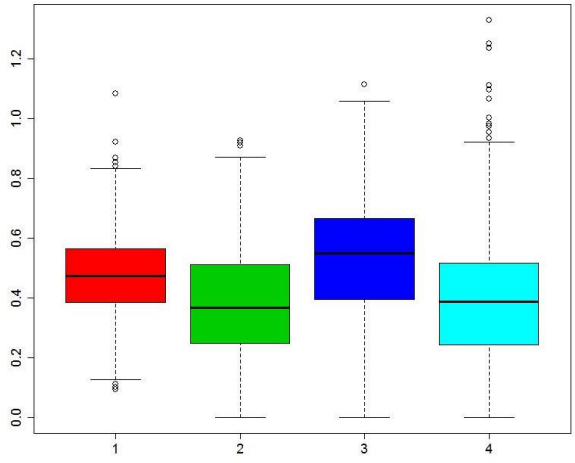
(i) turbidité (NTU)

(j) température (°C)

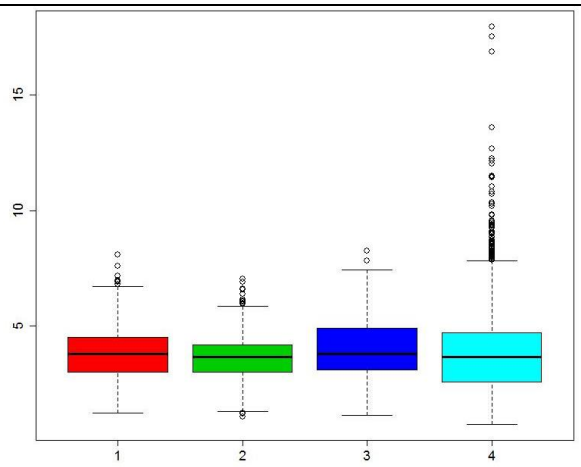
Figure 5.17 suite :



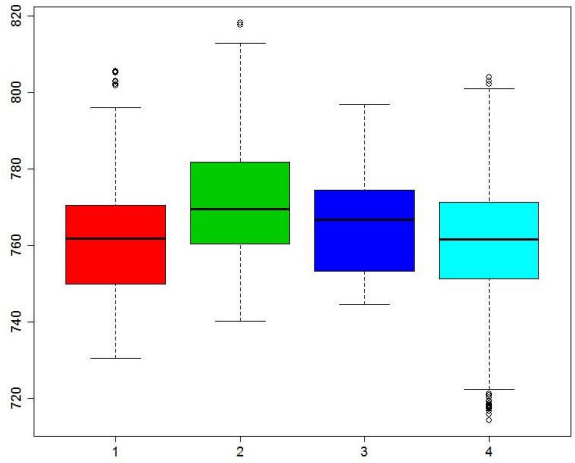
(k) pH (UpH)



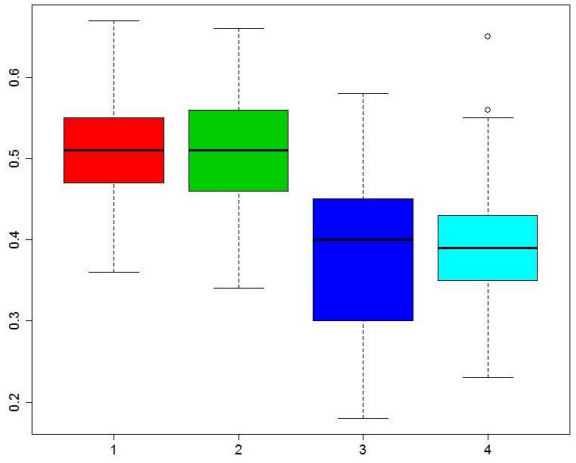
(l) concentration en cyanophycées (µg.L^{-1})



(m) concentration en Carbone Organique Total (mg.L^{-1})



(n) conductivité (mS.cm^{-1})



(o) concentration en PO_4^{3-} (µmol.L^{-1})

5.3.5. Conclusions - Perspectives

La classification spectrale de données acquises toutes les dix minutes au cours d'une période d'un mois en 2009 sur la rivière Deûle a permis de définir des groupes représentatifs d'états environnementaux. Combiné à une analyse en composante principale des données caractéristiques de chacun de groupe, il a été possible de hiérarchiser les contributions des variables et de définir les statistiques de base de ces variables pour chaque groupe. Cette approche permet de définir une dynamique temporelle de fonctionnement de cet environnement continental et d'identifier des événements particuliers comme, par exemple, des augmentations de turbidité liées à la navigation ayant des conséquences potentielles sur la qualité de l'eau (remise en suspension d'algues nuisibles, de contaminants chimiques – thématiques non abordée dans le cadre de cette étude, mais fortement liée).

Les perspectives d'utilisation de ce système MMC-NS hybridé avec une classification spectrale peuvent être résumées en quatre points :

1. L'amélioration des connaissances du fonctionnement des écosystèmes. Ceci passera également par l'intégration de nouveaux paramètres et / ou par l'amélioration de la capacité de discrimination des taxons phytoplanctoniques.
2. L'optimisation du traitement des données issues des systèmes de mesures à haute résolution.
3. La contribution aux réflexions visant à la mise en place d'un système de surveillance adapté et optimisé pour la surveillance du phytoplancton et des paramètres associés en eau douce.
4. La mise en place de stratégies de gestion de l'environnement et de systèmes d'alertes.

Conclusion générale et perspectives

1. Conclusion générale

Le fil conducteur de cette thèse a été de construire un système automatique d'estimation d'états environnementaux caractéristiques à partir des mesures acquises à haute résolution temporelle avec les aléas engendrés de données manquantes ou aberrantes. Aucune connaissance sur les états, leurs caractérisations et leur séquencement n'est apporté dans l'apprentissage du système, à lui d'apprendre automatiquement ces informations à partir des mesures uniquement.

Nous avons opté pour une approche de type modélisation de la dynamique des états sous-jacents à une série temporelle multidimensionnelle et non une approche discriminante des états. Nous avons démontré qu'il est possible d'adapter la structure d'un Modèle de Markov Caché par apprentissage Non Supervisé *sans processus itératif*.

Nous avons aussi montré que cette modélisation markovienne et son système d'estimation associé permet de répondre à des problématiques environnementales.

L'identification d'états environnementaux et la compréhension de la dynamique de ces états est une tâche complexe face à la multitude et la diversité des données collectées via les programmes de recherche, d'observation et de surveillance de l'environnement marin ou fluvial.

Dans le but d'accroître les performances de caractérisation des états, il est important d'apporter aux systèmes d'apprentissage une information utile et complète. Nous avons étudié différentes approches de prétraitements de séries à haute résolution temporelle et notamment mis l'accent sur la comparaison de méthodes d'imputation des données manquantes, données absentes ponctuelles ou sur des périodes de durée critique. Les expérimentations réalisées sur des jeux pédagogiques puis sur données réelles montre que la complétion des données manquantes par recherche de la séquence la plus vraisemblable par appariement élastique est une solution pertinente et rapide. N'ayant pas de certitude sur la distribution des données ainsi qu'une absence de contiguïté trop importante, cette approche n'a pas été implémentée sur les données de la station marine instrumentée MAREL-Carnot.

La deuxième phase de nos travaux a porté sur la construction d'un Modèle de Markov Caché, soit la définition de son architecture structurelle et probabiliste : nombre d'états, caractérisation des états, matrices de transition et émission et probabilité initiale des états. Ces modèles ont largement été approuvés dans un cadre supervisé pour modéliser des séries temporelles à partir de connaissance complète ou quasi-complète. La caractérisation de la structure étant connue a priori, des procédés itératifs d'Expectation-Maximization (EM) basés sur des calculs de maximum de vraisemblance sont utilisés pour déterminer la partie probabiliste du modèle. Lorsqu'aucune information n'est disponible, plusieurs approches adaptent la phase de détection des états de la phase de caractérisation de ces états puis

utilisent une approche EM pour estimer les paramètres probabilistes. Nous avons proposé d'optimiser globalement l'ensemble des paramètres par un processus séquentiel non itératif, pour aboutir à la définition d'un Modèle de Markov Caché Non Supervisé (MMC-NS). Pour cela, un processus de quantification vectorielle automatique est utilisé pour définir des symboles. A partir de ces symboles, une classification spectrale automatique permet d'extraire les états sous-jacents sans hypothèse sur la distribution des données fournies au classifieur. Les paramètres probabilistes sont calculés à partir de cette structuration, par sommation des transitions entre états et apparitions état-symbole dans la base d'apprentissage.
Les algorithmes de génération de symboles et d'états ont été adaptés afin de traiter des bases de données importantes et en temps limité avec pour objectif global de minimiser le nombre de paramètres du modèle et d'obtenir une structuration optimale des données en conservant l'ensemble de l'information haute fréquence. Nous avons veillé à automatiser toutes les procédures afin qu'aucun réglage ne soit nécessaire pour l'utilisateur. Cette approche markovienne hybridée avec une classification spectrale permet ainsi d'une part, une extraction intelligente des signatures caractéristiques d'états usuels et extrêmes dans une série temporelle sans autre connaissance *a priori* et, d'autre part, d'apporter une représentation graphique de la dynamique entre ces états.

L'approche a été utilisée sur trois jeux de données réelles de nature et de dimensions différentes. Le premier modèle a été appris sur une base de mesures, issues de la station MAREL-Carnot implantée dans la rade de Boulogne-sur-Mer, acquises sur une longue période (2005 à 2008), les données extraites ont une taille de 84 614 instants x 10 paramètres. Les données acquises en 2009 ont ensuite été traitées par le système d'estimation des états intégrant le modèle appris. Une premier modèle MMC-NS à 2-états fixés a été construit sur la base d'une segmentation experte, chaque étape de génération de la structure est comparée aux approches supervisées. Nous avons ainsi montré que ce concept de construction séquentielle globale est tout à fait opportun. Le séquencement obtenu sur des données (2009) n'ayant pas participé au modèle a permis de valider le système d'un point de vue numérique. D'un point de vue écologie numérique, un second modèle MMC-NS a été généralisé à N-états, le graphe obtenu répond à la problématique de caractérisation des conditions de déclenchement, de maintien et de fin des efflorescences phytoplanctoniques. Il apporte par ailleurs une plus-value importante puisqu'il a permis de caractériser sur ces données à la fois des évènements cycliques (efflorescences phytoplanctoniques, apports de nutriments,…), d'autres fugaces ou extrêmes (défaillances d'un capteur, ouverture de barrage,…), ensuite de distinguer une structuration particulière au sein de certains de ces événements (exemple : initiation, maximum et déclin de la production de biomasse).
Le second modèle a été appris sur une base de données acquise lors d'une campagne océanographique en Manche orientale au printemps 2012 (Leg 1 (de taille 1 567 instants x 4 paramètres) et Leg 2 (2 593 instants x 4 paramètres)) (fréquence d'acquisition : 1 min.) dans le contexte d'un développement massif le long des côtes françaises d'une microalgue nuisible

pour le fonctionnement des écosystèmes pélagiques et benthiques (*Phaeocystis globosa*). Le système a été validé sur la base de comparaisons entre les segmentations obtenues à dire d'expert et celles proposées par une approche classique (classification hiérarchique ascendante) et le modèle MMC-NS pour différentes parties de la campagne. Il apparaît ainsi que le modèle MMC-NS réalise une segmentation plus cohérente que celle basée sur un arbre de décision issue d'une classification hiérarchique vis-à-vis du dire d'experts. De plus, il a été montré que la prise en compte de la temporalité par le système MMC-NS via ses matrices probabilistes lui permettait d'estimer des états environnementaux de données issues de zones géographiques différentes. Le système permet d'obtenir une vision synoptique de la dynamique de succession d'états environnementaux caractérisés par la présence de différentes classes algales identifiées par fluorimétrie spectrale. Le modèle est capable d'estimer la probabilité d'appartenance à un état donné pour toute nouvelle série de données entrantes.
Le troisième modèle a été appris à partir d'une série (3 742 instants x 15 paramètres) (un mois – fréquence d'acquisition : 1 minute) acquise par une station fixe mise en œuvre temporairement sur la rivière Deûle afin de mieux comprendre le fonctionnement de ce système pour aider à définir un programme de surveillance du phytoplancton en eau douce. La taille de la base de données n'étant pas assez importante pour segmenter celle-ci en deux parties (une pour l'apprentissage et une pour le test), seule la classification spectrale a été utilisée sans volonté de définir la dynamique. La classification spectrale a permis de définir des états environnementaux caractéristiques d'une efflorescence importante de diatomées suivi de sa phase de déclin et de mettre en avant un état impacté par la navigation et pouvant avoir des conséquences en terme de qualité de l'eau.

2. Perspectives

Nous allons maintenant détailler les perspectives nouvelles tant d'un point de vue fondamental qu'applicatif et souligner les travaux en cours concernant l'amélioration du système global.

Optimisation du traitement des données manquantes.

- Indépendamment de la partie estimation, la valeur ajoutée d'une complétion des données manquantes d'une série monodimensionnelle par appariement élastique a été démontrée (chapitre 2) vis-à-vis des approches courantes qui ne permettent pas d'imputer de larges périodes d'absence de données. La comparaison de séries multidimensionnelles a fait ses preuves en classification par appariement élastique conjoint, c'est-à-dire l'acceptation d'une déformation locale conjointe sur l'ensemble des séries au lieu de déformations locales indépendantes sur chacune des séries. L'extension de cette méthode de comparaison de séquences par appariement élastique conjoint à l'imputation multi-paramètre d'une donnée manquante semble ainsi une piste pertinente. Cela permettrait, d'une part, d'enrichir la base d'apprentissage et par conséquent, améliorer la prédiction du système, d'autre part, d'envisager des

techniques robustes d'extraction et caractérisation de l'information (détaillées au chapitre 2).

- Une autre solution serait de compléter les données en aval de la partie estimation des états. Nous avons vu que le système MMC-NS offre un séquencement d'états permettant de reconstruire de manière fidèle des données n'ayant pas participé à l'apprentissage du modèle (validation de la modélisation temporelle, chapitres 4 et 5). La recherche du profil similaire à la séquence précédent la ou les données manquantes contiguës pourrait être contrainte au seul profil de séquencement similaire. De ce fait, l'apport de cette information réduirait le risque d'erreur de reconstruction. Sans biaiser le modèle MMC-NS avec ces données complétées, il pourrait ainsi estimer l'état d'une nouvelle donnée sans perte d'information.

Amélioration de l'architecture du système MMC-NS.

- Le système est capable d'enrichir ses paramètres par apprentissage dynamique de nouvelles données acquises. En effet, l'approche proposée permet de réadapter la structure (états, symboles) et les paramètres dynamiques en temps réel, celle-ci étant optimisée et sans itération EM.
- A partir du séquencement obtenu des observations, il est possible d'affiner les paramètres probabilistes du modèle MMC-NS notamment les paramètres d'émission état-symbole ($\boldsymbol{B}$) et de distribution initiale ($\boldsymbol{\pi}$) en considérant les cycles associés à chacun des états et symboles. Nous pouvons notamment citer les travaux parallèles à cette thèse : Derot Jonathan (thèse UMR 8187 LOG sous la direction de François Schmitt) utilisant des techniques de Décomposition Modale Empirique (EMD) et des Fonctions de Densité de Probabilité (PDF) pour caractériser les cycles.
- Le Modèle de Markov Caché a été construit par apprentissage totalement non supervisé au moyen d'une classification spectrale permettant de caractériser sa structure (états, symboles) à partir de séries HF. L'insertion de connaissances Basse Fréquence, dont notamment des informations de structures (caractérisation connue de certains états) ou des contraintes de distinction d'états entre observations pourront être introduites naturellement dans le système.

Applications environnementales

- L'exploitation de la méthode de recherche de séquences par appariement élastique (chapitre 2) sur le séquencement des états, en initialisant la requête sur les derniers états estimés, permettrait le passage d'une prédiction à l'instant t+1 à une prédiction de l'ordre d'une journée ou d'une semaine. L'analyse de la prédiction ferait office de système d'alerte afin d'adapter la stratégie d'échantillonnage en temps réel.
- Une meilleure définition du fonctionnement des milieux en tenant compte de toutes les échelles de variabilité afin de mettre en place des programmes de surveillance optimisé qui ne seront pas forcément à haute résolution.

- La définition d'états de référence à partir d'une classification non supervisée d'un cours d'eau type (1 modèle pour 1 typologie de masse d'eau) permettrait de fixer des seuils et de faire de nouvelles propositions d'indicateurs et des métriques associées. L'estimation dynamique des états d'un cours d'eau test serait comparée à ces seuils afin de connaître son état au sens DCE (ou DCSMM).

Bibliographie

Ait-Mohand, K., Heutte, L., Paquet, T., Ragot, N. (2010). Adaptation de modèles de Markov cachés - Application à la reconnaissance de caractères imprimés. *In Colloque International Francophone Sur l'Ecrit et Le Documents. CIFED 2010. Sousse (Tunisie)*, pages 12.

Alexandrov, V., Gerstein, M. (2004). Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures. In *BMC Bioinformatics, vol 5 (10)*, pages 10.

Arribas Gil, A. (2007). Estimation dans des modèles à variables cachées : alignement des séquences biologiques et modèles d'évolution. In *Thèse, Université de Paris 11*, pages 128.

Baudry, J.-P., Raftery, A.E., Celeux, G., Lo, K., Gottardo, R. (2010). Combining Mixture Components for Clustering. In *Journal of computational and graphical statistics, vol 9 (2)*, pages 332-353.

Benson, B.B., Krause, D. (1984). The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere. In *Limnology and Oceanography, vol 29 (3)*, pages 620–632.

Beutler, M., Wiltshire, K.H., Meyer, B., Moldaenke, C., Lüring, C., Meyerhöfer, M., Hansen, U.-P., Dau, H. (2002). A fluorometric method for the differentiation of algal populations in vivo and in situ. In *Photosynthesis research, vol 72 (1)*, pages 39–53.

Borcard, D., Gillet, F., Legendre, P. (2011). Numerical ecology with R. Éditeur Springer, pages 306.

Breton, E., Rousseau, V., Parent, J. (2006). Hydroclimatic modulation of diatom/Phaeocystis blooms in nutrient- enriched Belgian coastal waters(North Sea In *Limnology and Oceanography, vol 51 (3)*, pages 1401–1409.

Brockwell, P., Davis, R. (2002). Introduction to time series and forecasting. Seconde Edition, pages 434.

Brylinski, J., Lagadeuc, Y. (1990). L'interface eaux côtières /eaux du large dans le Pas-de-Calais : Une zone frontale. In *Comptes rendus l'Académie des Sciences de Paris, vol 311 (2)*, pages 535 – 540.

Brylinski, J.M. (1993). Ecohydrodynamique pélagique en Manche orientale. In *Habilitation à diriger des recherches, Université de Lille I,* pages 270.

Brylinski, J.M., Brunet, C., Bentley, D., Thoumelin, G., Hilde, D. (1996). Hydrography and Phytoplankton Biomass in the Eastern English Channel in Spring 1992. In *Estuarine, Coastal and Shelf Science, vol 43 (4)*, pages 507–519.

Brzezinski, M. (1985). The Si:C:N ration of marine diatoms: interspecific variability and the effect of some environmental variables. In *Journal of Phycology, vol 21*, pages 347–357.

Caillault, É., Hébert, P., Wacquet, G. (2009). Dissimilarity-based classification of multidimensional signals by conjoint elastic matching: Application to phytoplanktonic species recognition. In *Engineering Applications of Neural Networks*, vol 43, pages 153–164.

Caillault, E., Viard-Gaudin, C., Lallican, P.M. (2005). Training of hybrid ANN/HMM systems for on-line handwriting word recognition. In *Proceedings of 12th Conference of International Graphonomics Society (IGS 2005)*, pages 10.

Chatfield, C. (2004). The Analysis of Time Series An Introduction. Sixième Edition, pages 333.

Chatzis, S.P. (2010). Hidden Markov models with non-elliptically contoured state densities. In *IEEE transactions on pattern analysis and machine intelligence, vol 32 (12),*pages 2297–304.

Chen, Q., Mynett, A.E. (2006). Forecasting Phaeocystis globosa blooms in the Dutch coast by an integrated numerical and decision tree model. In *Aquatic Ecosystem Health & Management, vol 9 (3),* pages 357–364.

Chen, X., Cai, D. (2011). Large scale spectral clustering with landmark-based representation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence,* pages 313–318.

Cloern, J.E. and Jassby, A.D. (2008). Complex seasonal patterns of primary producers at the land-sea interface. In *Ecology Letters, vol 11*, pages 1294-1303.

Cortes, C., Vapnik, V. (1995). Support-vector networks. In *Machine Learning, vol 20 (3),* pages 273–297.

Dauvin, J., Lozachmeur, O. (2006). Mer côtière à forte pression anthropique propice au développement d'une Gestion Intégrée: exemple du bassin oriental de la Manche (Atlantique nord-est). In *VertigO - la revue électronique en sciences de l'environnement, vol 7 (3),* pages 1–24.

Dauvin, JC. (2008). The main characteristics, problems, and prospects for Western European coastal seas. *Marine Pollution Bulletin, vol 57 (1-5),* pages 22–40.

Davies, A.G., Madariaga, I. De, Bautista, B., Fernández, E., Harbour, D.S., Serret, P., Tranter, P.R.G. (1992). The ecology of a coastal Phaeocystis bloom in the north-western English Channel in 1990. In *Journal of the Marine Biological Association of the United Kingdom, vol 72 (3),* pages 691–708.

DCE - 2000/60/CE, Directive 2000/60/CE du Parlement Européen et du conseil, établissant un cadre pour une politique communautaire dans le domaine de l'eau. In *Journal officiel des Communautés européennes L 327/1*, pages 72.

Debyeche, M., Haton, J., Houacine, A. (2007). Improved Vector Quantization Approach for Discrete HMM Speech Recognition System. In *The International Arab Journal of Information Technology, vol 4 (4),* pages 338-344.

Dickey, D.A., Fuller, W.A. (1979). Distribution of the Estimates for Autoregressive Time Series with a Unit Root. *In Journal of the American Statistical Association, vol 74 (366)*, pages 427–431.

Dickey, T. (2003). Emerging ocean observations for interdisciplinary data assimilation systems. In *Journal of Marine Systems, vol 40-41*, pages 5–48.

Dolle, A., Oliva, P. (2001). Courantologie dans la rade du port de Boulogne sur Mer. In *rapport de THETIS oceanology*, pages 26.

Figueiredo, M.A.T., Jain, A.K. (2002). Unsupervised learning of finite mixture models. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 24 (3), pages 381–396.

Fine, S., Singer, Y., Tishby, N. (1998). The hierarchical hidden Markov model: Analysis and applications. In *Machine Learning, vol 32*, pages 41–62.

Forney, G.J. (1973). The viterbi algorithm. In *Proceedings of the IEEE, vol 61 (3)*, pages 268–278.

Glasson-Cicognani, M., Berchtold, A. (2010). Imputation des données manquantes: Comparaison de différentes approches. In *42èmes Journées de Statistique*, pages 6.

Goberville, E., Beaugrand, G., Sautour, B., Tréguer, P., Somlit, T. (2010). Climate-driven changes in coastal marine systems of western Europe. In *Marine Ecology Progress Series, vol 408*, pages 129–147.

Gorsky, G., Ohman, M.D., Picheral, M., Gasparini, S., Stemmann, L., Romagnan, J.-B., Cawood, A., Pesant, S., Garcia-Comas, C., Prejger, F. (2010). Digital zooplankton image analysis using the ZooScan integrated system. In *Journal of Plankton Research, vol 32 (3)*, pages 285–303.

Grosjean, P., Ibanez, F. (2002). Pastecs. Manuel de l'utilisateur de la librairie de fonctions pour R et pour S+. Pages 290.

Halverson, M.J., Pawlowicz, R. (2013). High-resolution observations of chlorophyll-a biomass from an instrumented ferry: Influence of the Fraser River plume from 2003 to 2006. In *Continental Shelf Research, vol 59*, pages 52–64.

Hartigan, J.A., Wong, M.A. (1979). A K-Means Clustering Algorithm. In *Journal of the Royal Statistical Society, Series C (Spplied Statistics)*, vol 28 (1), pages 100–108.

Hebert, C., Lefebvre, A. (2004). Circulation des masses d'eaux dans la rade de Boulogne-sur-Mer. In *rapport Ifremer/R.INT.DEL/BL/RST/04/08*, pages 18.

Holiday, D.M. (2009). Remote sensing of harmful algal blooms in the Mississippi Sound and Mobile Bay: Modelling and algorithm formation. In *Thèse, University of Southern Mississippi*, pages 365.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., Liu, H.H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol 454 (1971), pages 903–995.

Huang, Y., Schmitt, F.G. (2014). Time dependent intrinsic correlation analysis of temperature and dissolved oxygen time series using empirical mode decomposition. In *Journal of Marine Systems*, vol 130, pages 90–100

Ifremer environnement (2014a). REPHY : Réseau de surveillance du phytoplancton et des phycotoxines [en ligne]. In *URL http://envlit.ifremer.fr/surveillance/phytoplancton_phycotoxines/presentation, accédé le 23/06/14*.

Ifremer environnement (2014b). Directive Cadre sur l'eau: éléments de qualité [en ligne]. In *URL http://envlit.ifremer.fr/surveillance/directive_cadre_sur_l_eau_dce/elements_de_qualite, accédé le 27/06/14*.

Jaeger, S., Manke, S., Reichert, J., Waibel, A. (2001). Online handwriting recognition: The NPen++ recognizer. *In International Journal on Document Analysis and Recognition, vol 3*, pages 169–180.

Jain, a. K., Murty, M.N., Flynn, P.J. (1999). Data clustering: a review. In *ACM Computing Surveys, vol 31 (3),* pages 264–323.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. In *Pattern Recognition Letters, vol 31 (8),* pages 651–666.

Keogh, E.J., Pazzani, M.J. (2001). Derivative Dynamic Time Warping. In *First SIAM International Conference on Data Mining (SDM'2001)*, pages 1–11.

Ko, A.H., Sabourin, R., de Souza Britto Jr, A. (2008). A New HMM training and testing scheme. In *Pattern Recognition, ICPR 2008*, pages 4.

Kong, W., Hu, S., Zhang, J., Dai, G., 2013. Robust and smart spectral clustering from normalized cut. In *Neural Computing and Applications, vol 23*, pages 1503–1512.

Koo, J.M., Lee, H.S., Un, C.K. (1992). An improved VQ codebook design algorithm for HMM. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 1*, pages. 357–360.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y. (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root. In *Journal of Econometrics, vol 54*, pages 159–178.

Lamy, D., Artigas, L.F., Jauzein, C., Lizon, F., Cornille, V. (2006). Coastal bacterial viability and production in the eastern English Channel: A case study during a Phaeocystis globosa bloom. In *Journal of Sea Research, vol 56 (3)*, pages 227–238.

Lancelot, C., Keller, M.D., Rousseau, V., Smith, W.O. (1998). Autecology of the Marine Haptophyte Phaeocystis sp. In *NATO Advanced Study Institute, Physiological ecology of harmful algal blooms, vol 41*, pages 209–224.

Lefebvre, A., Guiselin, N., Barbet, F., Artigas, F.L. (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. In *ICES Journal of Marine Science, vol 68 (10),* pages 2029–2043.

Lefebvre A., Mégret C. (2014). Suivi Régional des Nutriments sur le littoral du Nord Pas de Calais Picardie. Bilan de l'année 2013. In *rapport Ifremer/RST.LER.BL/14.05, Laboratoire côtier de Boulogne-sur-Mer*, pages 195.

Lefebvre, A., Repecaud, M., Facq, J.-V., Lefebvre, G., Hitier, B. (2002). Projet d'implantation d'une station de mesures automatisées MAREL dans le port de Boulogne sur mer. In *rapport Ifremer/ R. INT.DEL/BL/RST/02/07*, pages 64.

Legendre, P., Legendre, L. (1998). Numerical ecology, Seconde Edition Anglaise, éditeur Elsevier, pages 870

Liao, X., Runkle, P., Carin, L. (2002). Identification of ground targets from sequential high-range-resolution radar signatures. In *IEEE Transactions on Aerospace and Electronic Systems, vol 38 (4),* pages 1230–1242.

Lorenzen C.J. (1966). A method for continuous measurement of in situ chlorophyll concentration. In *Deep Sea Research, vol 13*, pages 223-227.

Luxburg, U. (2007). A tutorial on spectral clustering. In *Statistics and Computing, vol 17 (4),* pages 395–416.

Margalef, R. (1978). Life-forms of phytoplankton as survival alternatives in an unstable environment. In *Oceanologica acta, vol 1 (4)*, pages 493–509.

Martin, A. P. (2003). Phytoplankton patchiness: the role of lateral stirring and mixing. In *Progress in Oceanography, vol 57 (2),* pages 125–174.

Milligan, G.W., Cooper, M.C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. In *Multivariate Behavioral Research, vol 21 (4),* pages 441–458.

Millie, D.F., Weckman, G.R., Pigg, R.J., Tester, P. a., Dyble, J., Wayne Litaker, R., Carrick, H.J., Fahnenstiel, G.L. (2006). Modeling Phytoplankton Abundance in Saginaw Bay, Lake Huron: Using Artificial Neural Networks To Discern Functional Influence of Environmental Variables and Relevance To a Great Lakes Observing System1. In *Journal of Phycology, vol 42 (2),* pages 336–349

Millot, G. (2011). Comprendre et réaliser les tests statistiques à l'aide de R. Manuel de biostatistique. Seconde Edition, éditeur De Boeck, pages 767.

Moon, T.K. (1996). The expectation-maximization algorithm. In *IEEE Signal Processing Magazine*, vol 13 (6), pages 47–60.

Müller, P., Li, X., Niyogi, K. (2001). Non-photochemical quenching. A response to excess light energy. In *Plant Physiology, vol 125 (4)*, pages 1558–1566.

Najmeddine, H., Suard, F., Jay, A., Maréchal, P., Marié, S. (2012). Mesures de similarité pour l ' aide à l ' analyse des données énergétiques de bâtiments. In *Reconnaissance des Formes et Intelligence Artificielle*, pages 8

Ng, A., Jordan, M., Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14, vol 2*, pages 849-856

Nzigou, A., Lefebvre, A. (2013). Suivi régional des nutriments sur le littoral Nord-Pas de Calais/Picardie. Bilan de l'année 2012. In *rapport Ifremer/RST/LER.BL/13.12*, pages 169.

OSPAR (2010). Bilan de Santé 2010. In *Commission OSPAR*, Londres, pages 176.

Petitjean, F., Ketterlin, A., Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. In *Pattern Recognition*, vol 44, pages 678–693.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257-286.

Rath, T.M., Manmatha, R. (2003). Word image matching using dynamic time warping. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol 2, pages 521–527.

Redfield, A.C. (1958). The biological control of chemical factors in the environment. In *American scientist, vol 46*, pages 205-221.

Reynaud, J.-Y., Tessier, B., Auffret, J.-P., Berné, S., Batist, M. De, Marsset, T., Walker, P. (2003). The offshore Quaternary sediment bodies of the English Channel and its Western Approaches. In *Journal of Quaternary Science, vol 18 (3-4)*, pages 361–371.

Reynolds, C.S., Huszar, V., Kruk, C., Naselli-Flores, L., Melo, S. (2002). Towards a functional classification of the freshwater phytoplankton. In *Journal of Plankton Research, vol 24 (5)*, pages 417–428.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. In *IEEE Transactions on Information, vol 30 (4)*, pages 629–636.

Rousseau, V., Vaulot, D., Casotti, R. (1994). The life cycle of Phaeocystis (Prymnesiophycaea): evidence and hypotheses. In *Journal of Marine Systems, vol 5 (1)*, pages 23–39.

Rousseeuw, K., Lefebvre, A., Poisson-Caillault, E., Hamad, D. (2013a). Detection of contrasted physico-chemical and biological environmental status using unsupervised classification tools. In *5Th FerryBox Workshop, Helsinki, Finland, 24-25 April 2013*.

Rousseeuw, K., Poisson-Caillault, E., Lefebvre, A., Hamad, D. (2013b). Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling In *IEEE International Geoscience and Remote Sensing Symposium – IGARSS 2013, Melbourne, VIC*, pages. 3962–3965.

Ruser, A., Popp, P., Kolbowski, J., Reckermann, M., Feuerpfei, P., Egge, B., Reineke, C., Vanselow, K.. (1999). Comparison of chlorophyll-fluorescence-based measuring systems for the detection of algal groups and the determination of chlorophyll-a concentrations. In *Berichte Forsch.-u. technologiezentr*. Vol 19, pages 27–38.

Sakoe, H., Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol 26, pages 43–49.

Sanguinetti, G., Laidler, J., Lawrence, N.D. (2005). Automatic determination of the number of clusters using spectral algorithms. In *IEEE Machine Learning for Signal Processing. 28-30 Sept 2005*, pages. 28–33.

Schmitt, F.G., Huang, Y,X. (2014). Utilisation de la Décomposition Modale Empirique (EMD) et transformation de Hilbert pour prendre en compte les valeurs manquantes dans l'analyse des propriétés multi-échelles des données MAREL. In *Colloque Instrumentation haute fréquence pour l'observation et la surveillance de l'environnement marin, 10 ans MAREL Carnot. 12 et 13 Juin 2014, Boulogne-sur-Mer – France.*

Seuront, L. (2005). Hydrodynamic and tidal controls of small-scale phytoplankton patchiness. In *Marine Ecology Progress Series, vol 302*, pages 93–101.

Seuront, L., Souissi, S. (2002). Climatic control of Phaeocystis spring bloom in the eastern English Channel (1991–2000). In *La mer, vol 40*, pages 41–51.

Shao, X., Xu, C., Kankanhalli, M.S. (2004). Unsupervised classification of music genre using hidden Markov model. In *IEEE International Conference on Multimedia and Expo 2004*, pages 2023–2026.

Shi, J., Malik, J. (2000). Normalized cuts and image segmentation. In *IEEE Transactions on Pattern Analysis and Machine, vol 22 (8)*, pages 888–905.

Shindler, M., Wong, A., Meyerson, A. (2011). Fast and accurate k-means for large datasets. In *Advances in Neural Information Processing Systems 17*, pages 9.

Singh, S.S., Chauhan, N.C. (2011). K-means v / s K-medoids : A Comparative Study. In *National Conference on Recent Trends in Engineering & Technology, 13-14 mai 2011, Gujarat, India*, pages 5.

Sournia, A., Brylinski, J., Dallot, S. (1990). Fronts hydrologiques au large des côtes françaises: Les sites-ateliers de programme Frontal. In *Oceanologica acta, vol 13 (4)*, pages 413–438.

Trainer, V.L., Bates, S.S., Lundholm, N., Thessen, A.E., Cochlan, W.P., Adams, N.G., Trick, C.G. (2012). Pseudo-nitzschia physiological ecology, phylogeny, toxicity, monitoring and impacts on ecosystem health. In *Harmful Algae, vol 14*, pages 271–300.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. In *IEEE Transactions on Information Theory, vol 13 (2)*, pages 260-269.

Volant, S., Bérard, C., Martin-Magniette, M.-L., Robin, S. (2013). Hidden Markov Models with mixtures as emission distributions. In *Statistics and Computing, vol 24 (4)*, pages 493-504.

Wacquet, G. (2011). Classification spectrale semi-supervisée. Application à la surveillance de l'écosystème marin. In *Thèse, Université du Littoral Côte d'Opale*, pages 219

Wacquet, G., Poisson Caillault, É., Hamad, D., Hébert, P-A. (2013). Constrained spectral embedding for K-way data clustering. In *Pattern Recognition Letters, vol 34 (9)*, pages 1009–1017.

Warren Liao, T. (2005). Clustering of time series data—a survey. In *Pattern Recognition, vol 38 (11)*, pages 1857–1874.

Won, K.-J., Prugel-Bennett, A., Krogh, A. (2006). Evolving the structure of hidden Markov models. In *IEEE Transactions on Evolutionary Computation, vol 10 (1)*, pages 39–49.

Wyatt, T. (2014). Margalef's mandala and phytoplankton bloom strategies. In *Deep Sea Research Part II: Topical Studies in Oceanography, vol 101*, pages 32–49.

Xiang, T., Gong, S. (2008). Spectral clustering with eigenvector selection. In *Pattern Recognition, vol 41*, pages 1012–1029.

Xie, Y., Wiltgen, B. (2010). Adaptive Feature Based Dynamic Time Warping. In *International Journal of Computer Science and Network Security*, vol 10 (1), pages 264–273.

Yan, D., Huang, L., Jordan, M.I. (2009). Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, pages 23.

Zelnik-Manor, L., Perona, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17, vol 2*, pages 1601-1608.

Zighed, D., Rakotomalala, R. (2000). Graphes d'Induction : Apprentissage et Data Mining, édition Hermes, pages 475.

Annexe 1 : Paramètres de la station MAREL-Carnot

A1.1. Oxygène dissous corrigé

Tableau A1.1 Statistiques de base de la concentration en oxygène dissous corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,47	7,94	8,25	9,34	16,38	0,47	1,69	$5{,}11e^{-3}$

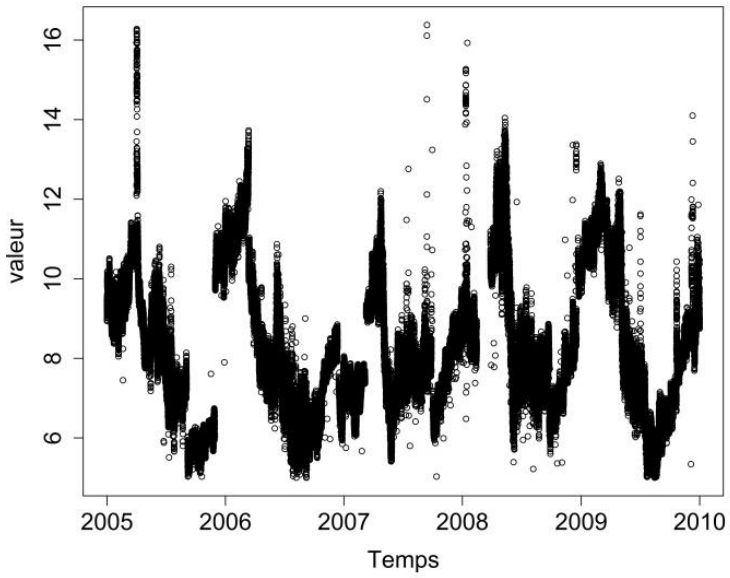


Figure A1.1. Représentation temporelle de la concentration en oxygène dissous corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009.

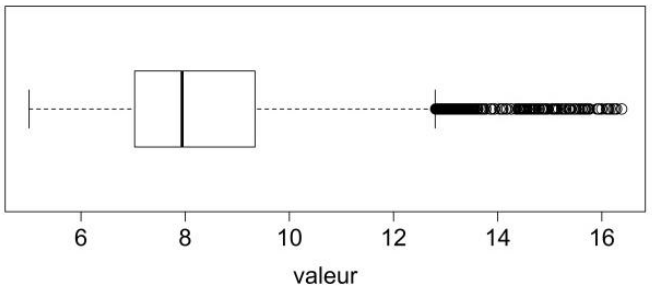


Figure A1.2. Boite de dispersion de la concentration en oxygène dissous corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009.

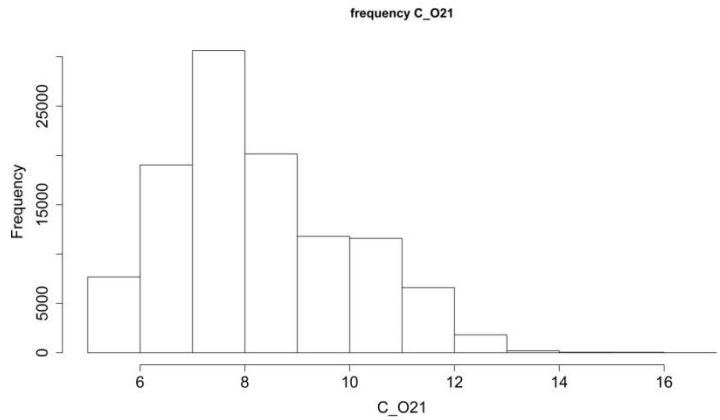


Figure A1. 3. Histogramme en fréquence de la concentration en oxygène dissous corrigé issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution gaussienne selon le test de Kurtosis (p-value***).*

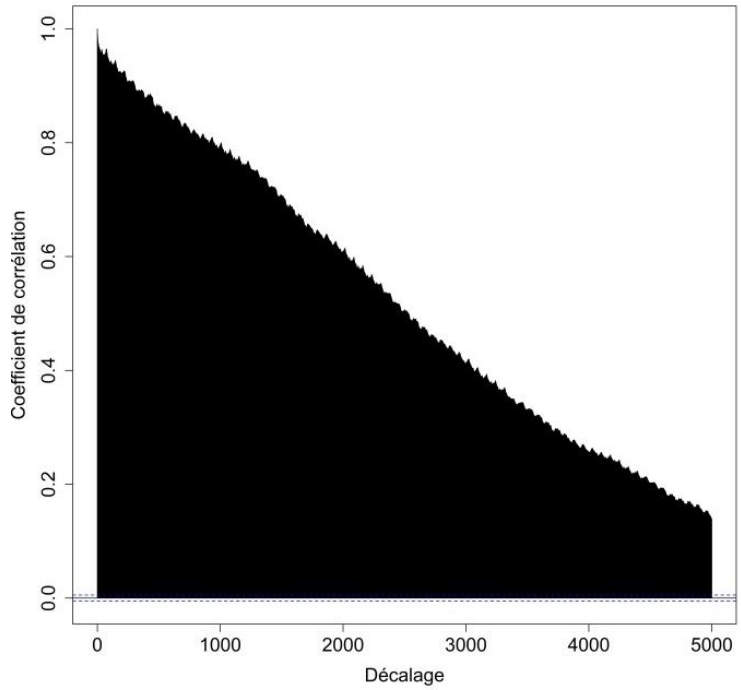


Figure A1.4. Corrélogramme de la concentration en oxygène dissous corrigé issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps.

A1.2. Oxygène dissous non corrigé

Tableau A1.2 Statistiques de base de la concentration en oxygène dissous non corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	5,04	8,49	9,59	10,00	11,37	19,96	1,11	$6{,}36e^{-3}$

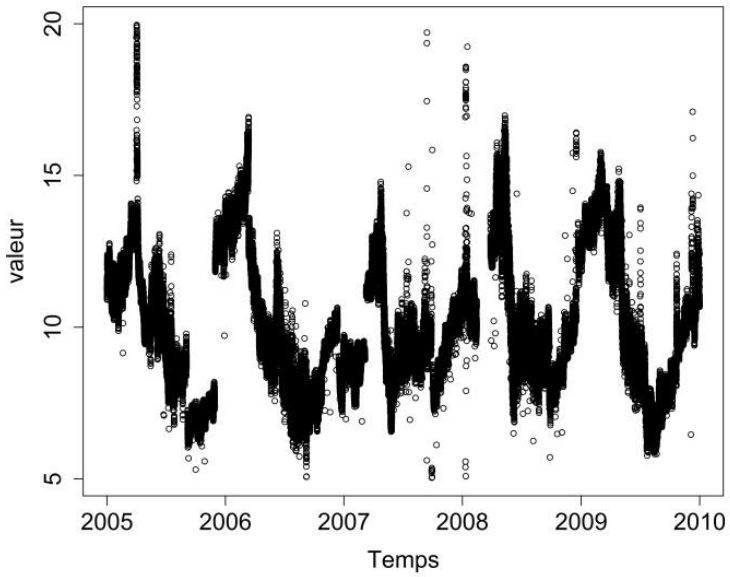


Figure A1.5. Représentation temporelle de la concentration en oxygène dissous non corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009.

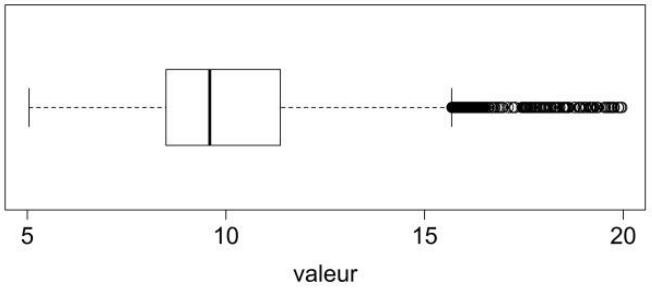


Figure A1.6. Boîte de dispersion de la concentration en oxygène dissous non corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009.

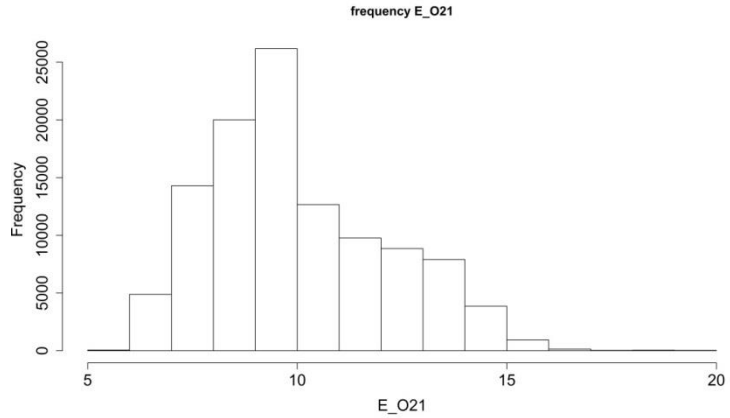


*Figure A1.7. Histogramme en fréquence de la concentration en oxygène dissous non corrigé (mg/L) issue de la station MAREL-Carnot sur la période 2005-2009. Distribution gaussienne selon le test de Kurtosis (p-value***)*

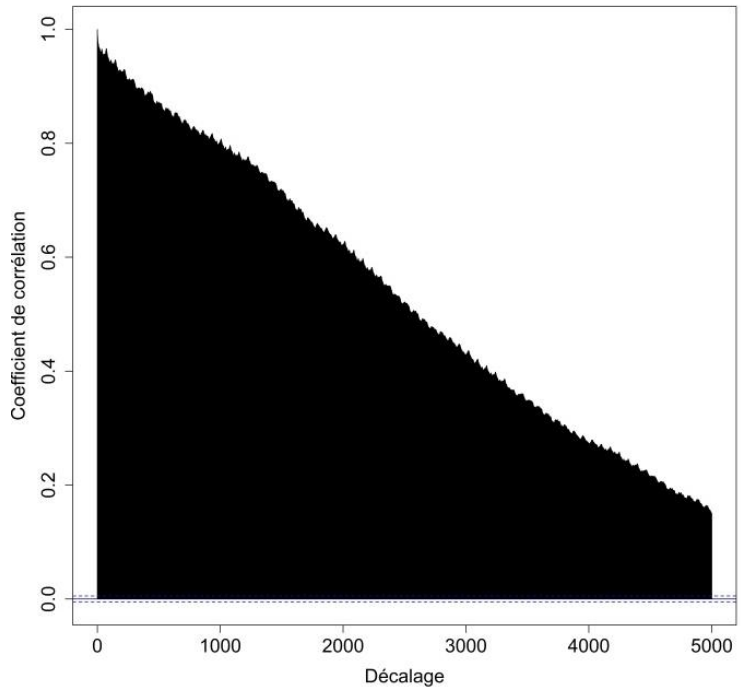


Figure A1.8. Corrélogramme de la concentration en oxygène dissous non corrigé issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps.

A1.3. Pourcentage de saturation en oxygène

Tableau A1.3 Statistiques de base de la saturation en oxygène (%) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	60.45	84,33	93,10	93,89	102,97	130,00	12,99	$3{,}96e^{-2}$

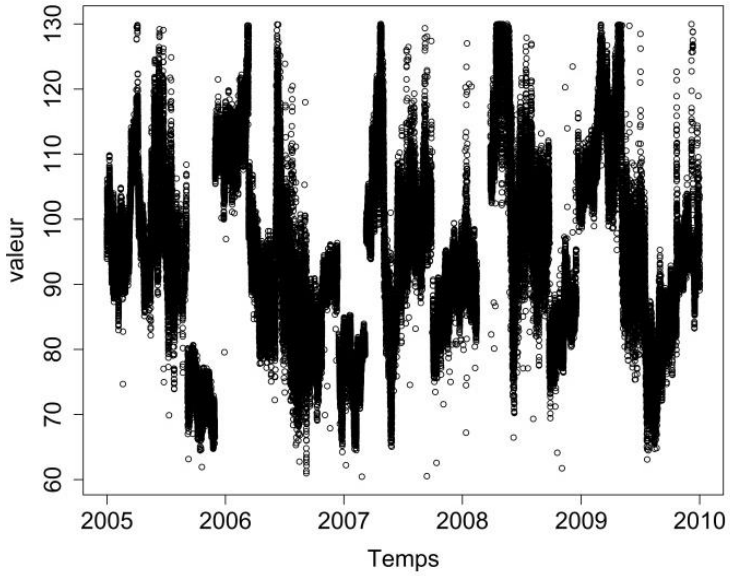


Figure A1.9. Représentation temporelle de la saturation en oxygène (%) issue de la station MAREL-Carnot sur la période 2005-2009

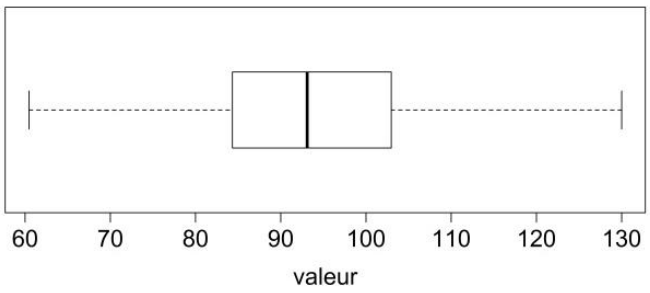


Figure A1.10. Boîte de dispersion de la saturation en oxygène (%) issue de la station MAREL-Carnot sur la période 2005-2009

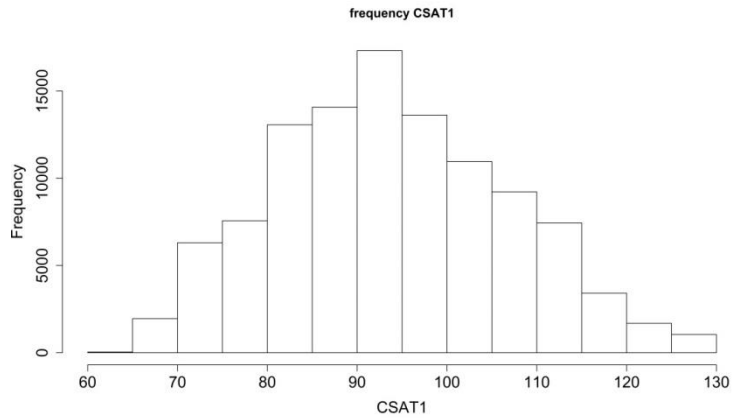


Figure A1.11. Histogramme en fréquence de la saturation en oxygène (%) issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution gaussienne selon le test de Kurtosis (p-value***)*

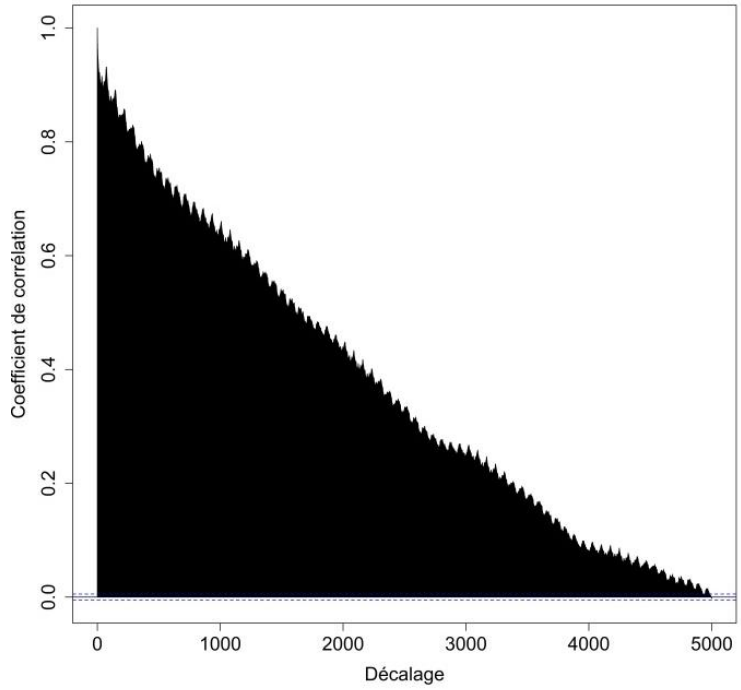


Figure A1.12. Corrélogramme de la saturation en oxygène issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.4. Fluorescence

Tableau A1.4 Statistiques de base de la fluorescence (FFU) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	0,33	0,61	1,38	1,25	45,99	2,69	$7{,}91e^{-3}$

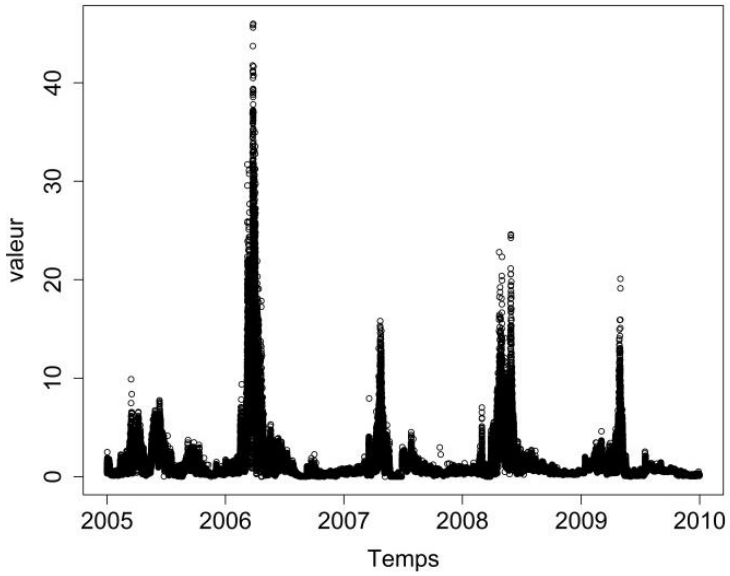


Figure A1.13. Représentation temporelle de la fluorescence (FFU) issue de la station MAREL-Carnot sur la période 2005-2009

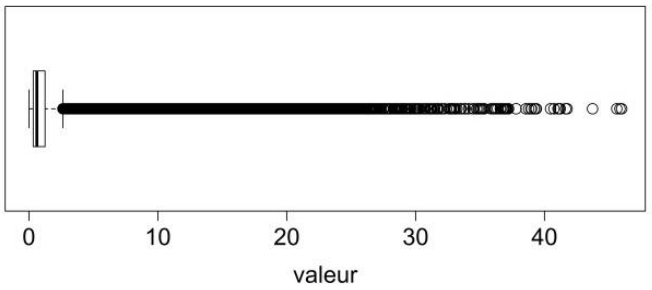


Figure A1.14. Boîte de dispersion de la fluorescence (FFU) issue de la station MAREL-Carnot sur la période 2005-2009

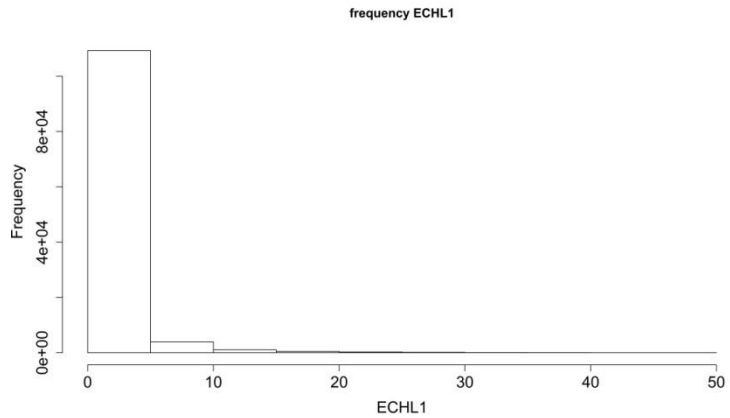


Figure A1.15. Histogramme en fréquence de la fluorescence issue de la station MAREL-Carnot sur la période 2005-2009
*Distribution χ^2 selon le test de Pearson (p-value***).*

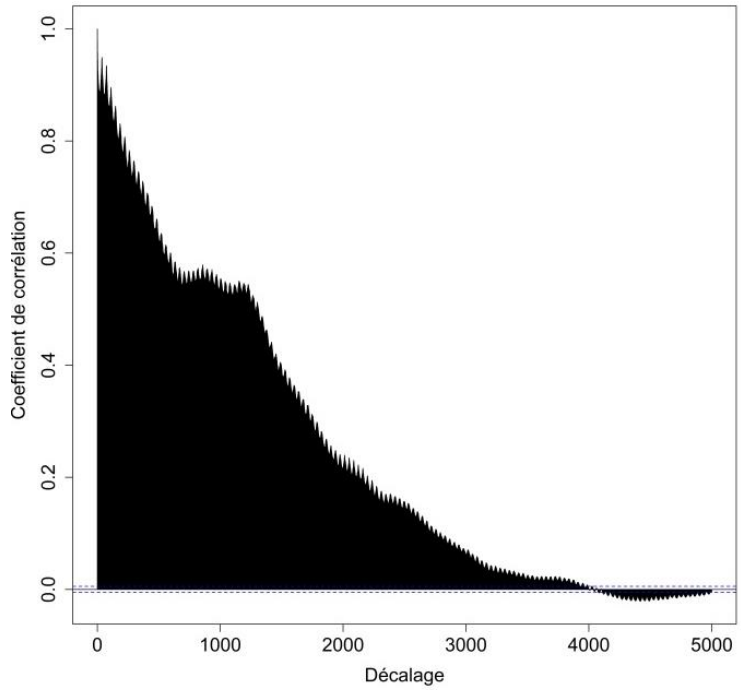


Figure A1.16. Corrélogramme de la fluorescence issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.5. pH

Tableau A1.5 Statistiques de base pH (UpH) issu de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	6,55	8,03	8,35	8,36	8,76	9,50	0,52	$1{,}68e^{-3}$

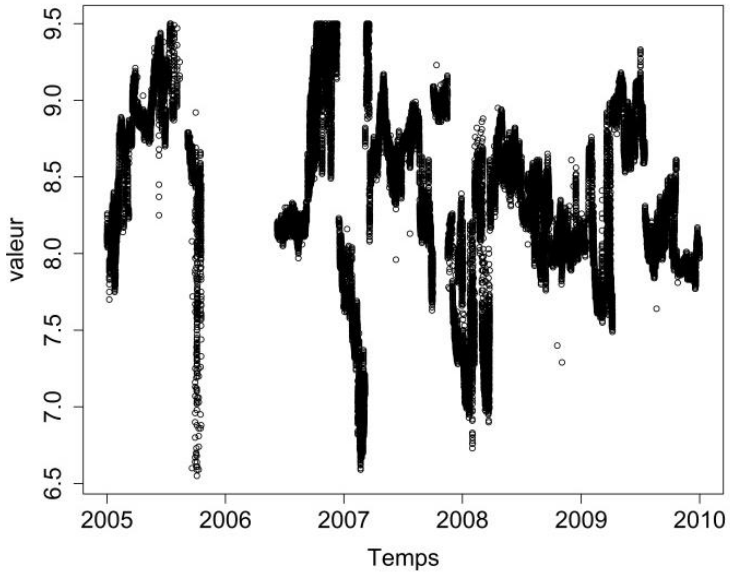


Figure A1.17. Représentation temporelle pH (UpH) issu de la station MAREL-Carnot sur la période 2005-2009

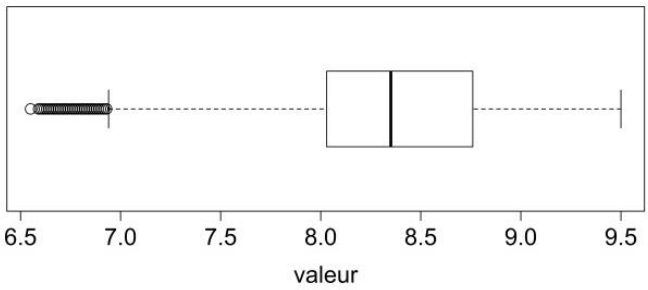


Figure A1.18. Boîte de dispersion du pH (UpH) issu de la station MAREL-Carnot sur la période 2005-2009

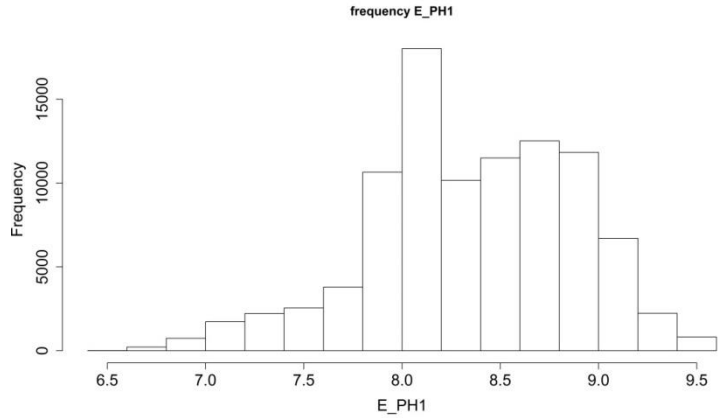


*Figure A1.19. Histogramme en fréquence du pH issu de la station MAREL-Carnot sur la période 2005-2009. Distribution gaussienne selon le test de Kurtosis (p-value***)*

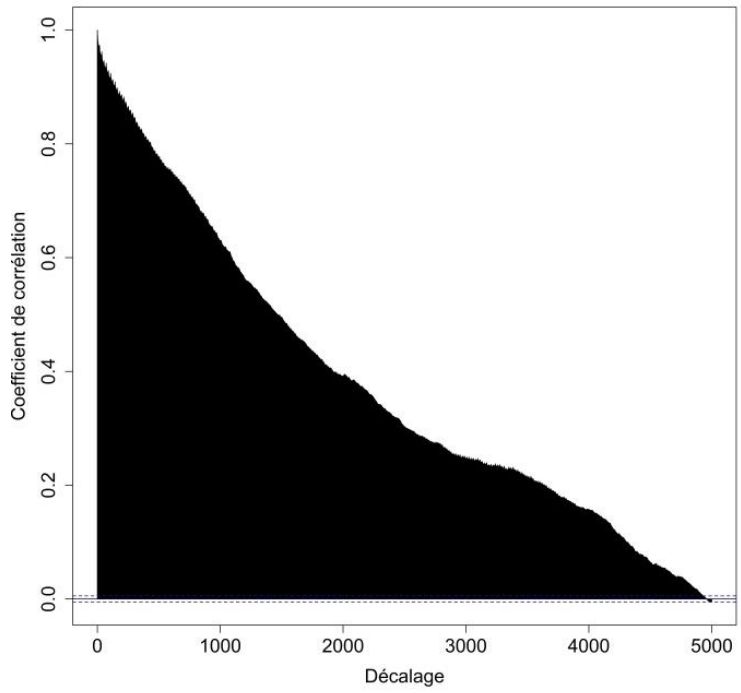


Figure A1.20. Corrélogramme du pH issu de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.6. Salinité

Tableau A1.6 Statistiques de base de la salinité (PSU) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	20,41	33,10	33,56	33,43	33,91	36,28	0,89	$2{,}61e^{-3}$

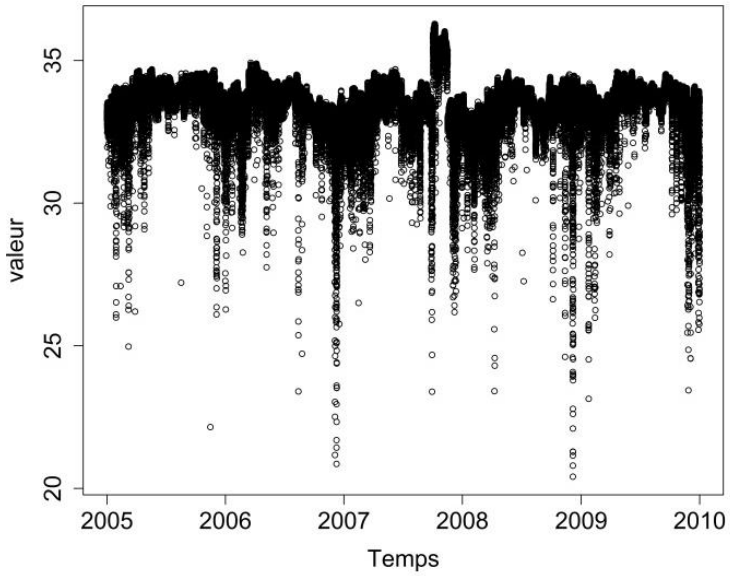


Figure A1.21. Représentation temporelle de la salinité (PSU) issue de la station MAREL-Carnot sur la période 2005-2009

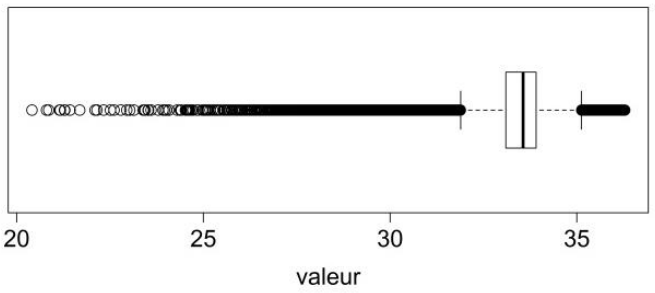


Figure A1.22. Boîte de dispersion de la salinité (PSU) issue de la station MAREL-Carnot sur la période 2005-2009

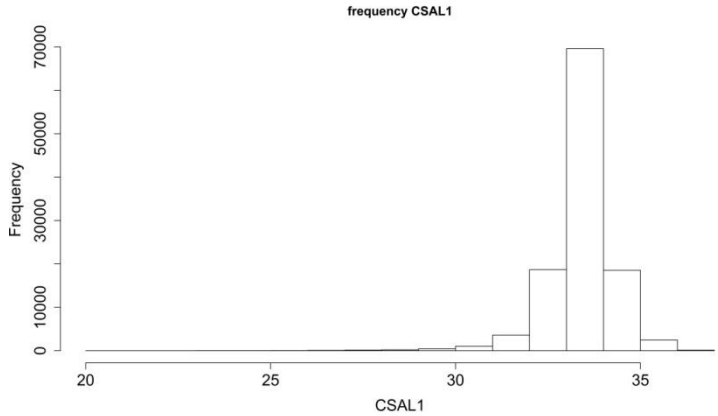


Figure A1.23. Histogramme en fréquence de la salinité issue de la station MAREL-Carnot sur la période 2005-2009

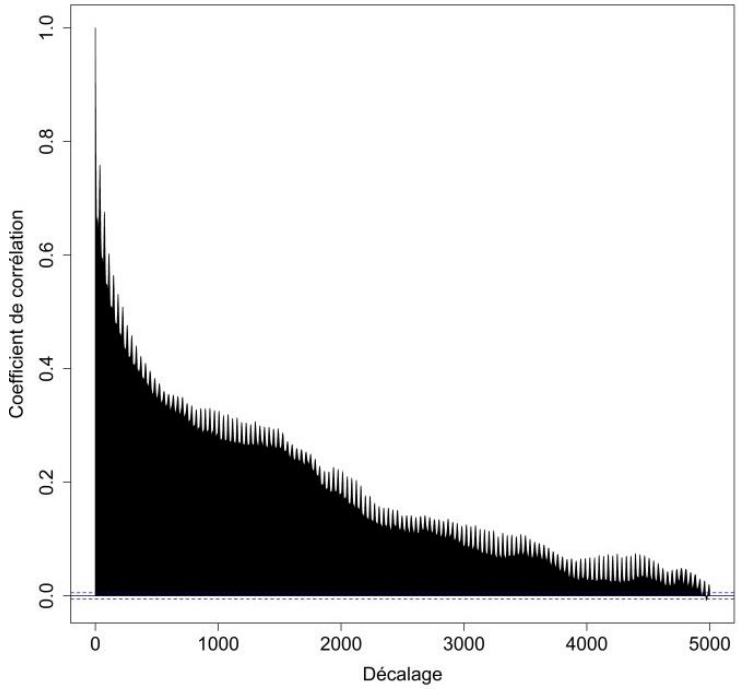


Figure A1.24. Corrélogramme de la salinité issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.7. Conductivité

Tableau A1.7 Statistiques de base de la conductivité (mS.cm^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	31,90	50,50	51,12	50,93	51,59	54,78	1,21	3,57e^{-3}

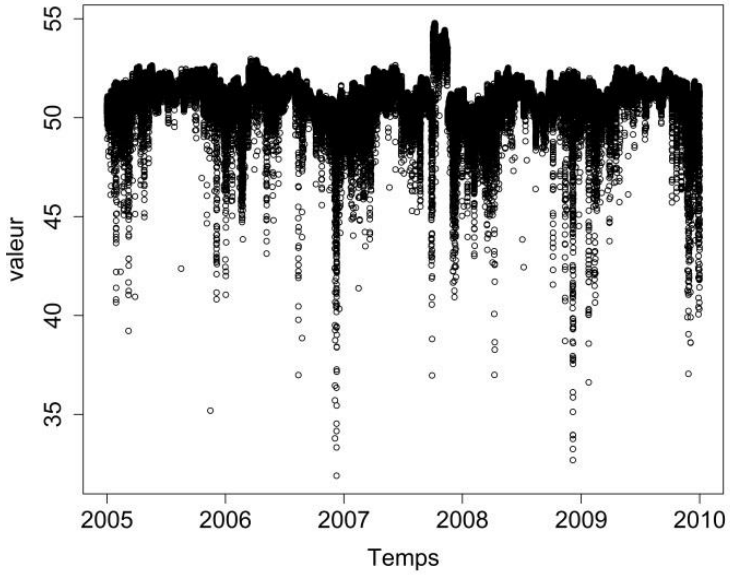


Figure A1.25. Représentation temporelle de la conductivité (mS.cm^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

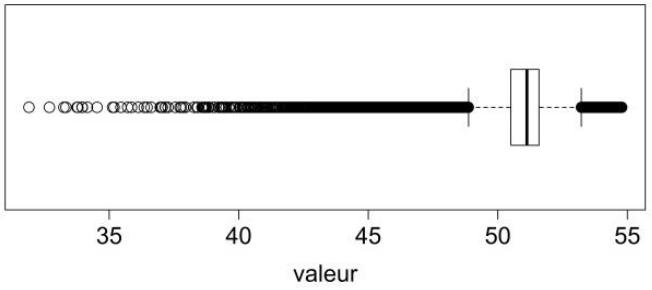


Figure A1.26. Boîte de dispersion de la conductivité (mS.cm^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

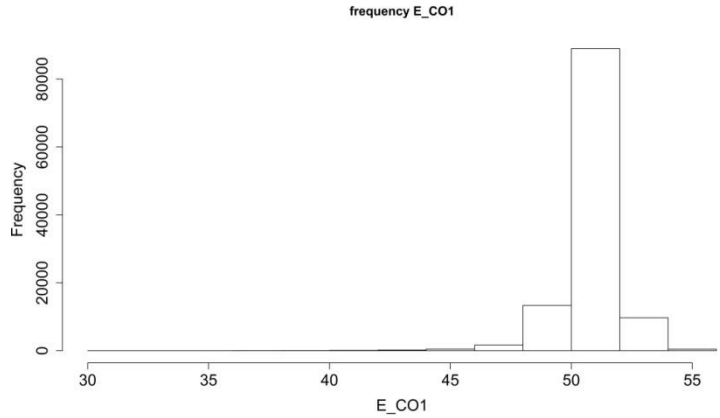


Figure A1.27. Histogramme en fréquence de la conductivité issue de la station MAREL-Carnot sur la période 2005-2009

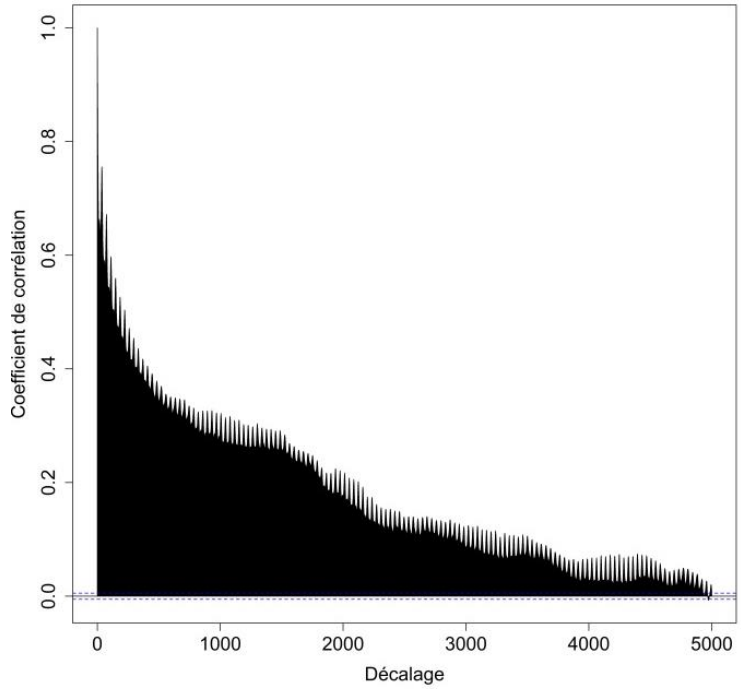


Figure A1.28. Corrélogramme de la conductivité issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.8. Température de l'air

Tableau A1.8 Statistiques de base de la température de l'air (°C) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	-4,27	8,18	12,47	12,12	16,54	29,83	3,38	$1{,}60e^{-2}$

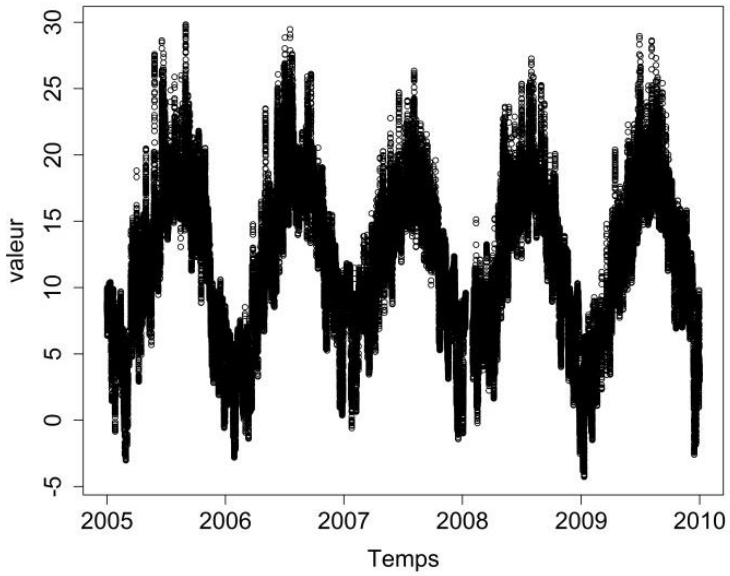


Figure A1.29. Représentation temporelle de la température de l'air (°C) issue de la station MAREL-Carnot sur la période 2005-2009

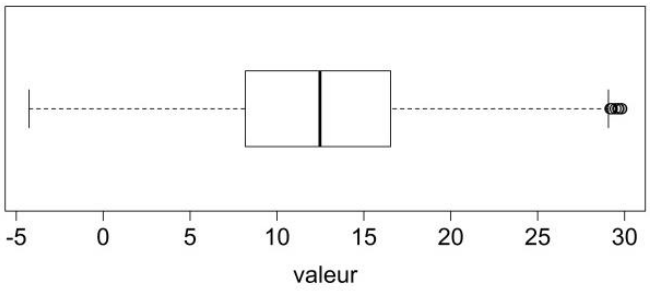


Figure A1.30. Boîte de dispersion de la température de l'air (°C) issue de la station MAREL-Carnot sur la période 2005-2009

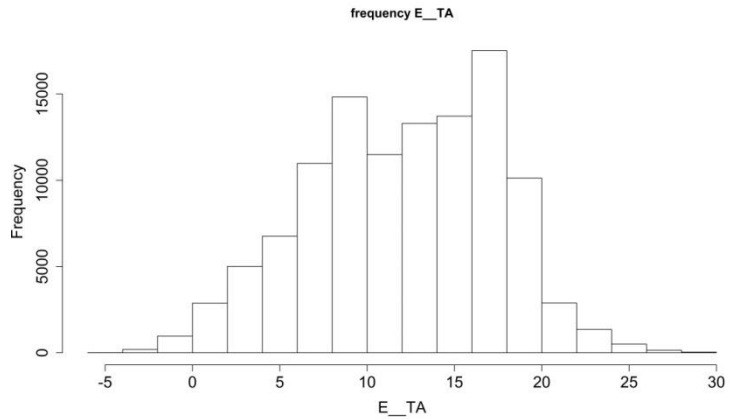


Figure A1.31. Histogramme en fréquence de la température de l'air issue issu de la station MAREL-Carnot sur la période 2005-2009.
*Distribution gaussienne selon le test de Kurtosis (p-value***)*

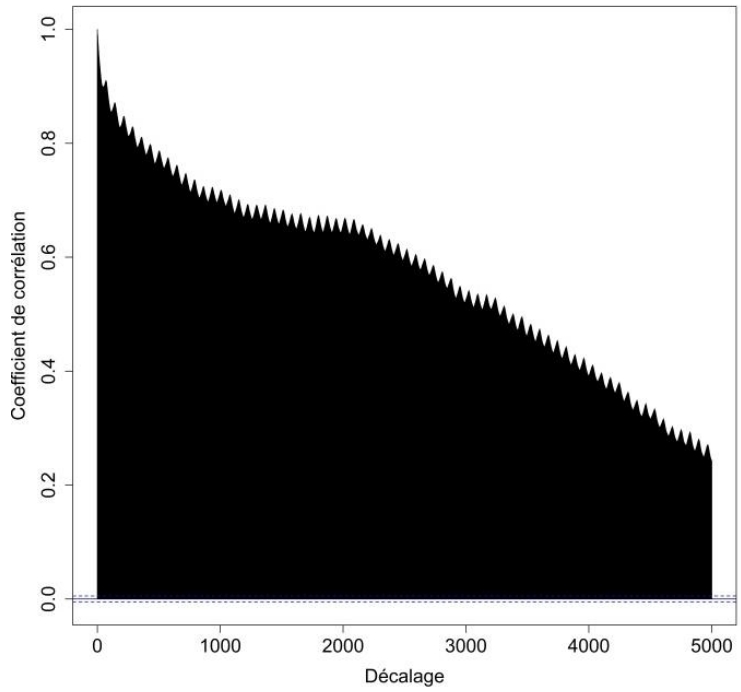


Figure A1.32. Corrélogramme de la température de l'air issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.9. Température de l'eau

Tableau A1.9 Statistiques de base de la température de l'eau (°C) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	3,60	8,60	12,70	12,65	17,10	21,40	4,57	$1,35e^{-2}$

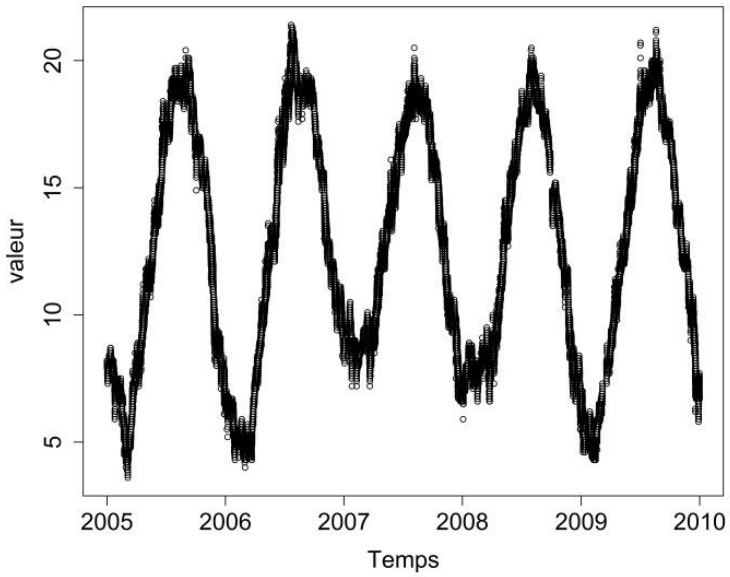


Figure A1.33. Représentation temporelle de la température de l'eau (°C) issue de la station MAREL-Carnot sur la période 2005-2009

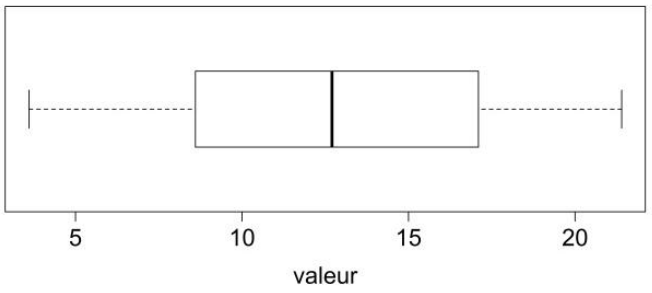


Figure A1.34. Boîte de dispersion de la température de l'eau (°C) issue de la station MAREL-Carnot sur la période 2005-2009

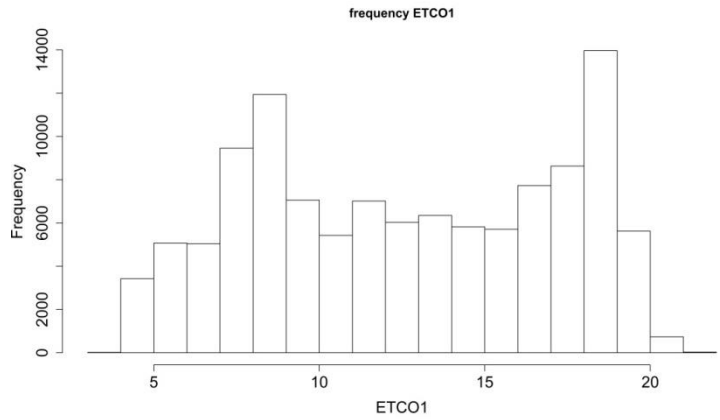


Figure A1.35. Histogramme en fréquence de la température de l'eau issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution : somme de deux gaussienne selon le test de Kurtosis (p-value***)*

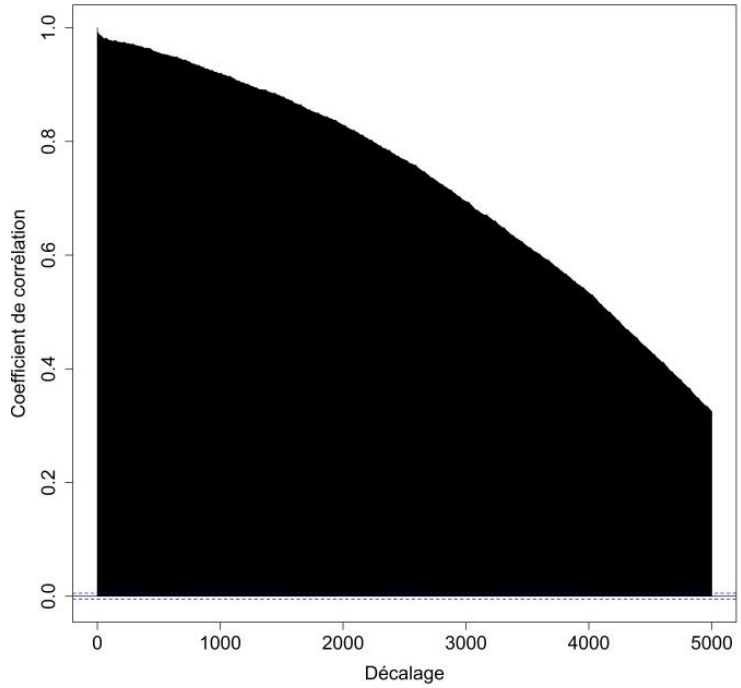


Figure A1.36. Corrélogramme de la température de l'eau issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.10. Hauteur d'eau

Tableau A1.10 Statistiques de base de la hauteur d'eau (m) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,52	2,96	4,86	4,91	6,86	9,26	2,19	$6{,}75.\ 10^{-3}$

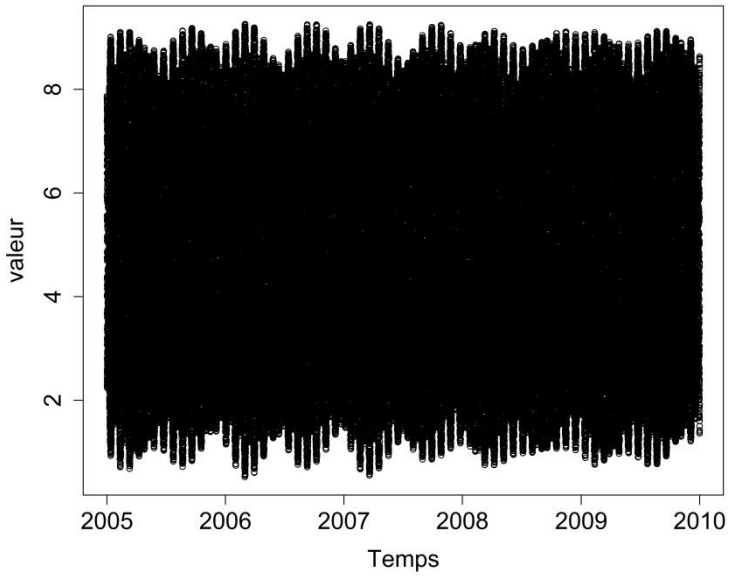


Figure A1.37. Représentation temporelle de la hauteur d'eau (m) issue de la station MAREL-Carnot sur la période 2005-2009

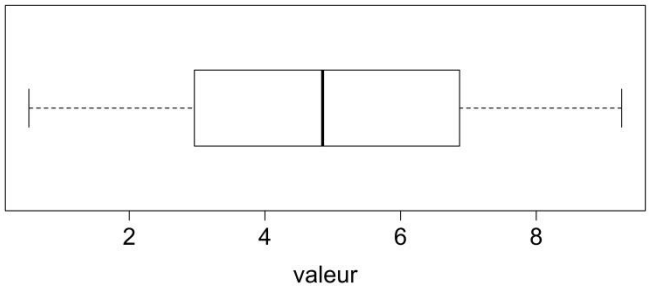


Figure A1.38. Boîte de dispersion de la hauteur d'eau (m) issue de la station MAREL-Carnot sur la période 2005-2009

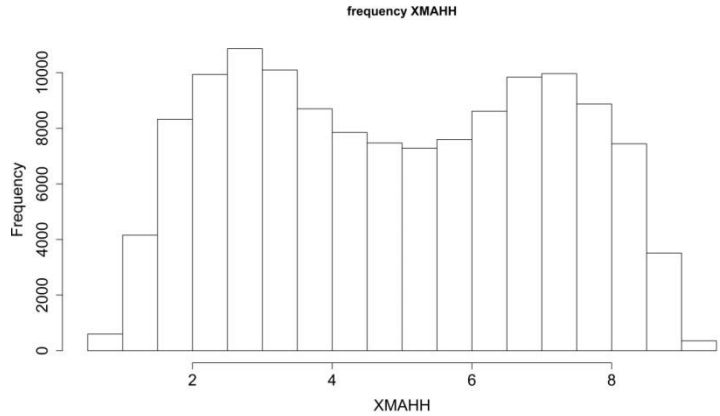


Figure A1.39. Histogramme en fréquence de la hauteur d'eau issue de la station MAREL-Carnot sur la période 2005-2009
*Distribution : somme de deux gaussienne selon le test de Kurtosis (p-value***)*

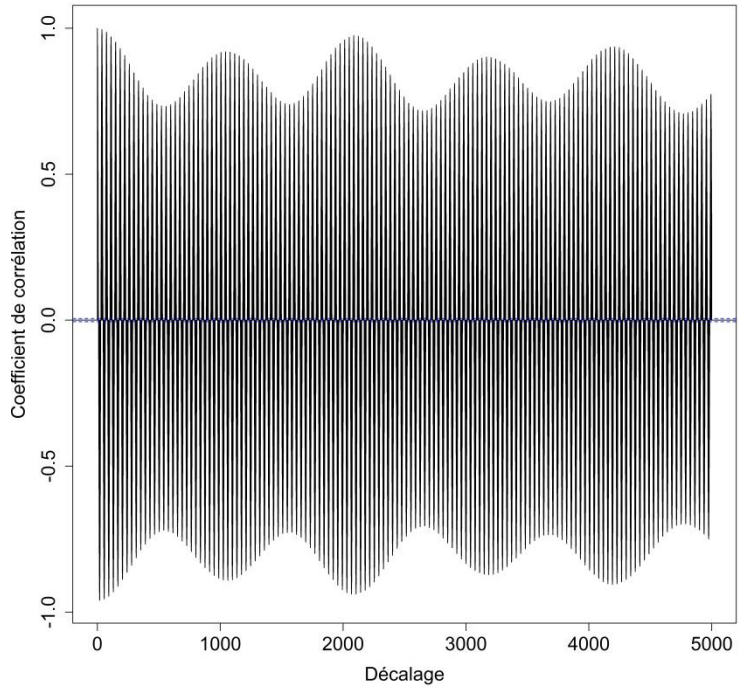


Figure A1.40. Corrélogramme de la hauteur d'eau issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.11. Vitesse du vent en moyenne

Tableau A1.11 Statistiques de base de vitesse du vent en moyenne ($m.s^{-1}$) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	6,00	9,52	10,14	13,96	40,51	5,43	$1,57e^{-2}$

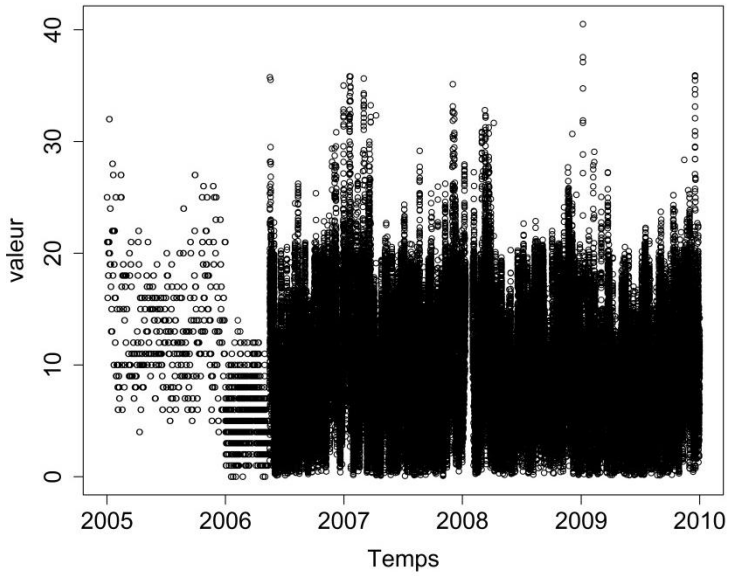


Figure A1.41. Représentation temporelle de la vitesse du vent en moyenne ($m.s^{-1}$) issue de la station MAREL-Carnot sur la période 2005-2009

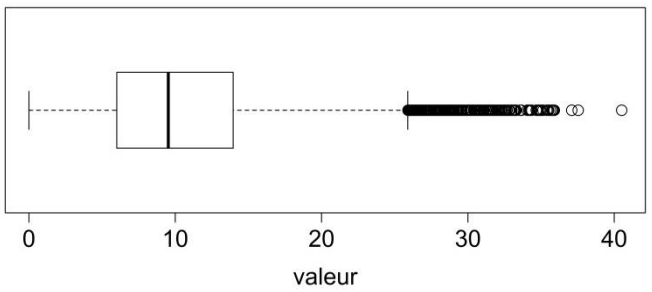


Figure A1.42. Boîte de dispersion de la vitesse du vent en moyenne ($m.s^{-1}$) issue de la station MAREL-Carnot sur la période 2005-2009

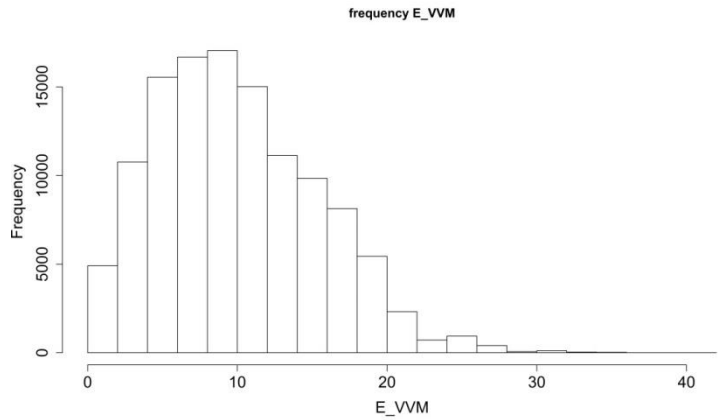


Figure A1.43. Histogramme en fréquence de la vitesse du vent en moyenne issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

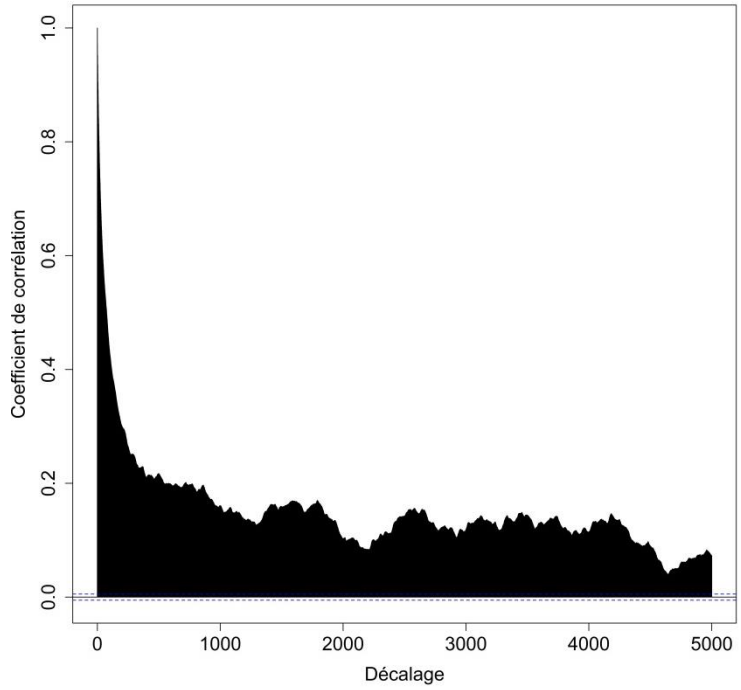


Figure A1.44. Corrélogramme de la vitesse du vent en moyenne issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.12. Vitesse du vent en rafale

Tableau A1.12 Statistiques de base de la vitesse du vent en rafale ($m.s^{-1}$) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	6,00	9,52	10,14	13,96	40,51	5,43	$1,57e^{-2}$

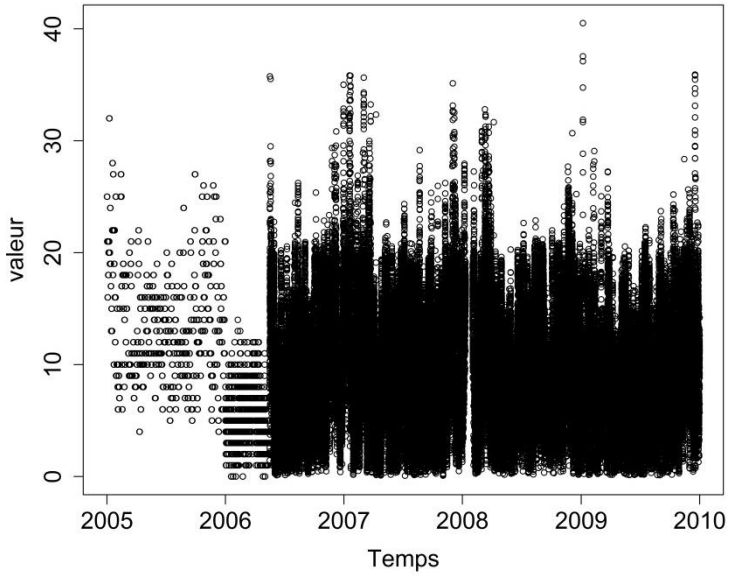


Figure A1.45. Représentation temporelle de la vitesse du vent en rafale ($m.s^{-1}$) issue de la station MAREL-Carnot sur la période 2005-2009

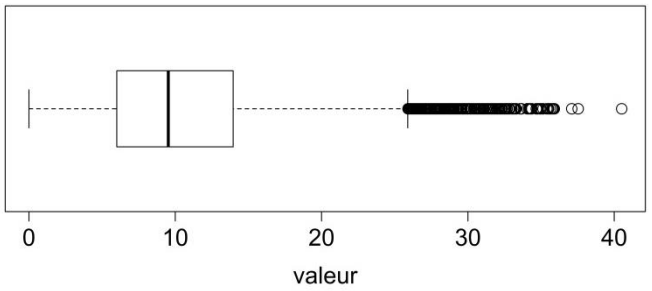


Figure A1.46. Boîte de dispersion de la vitesse du vent en rafale ($m.s^{-1}$) issue de la station MAREL-Carnot sur la période 2005-2009

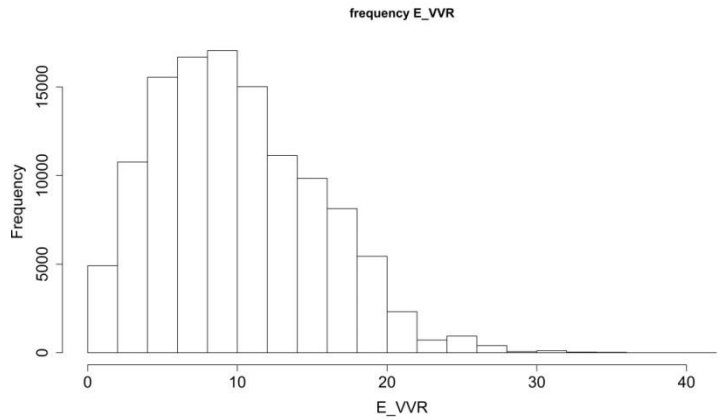


Figure A1.47. Histogramme en fréquence de la vitesse du vent en rafale issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

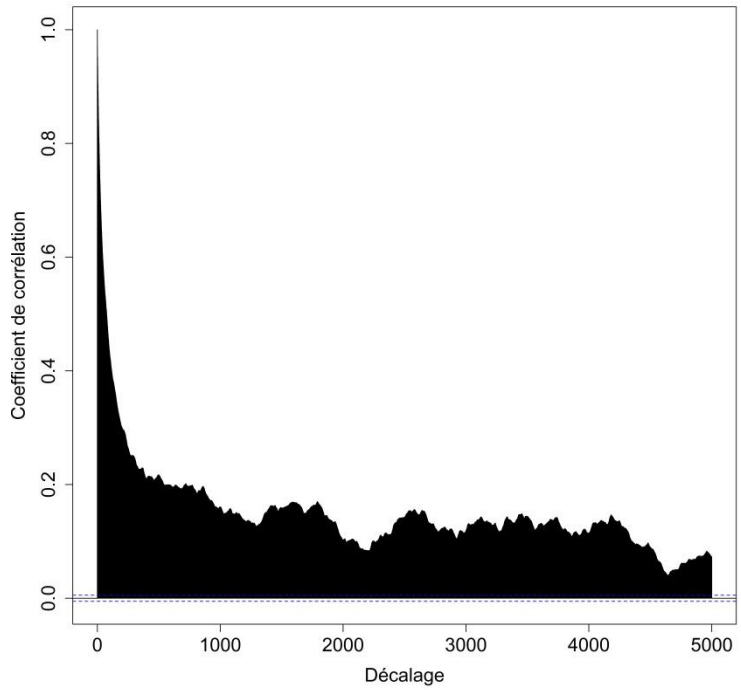


Figure A1.48. Corrélogramme de la vitesse du vent en rafale issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.13. Direction du vent

Tableau A1.13 Statistiques de base de la direction du vent (degré) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	97,0	209,0	184,7	250,0	360,0	95,25	0,28

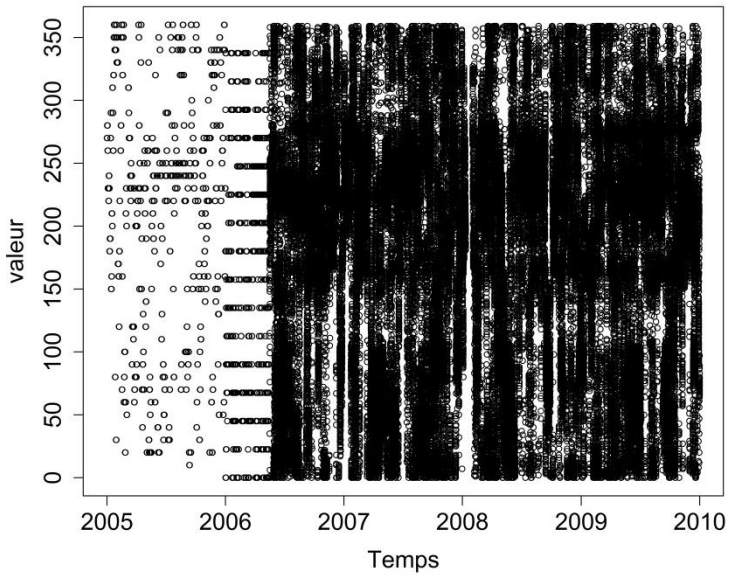


Figure A1.49. Représentation temporelle de la direction du vent (degré) issue de la station MAREL-Carnot sur la période 2005-2009

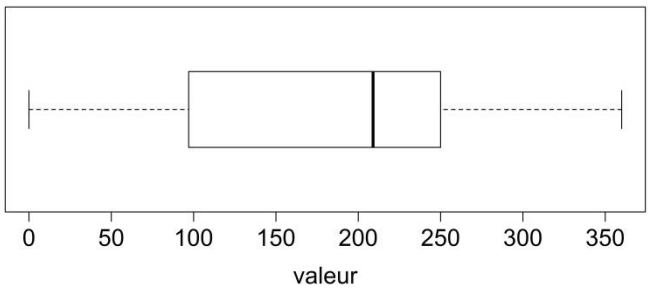


Figure A1.50. Boîte de dispersion de la direction du vent (degré) issue de la station MAREL-Carnot sur la période 2005-2009

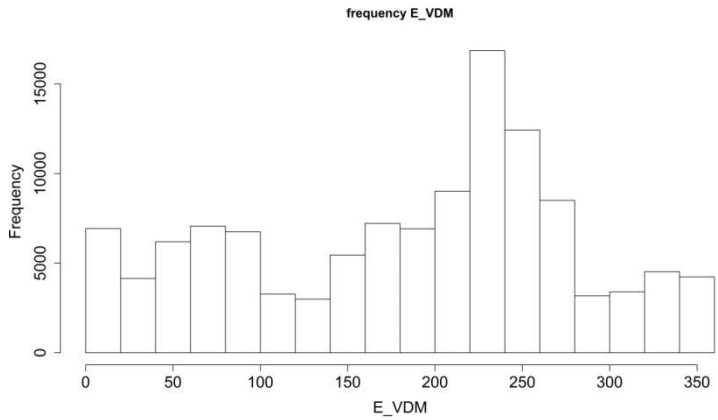


Figure A1.51. Histogramme en fréquence de la direction du vent issue de la station MAREL-Carnot sur la période 2005-2009

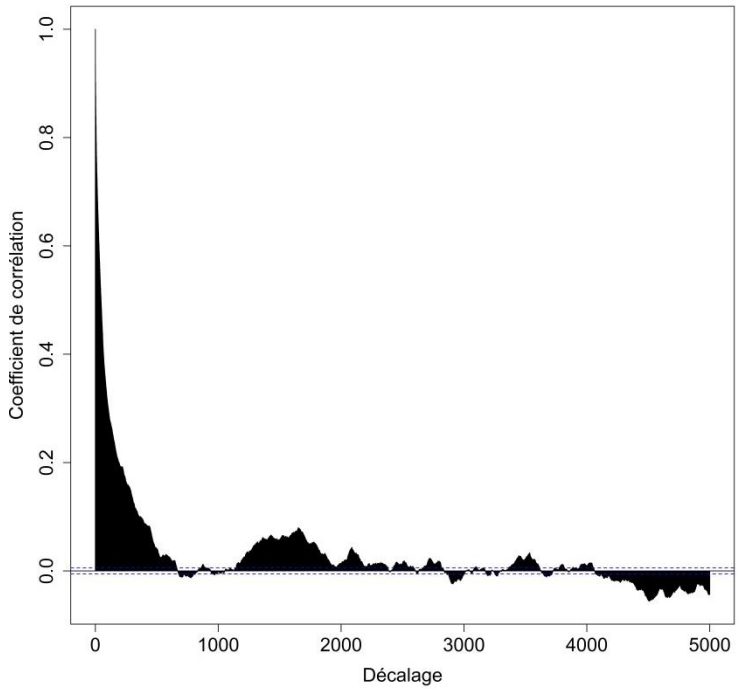


Figure A1.52. Corrélogramme de la direction du vent issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.14. Photosynthetically Active Radiation (P.A.R.)

Tableau A1.14 Statistiques de base du PAR (µmol de photons .s^{-1}.m^{-2}) issu de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	0,00	13,30	292,20	428,30	2487,80	466,65	1,38

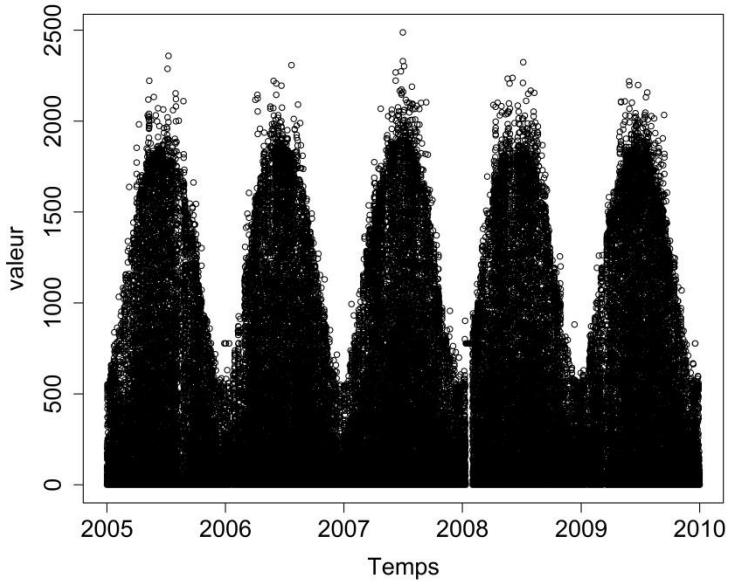


Figure A1.53. Représentation temporelle du PAR (µmol de photons .s^{-1}.m^{-2}) issu de la station MAREL-Carnot sur la période 2005-2009

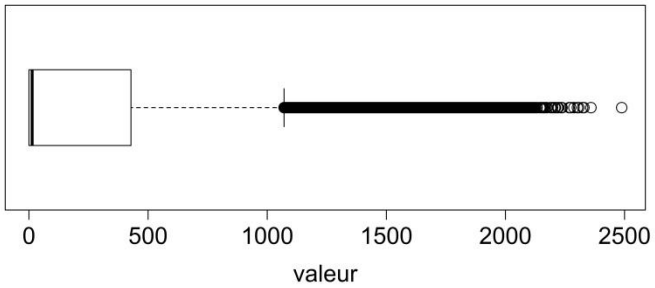


Figure A1.54. Boîte de dispersion du PAR (µmol de photons .s^{-1}.m^{-2}) issu de la station MAREL-Carnot sur la période 2005-2009

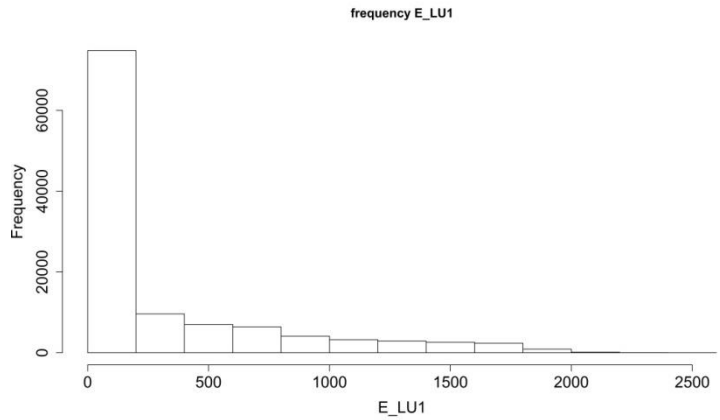


Figure A1.55. Histogramme en fréquence du PAR issu de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

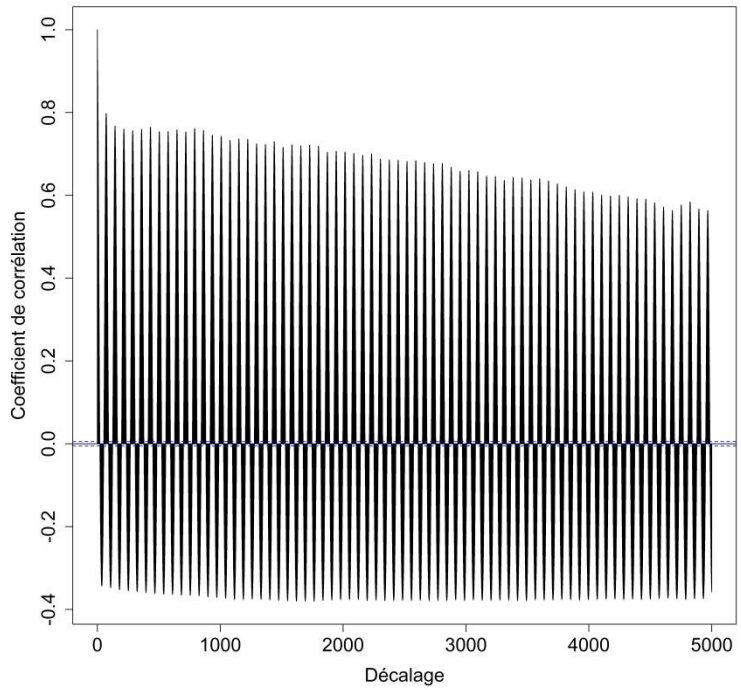


Figure A1.56. Corrélogramme du PAR issu de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.15. Turbidité

Tableau A1.15 Statistiques de base de la turbidité (NTU) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	4,30	7,70	12,31	14,50	148,90	14,27	$4{,}22e^{-2}$

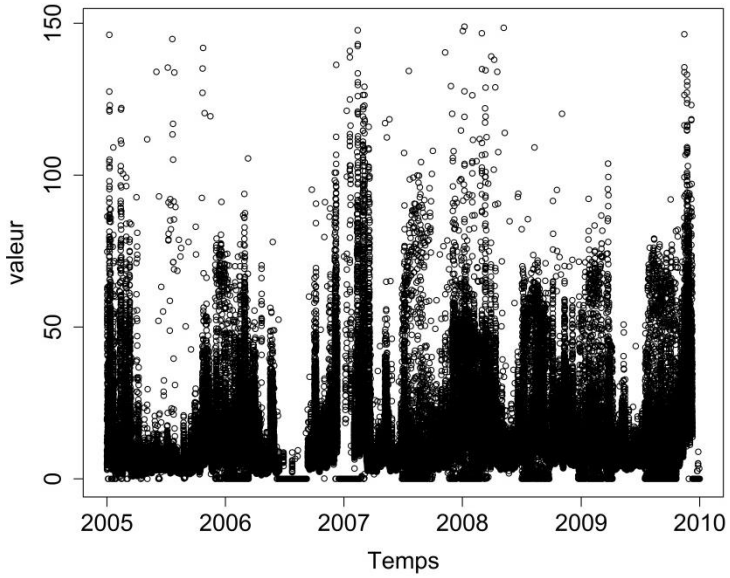


Figure A1.57. Représentation temporelle de la turbidité (NTU) issue de la station MAREL-Carnot sur la période 2005-2009

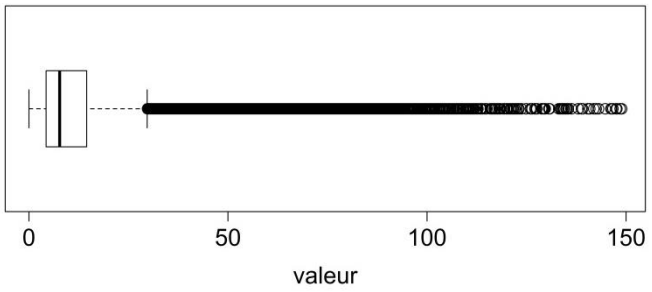


Figure A1.58. Boîte de dispersion de la turbidité (NTU) issue de la station MAREL-Carnot sur la période 2005-2009

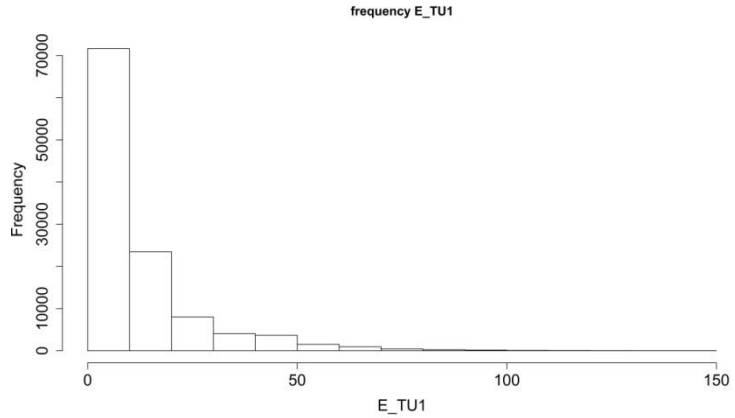


Figure A1.59. Histogramme en fréquence de la turbidité issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

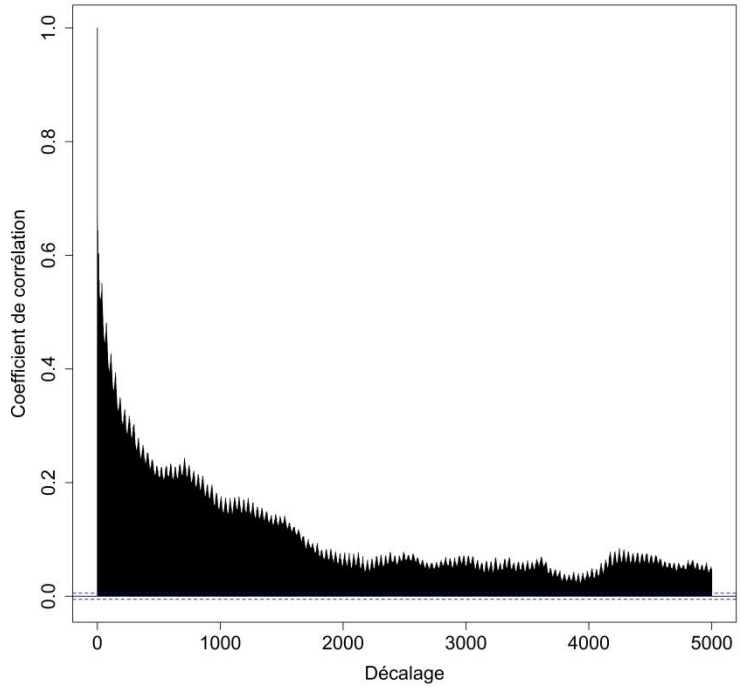


Figure A1.60. Corrélogramme de la turbidité issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.16. Concentration en nitrate

Tableau A1.16 Statistiques de base de la concentration en nitrate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,01	5,90	13,58	18,24	26,11	99,54	16,19	0,30

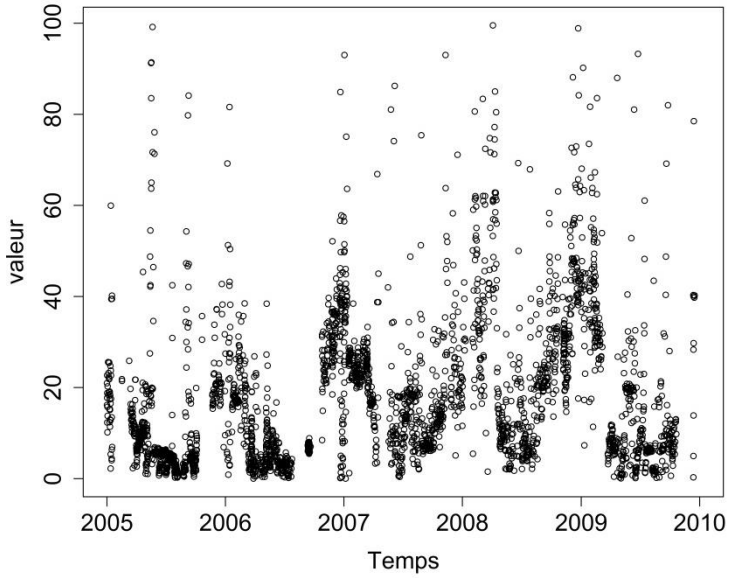


Figure A1.61. Représentation temporelle de la concentration en nitrate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

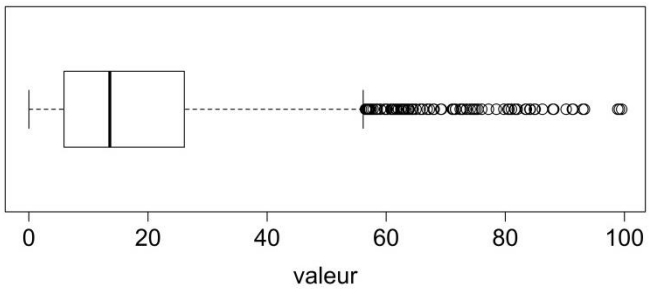


Figure A1.62. Boîte de dispersion de la concentration en nitrate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

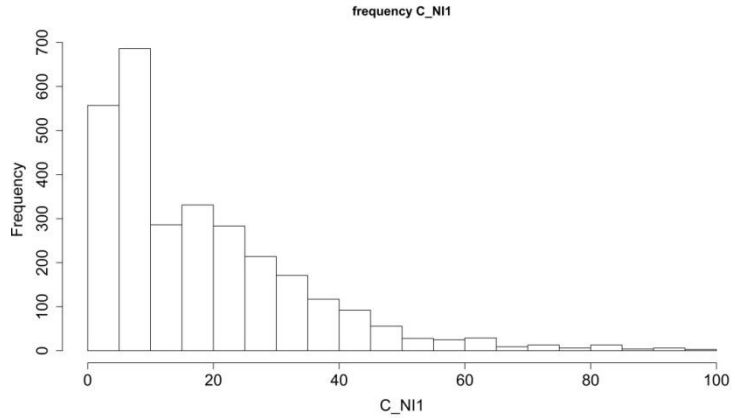


Figure A1.63. Histogramme en fréquence de la concentration en nitrate issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

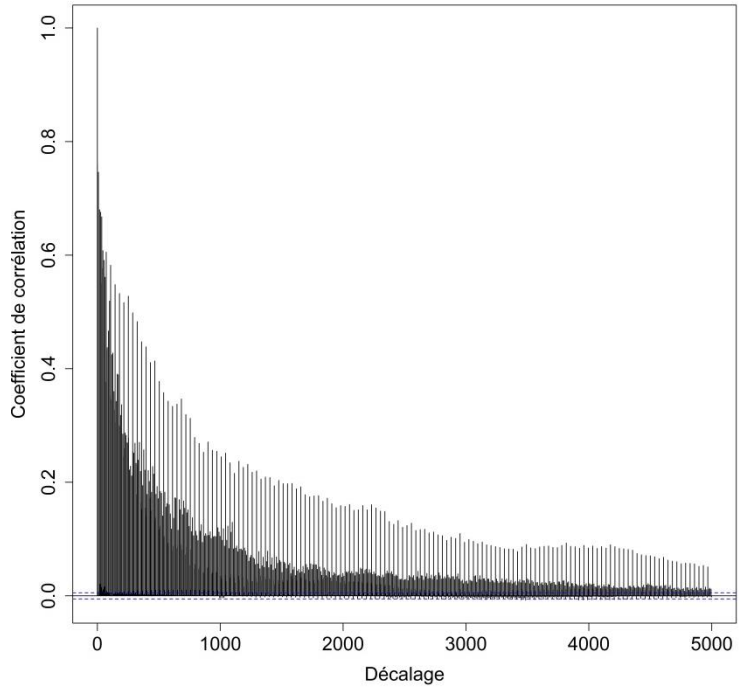


Figure A1.64. Corrélogramme de la concentration en nitrate issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps.

A1.17. Concentration en phosphate

Tableau A1.17 Statistiques de base de la concentration en phosphate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,00	0,47	0,74	2,03	1,06	94,00	6,41	0,12

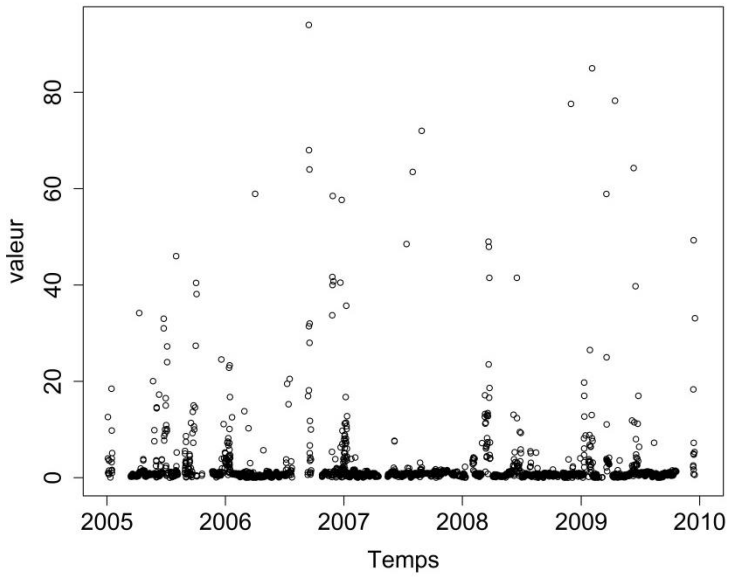


Figure A1.65. Représentation temporelle de la concentration en phosphate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

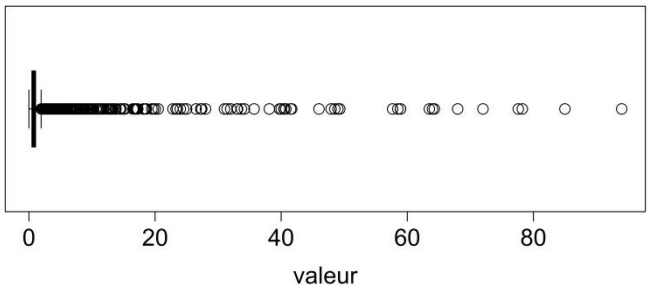


Figure A1.66. Boîte de dispersion de la concentration en phosphate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

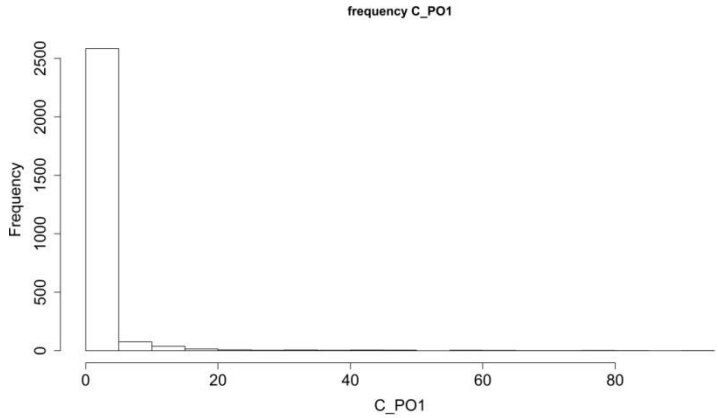


Figure A1.67. Histogramme en fréquence de la concentration en phosphate issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

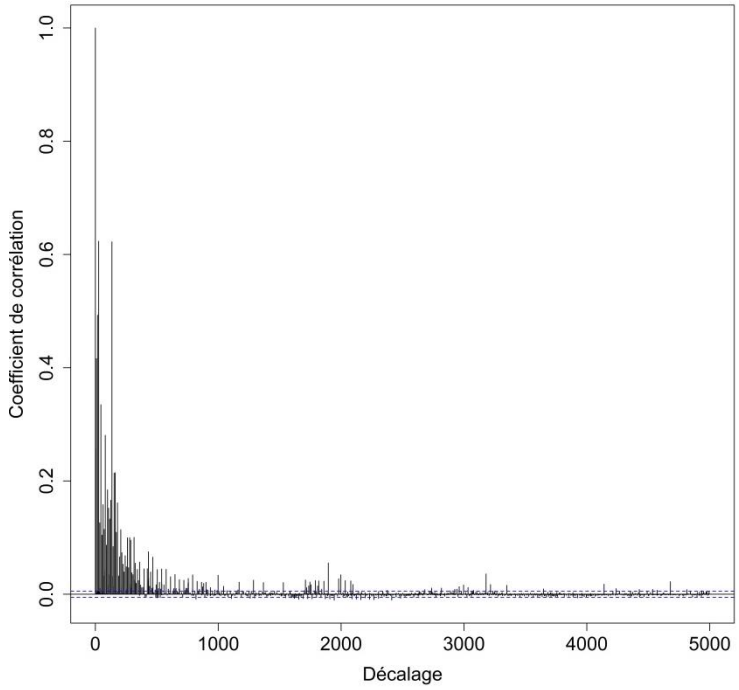


Figure A1.68. Corrélogramme de la concentration en phosphate issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps

A1.18. Concentration en silicate

Tableau A1.18 Statistiques de base de la concentration en silicate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009, avec N le nombre de données, Q1 le premier quantile et Q3 le troisième quantile.

N	Minimum	Q1	Médiane	Moyenne	Q3	Maximum	Ecart-type	Erreur standard de la moyenne
131 472	0,01	1,74	4,28	5,88	8,32	95,29	6,29	0,11

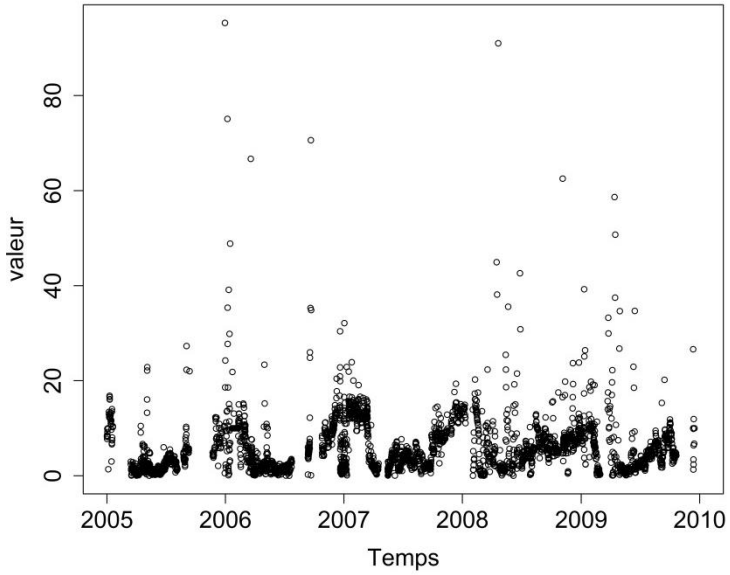


Figure A1.69. Représentation temporelle de la concentration en silicate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

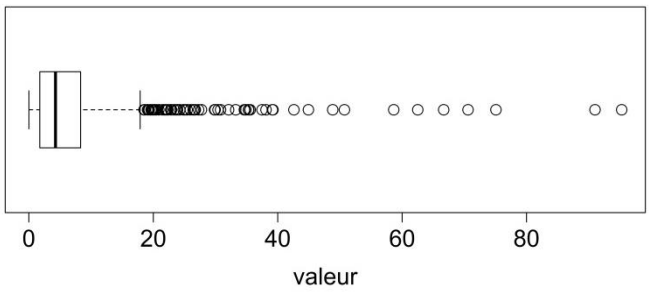


Figure A1.70. Boîte de dispersion de la concentration en silicate (µmol.L^{-1}) issue de la station MAREL-Carnot sur la période 2005-2009

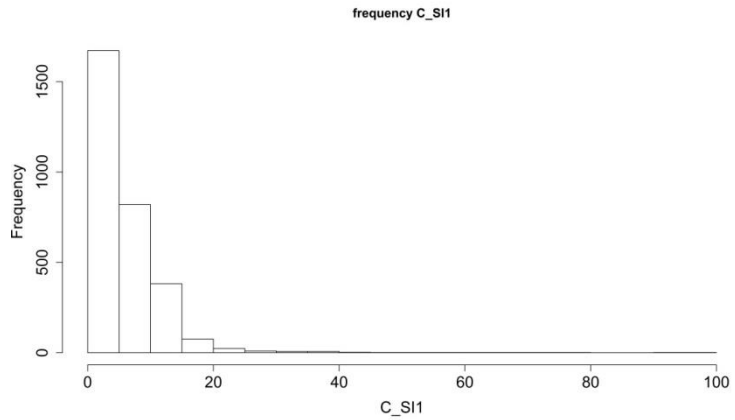


Figure A1.71. Histogramme en fréquence de la concentration en silicate issue de la station MAREL-Carnot sur la période 2005-2009.
*Distribution χ^2 selon le test de Pearson (p-value***).*

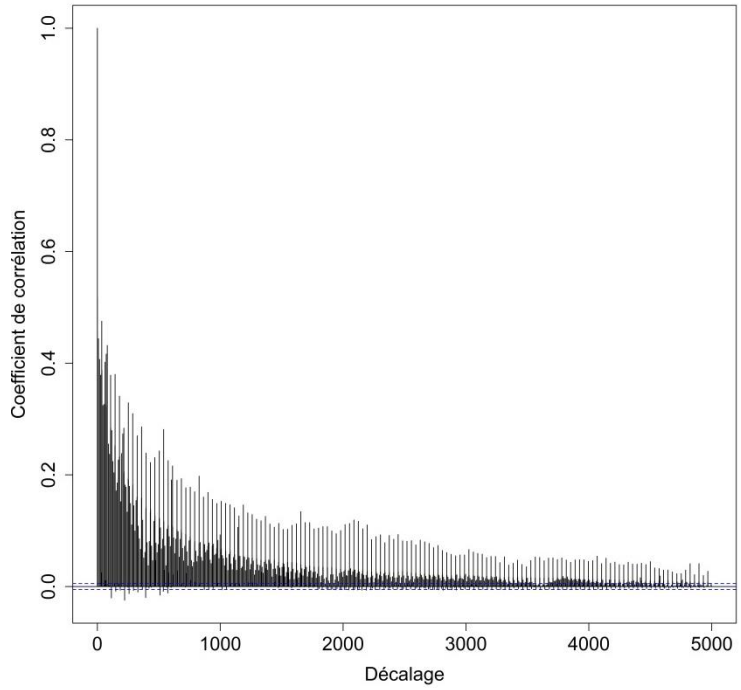


Figure A1.72. Corrélogramme de la concentration en silicate issue de la station MAREL-Carnot sur la période 2005-2009 avec un décalage allant de 1 à 5000 pas de temps.

Annexe 2 : Stationnarité et résultats de complétion

A2.1. Stationnarité

Tableau A2.1. Résultats des tests des stationnarités fortes et faibles sur les données de la station MAREL-Carnot sur la période 2005 à 2009.

Paramètre	Stationnarité faible				
	Critère 1		Critère 2		Critère 3
	Moyenne	Écart-type	Variance	Écart-type	Écart-type autocovariance
Oxygène dissous corrigé	8,25	0,43	2,86	0,88	0,09
Oxygène dissous non corrigé	10,00	0,55	4,44	1,37	0,13
Saturation en oxygène	93,89	2,18	168,67	51,39	9,56
Fluorescence	1,38	0,36	7,21	5,02	0,62
pH	8,36	0,16	27,05	0,09	0,02
Salinité	33,43	0,13	0,79	0,17	0,08
Conductivité	50,93	0,18	1,47	0,31	0,15
Température de l'eau	12,65	1,49	20,86	6,14	0,17
Température de l'air	12,12	1,45	28,98	6,55	1,44
Hauteur d'eau	4,91	0,03	4,79	0,12	2,96
Vitesse du vent en moyenne	10,14	1,98	29,49	3,57	4,84
Vitesse du vent en rafale	10,14	1,98	29,49	3,57	4,84
Direction du vent	184,7	13,81	9074,35	1060,97	1628,60
P.A.R.	292,20	48,72	217764,90	43584,34	87602,25
Turbidité	12,31	2,42	203,61	56,34	20,78
Concentration en Nitrate	18,24	2,98	262,21	50,67	56,05
Concentration en Phosphate	2,03	0,46	41,03	10,22	5,42
Concentration en Silicate	5,88	1,53	39,51	10,96	6,22

A2.2. Résultats de la DTW

Rappel :

E1. La requête R et la série x non bruitées : correspond au test de la série brute de la hauteur d'eau. Les signaux sont rarement aussi lisses, on peut donc assimiler ce signal à un signal filtré.

E2. Requête R bruitée uniquement : ce test peut être assimilé à un filtrage de la série complète x. La requête ne l'étant pas pour conserver la variabilité existante avant les données manquantes.

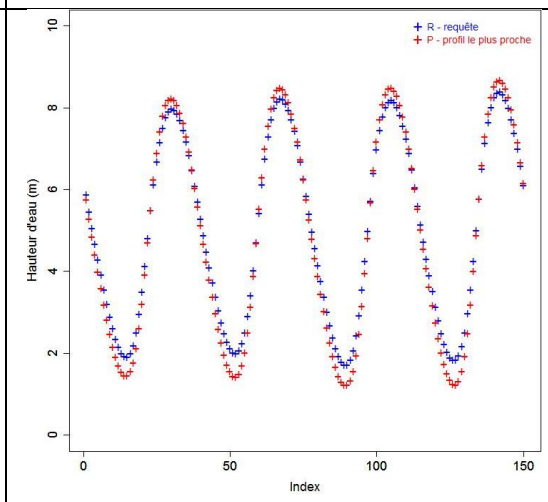
E3. Série x bruitée uniquement : pour celle-ci, seule la requête n'est pas bruitée. Par analogie, on peut dire qu'un filtre a été appliqué sur la requête pour améliorer les chances de correspondance.

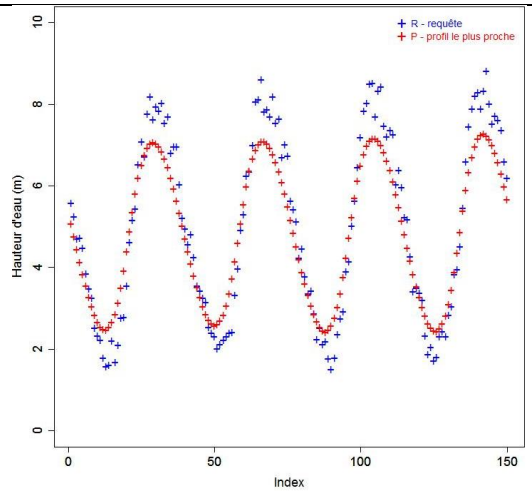
E4. Requête R et série x bruitées : cette expérience équivaut à tester un cas réaliste de signal totalement bruité.

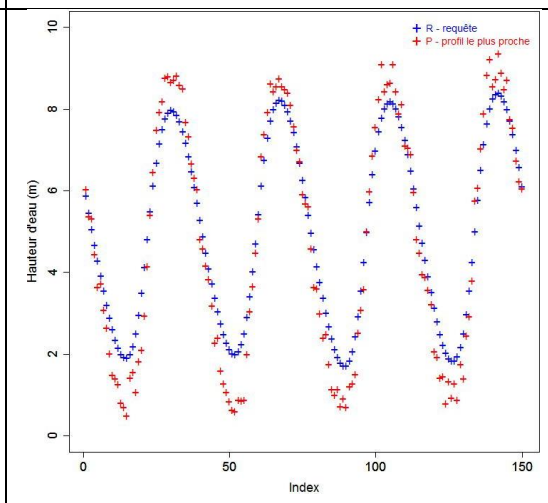
Tableau A2. 2. Résultats du calcul du coefficient de détermination, de similarité et de l'erreur de déformation moyenne pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur les données situées après les valeurs manquantes entre la requête R et du profil P le plus le plus proche.

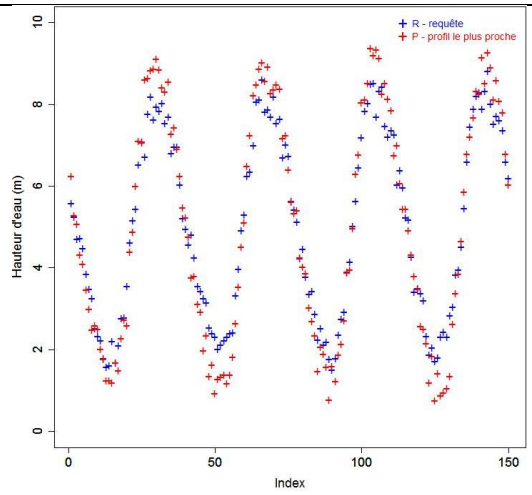
Expérience sur x_{Ap}	R^2	$Sim(R,P)$	$Err(R,P)$
E1	0,99***	0,79	0,06
E2	0,96***	0,65	0,13
E3	0,99***	0,64	0,14
E4	0,97***	0,69	0,12

Figure A2.1. Représentation, pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur les données situées après les valeurs manquantes entre la requête R (en bleu) et du profil P le plus le plus proche (en rouge).

<table>
<tr><th></th><th>Requête Brute</th><th>Requête Bruitée</th></tr>
<tr><td>Ref Brute</td><td> 

E1</td><td> 

E2</td></tr>
<tr><td>Ref Bruitée</td><td> 

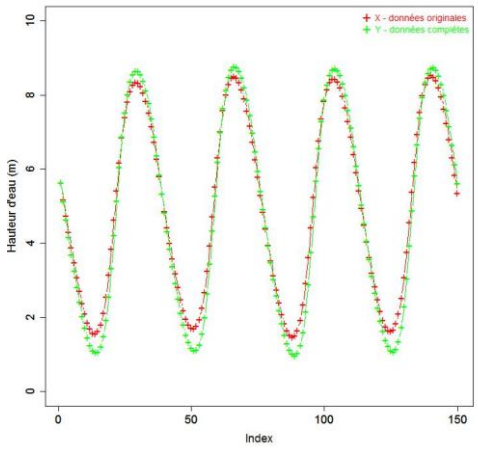
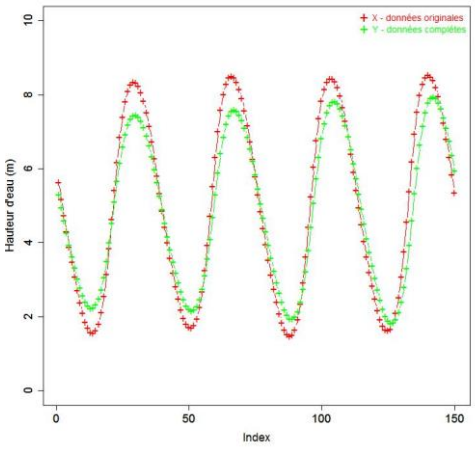
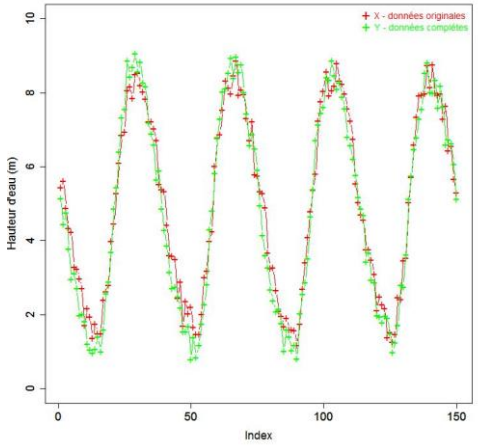
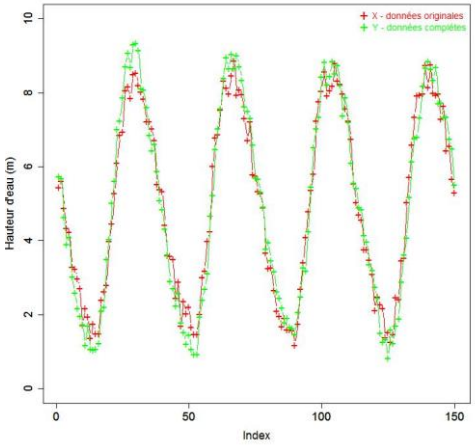
E3</td><td> 

E4</td></tr>
</table>

Tableau A2.3. Résultats du calcul du coefficient de détermination,de similarité et de l'erreur de déformation moyenne pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur les données situées après les valeurs manquantes entre les données originales ***X*** *et les données complétées* ***Y*****.**

Expérience sur x_{Ap}	R^2	$Sim(X,Y)$	$Err(Y,X)$
E1	0,995***	0,779	0,067
E2	0,961***	0,687	0,113
E3	0,970***	0,727	0,094
E4	0,968***	0,726	0,092

*Figure A2.2. Représentation, pour chacune des quatre expériences, du meilleur résultat après utilisation de l'appariement élastique sur les données situées après les valeurs manquantes entre les données originales **X** (en noir) et les données complétées **Y** (en vert).*

	Requête R Brute	Requête R Bruitée
Profil P Brut	 E1 : X vs Y	 E2 : X vs Y
Profil P Bruité	 E3 : X vs Y	 E4 : X vs Y

A2.3. Complétion multi-conjointe de la température de l'eau

Le plus proche voisin

Tableau A2. 4. Résultats des trois critères pour la complétion de la température de l'eau par la méthode du plus proche voisin.

R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
0,016 (0,129)	0,59	0,27

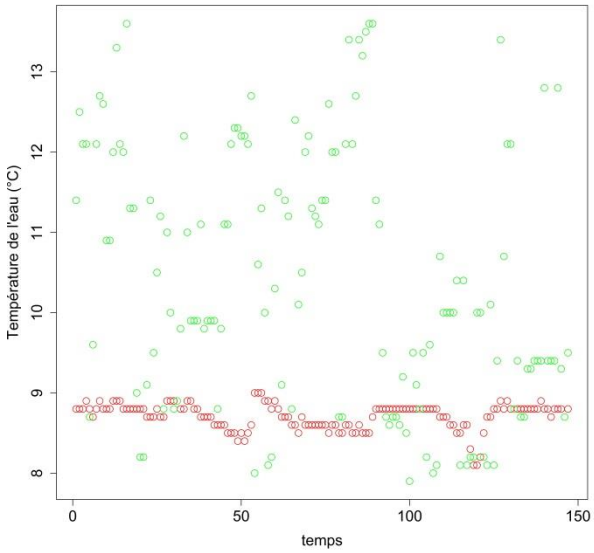


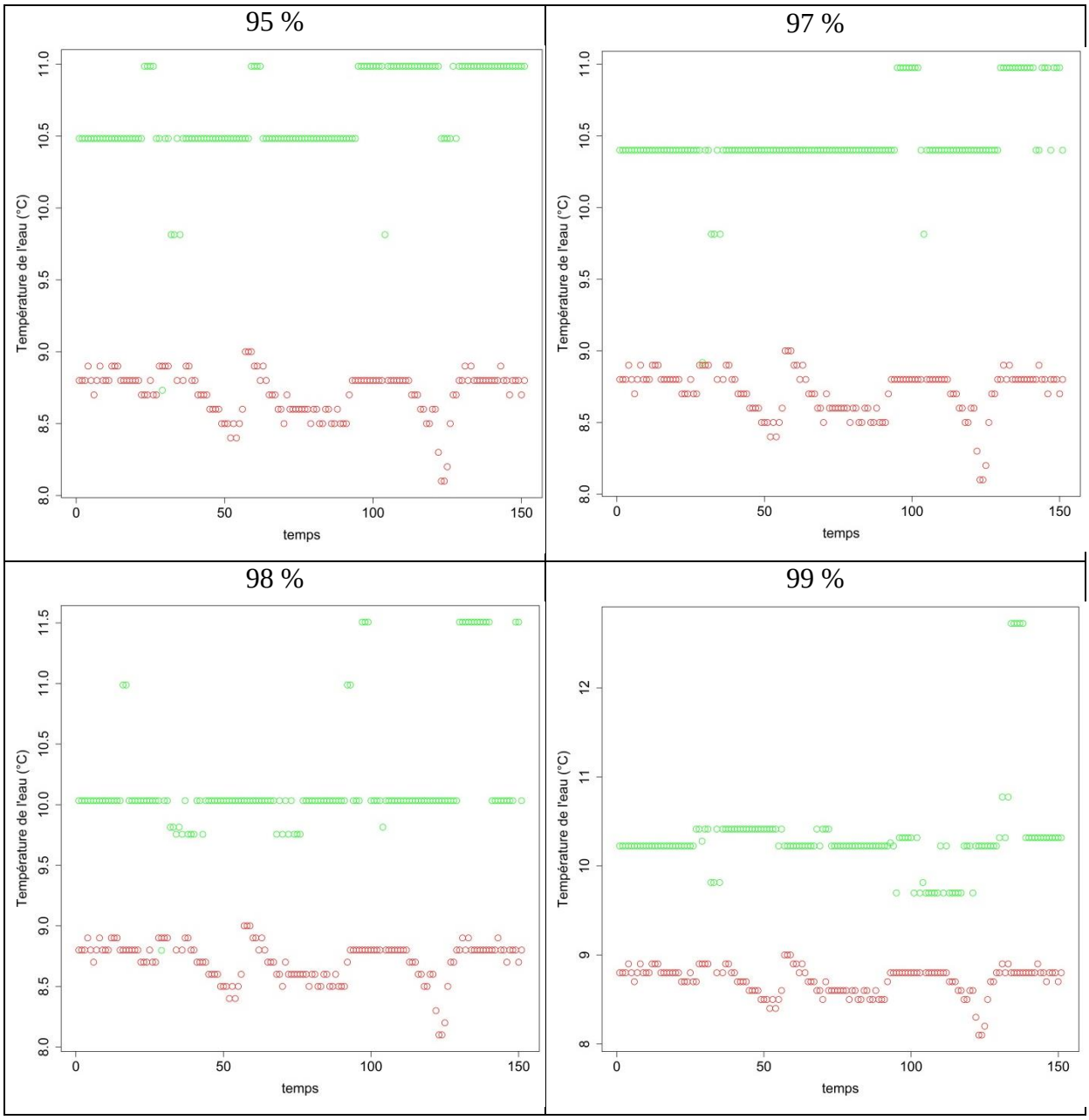
Figure A2.3. Représentation des données supprimées de la température de l'eau (en rouge) et du remplacement de celles-ci par leur plus proche voisin (en vert)

A2.4. Imputation par voisinage dans l'espace D-1 réduit par classification non supervisée

Tableau A2. 5. Résultats des trois critères pour la complétion de la température de l'eau par l'imputation par voisinage dans l'espace D-1 réduit par classification non supervisée.

Variance expliquée	R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
0,95	0,03 (0,03)	0,46	0,23
0,97	0,03 (0,04)	0,49	0,21
0,98	0,19 (0,02)	0,58	0,18
0,99	0.007 (0,29)	0,55	0,19

Figure A2.4. Représentation, pour chaque pourcentage de variance expliquée, des données supprimées de la température de l'eau (en rouge) et du remplacement de celles-ci par l'imputation par voisinage dans l'espace D-1 réduit par classification non supervisée (en vert).



A2.5. Imputation par voisinage dans l'espace $N_p \times D$ d'une base réduite par classification non supervisée

Tableau A2.6. Résultats des trois critères pour la complétion de la température de l'eau par l'imputation par voisinage dans l'espace $N_p \times D$ réduit par classification non supervisée

Variance expliquée	R^2 (p value)	$Sim(Y,X)$	$Err(Y,X)$
0,99	0,016 (0,12)	0,84	0,07

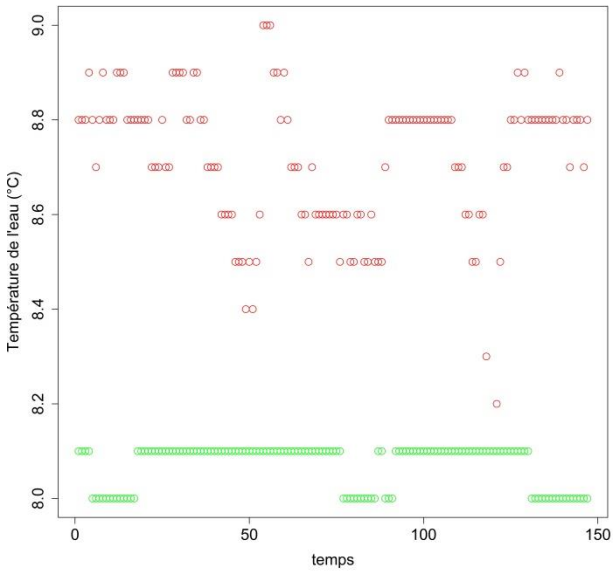


Figure A2.5. Représentation, pour une variance expliquée de 0,99, des données supprimées de la température de l'eau (en rouge) et du remplacement de celles-ci par l'imputation par voisinage dans l'espace D-1 réduit par classification non supervisée (en vert).

Annexe 3 : Tableaux des coefficients de corrélations entre paramètres et états

Tableau A3. 1. Coefficients de corrélations entre les paramètres et les états $c_i(2005)$ déterminés à partir d'une classification non supervisée sur les paramètres non corrélés (NC) issus de la station MAREL-Carnot sur l'année 2005 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 par état : seuils arbitraires).

État	$c_1(2005)$	$c_2(2005)$	$c_3(2005)$	$c_4(2005)$	$c_5(2005)$	$c_6(2005)$	$c_7(2005)$
Couleur	rouge	vert	bleu	cyan	rose	jaune	gris
C_NI1	-0,06	0,16	**-0,30**	-0,15	**0,64**	**0,21**	-0,19
C_O21	0,03	0,26	**-0,28**	**0,32**	-0,11	**0,28**	0,00
C_PO1	**0,48**	-0,05	-0,04	-0,24	-0,07	-0,05	-0,07
C_SI1	**0,18**	**0,32**	-0,01	**-0,32**	-0,01	0,15	-0,14
CSAL1	0,02	**-0,36**	0,15	0,13	0,13	-0,11	0,12
E_LU1	-0,01	-0,14	-0,11	-0,08	0,07	-0,12	**0,72**
E_TU1	-0,01	**0,45**	-0,17	-0,22	-0,09	0,01	-0,13
E_VVR	0,03	**0,36**	-0,13	-0,15	-0,08	-0,01	-0,10
ETCO1	0,04	-0,25	**0,48**	-0,22	0,13	**-0,27**	0,16
XMAHH	0,00	0,00	-0,02	-0,05	0,00	0,02	0,06

Tableau A3.2. Coefficients de corrélations entre les paramètres et les états $c_i(2006)$ déterminés à partir d'une classification non supervisée sur les paramètres non corrélés (NC) issus de la station MAREL-Carnot sur l'année 2006 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 par état : seuils arbitraires).

État	$c_1(2006)$	$c_2(2006)$	$c_3(2006)$	$c_4(2006)$	$c_5(2006)$	$c_6(2006)$	$c_7(2006)$
Couleur	rouge	vert	bleu	cyan	rose	jaune	gris
C_NI1	**-0,38**	0,26	**-0,35**	0,12	**0,35**	**-0,29**	**0,42**
C_O21	-0,23	**0,78**	0,08	-0,08	0,14	-0,06	-0,16
C_PO1	0,00	-0,04	-0,14	**0,86**	-0,05	-0,07	-0,11
C_SI1	0,06	0,21	-0,26	0,01	0,16	-0,21	0,01
CSAL1	0,15	-0,06	**0,32**	-0,03	**-0,55**	0,20	-0,23
E_LU1	-0,08	-0,18	-0,11	-0,05	-0,09	**0,75**	-0,16
E_TU1	**-0,24**	0,15	-0,07	0,01	**0,60**	-0,12	0,05
E_VVR	-0,05	-0,26	-0,26	0,08	0,14	-0,05	**0,36**
ETCO1	**0,40**	**-0,61**	**-0,31**	0,03	-0,16	0,10	-0,01
XMAHH	-0,01	0,00	-0,02	0,00	0,01	0,00	0,02

Tableau A3.3. Coefficients de corrélations entre les paramètres et les états $c_i(2007)$ déterminés à partir d'une classification non supervisée sur les paramètres non corrélés (NC) issus de la station MAREL-Carnot sur l'année 2007 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 par état : seuils arbitraires).

État	$c_1(2007)$	$c_2(2007)$	$c_3(2007)$	$c_4(2007)$	$c_5(2007)$	$c_6(2007)$	$c_7(2007)$
Couleur	rouge	vert	bleu	cyan	rose	jaune	gris
C_NI1	-0,02	-0,07	-0,18	**0,57**	0,16	**-0,37**	0,10
C_O21	**0,66**	**-0,38**	-0,12	-0,11	-0,09	-0,09	0,08
C_PO1	-0,07	-0,03	-0,04	**0,45**	-0,06	-0,07	-0,06
C_SI1	**-0,31**	0,04	**-0,27**	-0,07	**0,66**	**-0,43**	**0,34**
CSAL1	0,04	**0,65**	0,03	-0,12	-0,23	-0,08	**-0,36**
E_LU1	0,11	-0,12	**0,70**	-0,10	-0,20	-0,18	-0,11
E_TU1	-0,12	-0,11	-0,08	-0,13	-0,05	-0,18	**0,76**
E_VVR	-0,16	-0,18	-0,05	0,18	-0,07	-0,03	**0,34**
ETCO1	-0,27	0,13	**0,37**	-0,14	**-0,53**	**0,63**	-0,28
XMAHH	-0,01	0,04	0,04	0,00	0,01	-0,05	-0,01

Tableau A3. 4. Coefficients de corrélations entre les paramètres et les états $c_i(2008)$ déterminés à partir d'une classification non supervisée sur les paramètres non corrélés (NC) issus de la station MAREL-Carnot sur l'année 2008 (les valeurs en gras correspondent aux corrélations les plus proches de 1 ou -1 par état : seuils arbitraires).

État	$c_1(2008)$	$c_2(2008)$	$c_3(2008)$	$c_4(2008)$	$c_5(2008)$	$c_6(2008)$	$c_7(2008)$
Couleur	rouge	vert	bleu	cyan	rose	jaune	gris
C_NI1	**0,55**	-0,18	-0,03	-0,08	-0,09	-0,24	**-0,29**
C_O21	0,25	**-0,32**	-0,02	**-0,32**	**0,33**	-0,08	**0,53**
C_PO1	-0,01	-0,07	-0,01	-0,09	-0,05	-0,02	-0,08
C_SI1	0,06	-0,09	0,02	-0,10	**0,77**	-0,07	-0,11
CSAL1	-0,13	0,14	**-0,26**	0,23	0,10	0,15	0,16
E_LU1	-0,13	-0,15	-0,05	-0,21	0,02	**0,66**	-0,01
E_TU1	-0,12	-0,14	**0,63**	-0,14	-0,09	-0,11	-0,18
E_VVR	-0,15	-0,21	**0,33**	0,05	0,00	-0,07	-0,12
ETCO1	**-0,50**	**0,40**	-0,02	0,27	-0,06	**0,33**	0,00
XMAHH	-0,04	**-0,38**	-0,03	**0,38**	0,00	0,06	-0,02

Annexe 4 : Article accepté dans “*IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2013)”*

K. Rousseeuw, É. Caillault Poisson, A. Lefebvre, D. Hamad, 2013. Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2013)*, pp. 3962-3965, Melbourne, 21-26 July 2013. http://dx.doi.org/10.1109/IGARSS.2013.6723700

Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling

K. Rousseeuw[1,2], É. Poisson Caillault[2], A. Lefebvre[1], D. Hamad[2]

Abstract

The paper deals with a monitoring system combining K-means classifier and one Hidden Markov Model in order to detect phytoplankton blooms and to understand their dynamics. The states of the Hidden Markov Model and codebook symbols are computed without *a priori* knowledge thanks to K-means algorithms. The system is tested on database signals from the Marel-Carnot station that registers water characteristics at high frequency resolution. The experiments show that, when the states are set to two, these correspond to phytoplankton productive and non-productive periods. Moreover, when states are set to five, these correspond to the dynamics of phytoplankton blooms.

Index Terms

Monitoring, phytoplankton bloom, K-means, HMM, high frequency resolution.

I. Introduction

In the framework of coastal and river water quality assessment and management, the phytoplankton plays an important role as indicator of short-term and long-term changes in water quality. This is because the phytoplankton cells are capable of integrating natural and human induced disturbances, by changing their physiology. Thus, it is important to prevent and early detect phytoplankton blooms and understand their physical and nutrients conditions of outbreak [1].

For this purpose, many fixed buoys and ferry boxes were implemented for water monitoring. These systems allow characterizing water quality by several measurements at high frequency resolution: temperature, salinity, turbidity, Chlorophyll a, etc. This leads to the generation of a large amount of data signals which have to be automatically and finely analyzed in order to achieve an effective monitoring system. Such system will be able to characterize the physico-chemical and biological conditions at high resolution during phytoplankton blooms. In this context, we propose a monitoring system based on temporal data clustering and modeling.

In the last years, many research efforts concern time series clustering [2]. Time series could be considered if they share similar shapes or similar temporal models such as Hidden Markov Model HMM [3]. In [4][5][6], the authors combine HMM and clustering. The clustering is used to segment the time signals and to reduce the computational complexity. In this paper, the phytoplankton dynamic is modeled by a HMM and, two K-means algorithms are used: the first gives the K states of HMM and the second is used to generate the codebook symbols.

The proposed monitoring system is experimented on the Marel-Carnot station (automatic monitoring network) located in the harbor of Boulogne sur Mer (France) which records seventeen water characteristics at high frequency resolution (http://www.ifremer.fr/difMarelCarnot/).

The following section II presents the monitoring system architecture. And the section III validates the clustering part of the system on these data for the period 2005-2008 and highlights the use of HMM to integrate time modeling.

II. Monitoring system architecture

The monitoring tool is designed in order to answer the following questions:
- is the system able to detect phytoplankton productive period and/or blooms?
- in case the database signals is partitioned into 5 classes, do they correspond to the dynamics of phytoplankton: no productive, pre-bloom, bloom, post-bloom, and rare events?
- is it possible to forecast phytoplankton blooms?

A. General Scheme of the Monitoring System

The general scheme of the monitoring system is illustrated in Figure 1. It includes data pre-treatment level, clustering and HMM levels.

The step of unsupervised classification by K-means will allow to segment temporal data in two or more classes to answer at the previous requirements as well as to build the HMM parameters. Then elaborated HMM is used to predict a new or future potentially phytoplankton bloom.

This work is funded by IFREMER and the Artois-Picardie Water Agency with a PhD grant. The IFREMERś Department REM and its unit Research and Technological Development have operational control for servicing the MAREL system.

[1] IFREMER Centre Manche - Mer du Nord, BP 699, FR-62321 BOULOGNE SUR MER Cedex.
[2] LISIC - Université Lille Nord de France - ULCO, BP 719, FR-62228 CALAIS

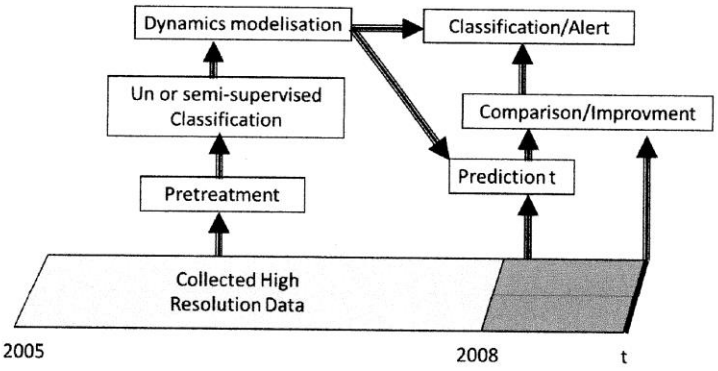


Fig. 1. General scheme of the monitoring system

K-means algorithm
Input: K, N data points
Output: K centers, label
1. Initialize the K centers
if $N < 20000$ select K distinct centers randomly
Otherwise
Partition the data in $n = 10$ subsamples of $nP = 20000$ pts
Compute K-means on each subsample
Select the K clusters centers from the best partition
according within-cluster sum of squares similarity
2. Decide the class memberships of the N data points by
assigning them to the nearest cluster center.
3. Re-estimate the k cluster centers, by assuming
the new memberships found
4. If none of N point changed membership in the last
iteration, Otherwise goto 2.

TABLE I
OUTLINE OF FAST K-MEANS ALGORITHM

B. Data Preprocessing and Clustering

Data signals from the various sensors are processed in order to respect the sensor ranges and to align the different acquisition time of all sensors. When one of the measurements do not meet the sensor range, this measure is removed. The time alignment is obtained by a simple time scaling (duration: 20 minutes) and for nutrients measured each 12 hours value are duplicated. In case of sensor failure, its signals are not retained and therefore, its related attribute values are not considered in the clustering process.

The unsupervised classification method employed is performed by K-means, using the Hartigan-Wong algorithm [7]. The principle is described in Table I. In the context of large database ($N > 20000$) it is more appropriate to partition the dataset into n subsamples ($n = 10$) in order to speed up the convergence process of K-means algorithm.

C. Building HMM Representation

Unsupervised classification used does not take into account the temporal dimension but can discover here similar behavior states (=clusters) that we expect to assimilate with the second question. We assume that Marel data has Markov property and may be view as the result of a probabilistic walk along the states obtained by K-means. So we use a first order Hidden Markov Model whose the number of states $S = (S_1, S_2, \ldots, S_K)$ correspond to the different clusters of the K-means. The method used is shown schematically in Figure 2. The HMM noted $\lambda = K, K2, A, B, \Pi$ is defined with 3 computed sets of probabilities from K-means algorithm:

1) Π of size K, defines the initial probability distribution. There is no information to determine the state that will be predominant during data acquisition. Therefore, the initial probability vector is equiprobable.
2) A of size $n * n$, defines the transition matrix, a_{ij} is the conditional probability $P(S_t = S_i | S_{t-1} = S_j)$. So the number of times that one pass from a cluster i to cluster j is estimated and then matrix A is normalized in row (for each state).
3) B of size $n * K2$, defines the emission probability. We consider that it is not feasible to define HMM states by exhaustively directly enumerating values but can be defined by a symbol. Data space is so represented by a codebook defined by a second K-means method with a fixed number of groups $K2$ by user. So the probability b_{ij} is the emission probability $P(CG_t = CG_i | S_t = S_j)$. To compute this matrix, the number of times that one point (Observation O_t) is both in a

state S_j (first classification) and in a cluster defined by the centers CG_i(second classification) is counted and then B normalized in row.

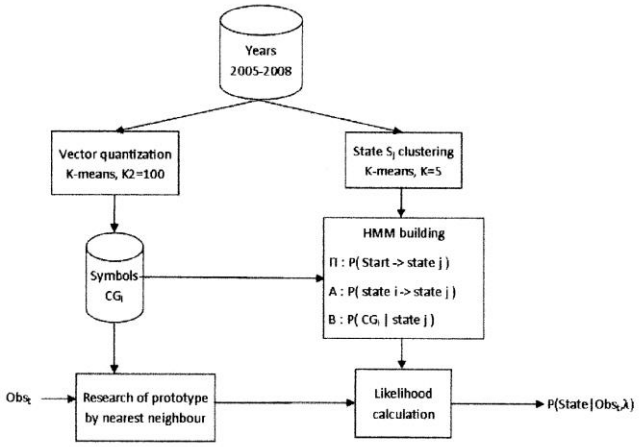


Fig. 2. unsupervised HMM

When new data denoted Obs_t are collected with the Marel Carnot station, they should be classified quickly and in the most optimal possible way. Each cluster center CG_i corresponds to a symbol. Each new data is associated to its nearest symbol by a 1-nearest neighboring algorithm. The Viterbi algorithm, using the built HMM representation λ and the sequence of symbols (current symbol and past symbols), calculated the likelihood that this new data belongs to one of the five environmental conditions.

III. Experimentation and discussion

A. Data Presentation and Protocols

The Marel-Carnot station registers 17 data signals: 14 water characteristics every 20 minutes and 3 nutrients levels every 12h. The sensor ranges are adapted to the ecosystems of Boulogne-sur-Mer. To build the state partition and the HMM representation, the period 2005-2008 is considered. After a sensor time alignment, signal database for 2005-2008 contains 105 192 points with 17 dimensions. Only half data (48 157) have no missing attributes (due to sensor default). To reduce missing data, a moving average on one week (temporal scale of a bloom), is applied and so the number of completed data is 84 614 points. To test the time modeling, data from 2009 will be tested with the same pretreatments protocol.

In this paper, the attributes used for the clustering process are only nutrients (Nitrate, Silicate, Phosphate), turbidity and corrected dissolved oxygen. Fluorescence signal is not take into account but are used to validate the clustering results since we have no ground truth. And the others parameters are used only to explain the obtained states.

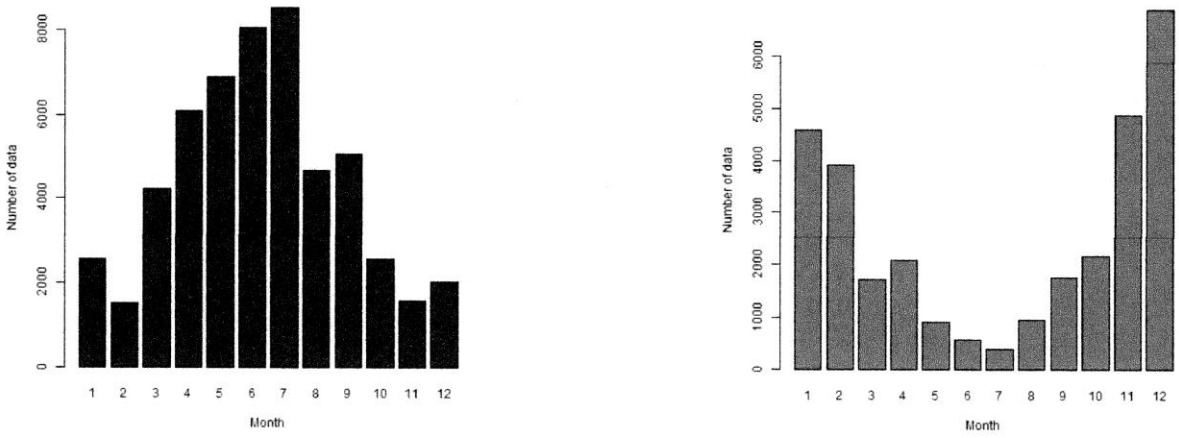
HMM representation is built with the results of two K-means. The first K-means used to segment in K states, first K=2 and K=5 class number is applied. The second K-means used K2=100 to generate the codebook symbols for HMM. The classification test of the system is cut into 2 parts according the questions of the section II.

B. States Characterization

To answer the first question on how to discriminate between the productive and the non productive periods, a K-means with K=2 classes is performed. The distribution of data obtained in each class is shown in Figure 3. We can conclude that the system is able to detect this two environmental states. Class A corresponds to the March-October months (productive period) and class B to the November-February (non productive period) as cited in the EU Water Directive Framework [1].

Changes in the fluorescence signal (characterizing the phytoplankton biomass) for the year 2008, confirm also this result (Figure 4). In black are represented data whose one attributes is missing and do not contribute in the clustering process.

For the next question, ideally, the environmental states would be: pre-bloom, bloom, post-bloom, no productive and rare events. The K-means is performed with K=5 classes. Then it is possible to label 2 environnemental states within a given bloom (Figure 5). These two states correspond to a succession of two different community of phytoplankton. The class B and D are these two states (classes are labeled from letter A to E: A in the bottom and E in the top of Figure 5 (a)). This is validated by the temporal changes of the fluorescence (Figure 5 (b)) since a high level of fluorescence means a productive period of phytoplankton. The labeling of the others classes is not possible with the intuitive way, the analysis of the relation between the classes and the statistical of each parameters will permit the labeling biologically correct.



(a) class A (b) class B

Fig. 3. Distribution of data in each class

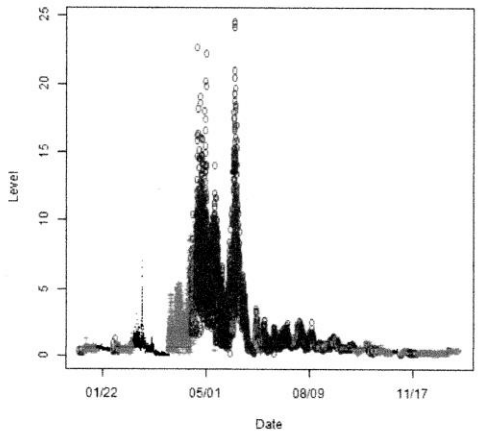


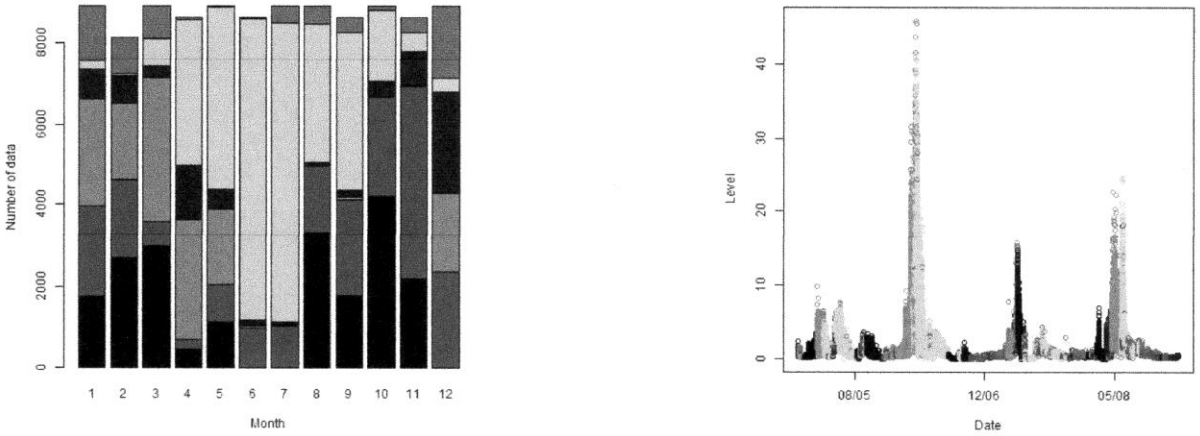
Fig. 4. Temporal change of the fluorescence for the year 2008

C. Time Modeling

HMM parameters λ ($K = 5$ states, $K2 = 100$ symbols, A, B, Π) is computed from the 2 K-means partitions from the data signals in the period 2005-2008. First in order to test the robustness of the training system, each trained period is targeted by the Viterby algorithm. Then the year 2009, never used in HMM building, will be tested in order to see the generalization performance of the system to predict blooms.

Table II shows the percentage of same labeling by the clustering methods and the time modeling for each year. The results are very encouraging. Actually missing data are not taken into account for the moment and so they may disturb the probabilities of transitions. Note that the years 2006 and 2007 are very corrupted by the sensor defaults.

Figure 6 illustrates the state labeling on the fluorescence signal for year 2009. Unlike years 2005-2008, a unique bloom (class B) is detected in the year 2009, so the system is able to detect if one species is dominant. According to environmental point of view, 2009 is therefore a particular year.



(a) (b)

Fig. 5. (a) The distribution of data in each class (black color is not a real class but correspond to missing data, then class is labeled in order A to E, E at the top) and (b) Labeling of the fluorescence from the year 2005 to 2008

Year	2005	2006	2007	2008	2009
Recognition Rate	89.16	74.54	71.64	92.70	87.68

TABLE II
PERCENTAGE OF SAME LABELING BY K-MEANS OR TIME MODELING

IV. CONCLUSION

In this paper, a new HMM-based K-means time series modeling algorithm is proposed. Here two K-means process are used: one to build HMM states and the second to build codebook symbols without any *a priori* knowledge. The system was experimented on the database of Marel-Carnot station that registers water characteristics at high frequency resolution. The monitoring system, with this combination K-means and HMM-representation, can detect and forecast phytoplankton blooms. The results are very hopeful. In order to perfect the system, the next step is to find an optimal way to set the HMM parameters

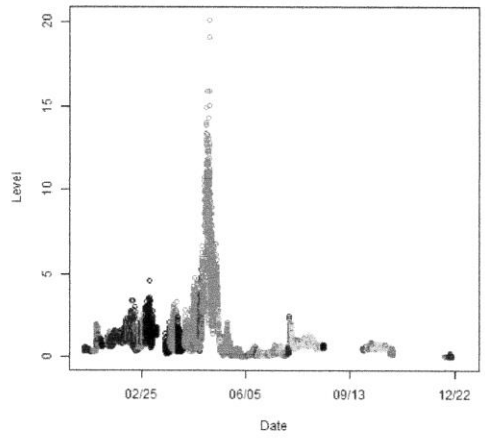


Fig. 6. Labeling of the fluorescence for 2009

(number of states, numbers of symbols) and the taking account of the missing data in the clustering process. Another perspective is to apply an semi-supervised classification scheme directed by other systems able to detect phytoplankton composition like a flow cytometer.

REFERENCES

[1] *Directive 2000/60/EC of the European Parliament and of the Council. Establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327/1.*, 2000.

[2] T. Warren Liao, "Clustering of time series data-a survey," *Pattern Recogn.*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005.

[3] Lawrence R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[4] Xi Shao, Changsheng Xu, and Mohan S. Kankanhalli, "Unsupervised classification of music genre using hidden markov model," in *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan*. 2004, pp. 2023–2026, IEEE.

[5] Li-Li Wei and Jing-Qiang Jiang, "A hidden markov model-based k-means time series clustering algorithm," in *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, oct. 2010, vol. 3, pp. 135 –138.

[6] Tim Schlüter and Stefan Conrad, "Hidden markov model-based time series prediction using motifs for detecting inter-time-serial correlations," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, New York, NY, USA, 2012, SAC '12, pp. 158–164, ACM.

[7] J.A. Hatrigan and M.A. Wong, "A k-means clustering algorithm," Yale University, New Haven, Connecticut, USA, 1979, vol. 28, pp. 100–108.

Annexe 5 : Article accepté dans "*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS 2014)"*

K. Rousseeuw, E. Poisson-Caillault, A. Lefebvre, D. Hamad. Hybrid Hidden Markov Model for Marine Environment Monitoring. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS 2014)*. DOI: 10.1109/JSTARS.2014.2341219.

Hybrid Hidden Markov Model for Marine Environment Monitoring

Kévin Rousseeuw, *Ph.D. Student Member, IEEE,* Émilie Poisson-Caillault, Alain Lefebvre, and Denis Hamad

Abstract

Phytoplankton is an important indicator of water quality assessment. To understand phytoplankton dynamics, many fixed buoys and ferry boxes were implemented, resulting in the generation of substantial data signals. Collected data are used as inputs of an effective monitoring system. The system, based on unsupervised Hidden Markov Model (HMM), is designed not only to detect phytoplancton blooms but also to understand their dynamics. HMM parameters are usually estimated by an iterative Expectation-Maximisation approach. We propose to estimate HMM parameters by using spectral clustering algorithm. The monitoring system is assessed on database signals from MAREL-Carnot station (Boulogne-sur-Mer, France). Experiment results show that the proposed system is efficient to detect environmental states such as phytoplankton productive and non productive periods without *a priori* knowledge. Furthermore, discovered states are consistent with biological interpretation.

Index Terms

Hybrid Hidden Markov Model, marine water monitoring, Phytoplankton blooms, spectral clustering.

I. Introduction

IN the framework of coastal and river water quality assessment and management, phytoplankton plays an important role as an indicator of short- and long-term changes in water quality. Indeed phytoplankton cells are capable of integrating natural and human induced disturbances by changing their physiology. The Marine Strategy Framework Directives (MSFD) underlined the importance to prevent and early detect phytoplankton blooms (harmful, and non-toxic as well), and to understand their physical and outbreak nutrient conditions [1], [2].

Advances in monitoring systems arise from the evolution of computer technology, the availability of effective low-cost sensors, and the deployment of remote sensing generating multidimensional signals. Mathematical models and powerful tools are therefore needed to effectively monitor complex systems with multivariate time series. Recently, machine learning approaches are used to detect harmful algae blooms thanks to available information on cell taxonomy. Such systems are trained from global observation by remote sensing (Support Vector Machine [3], Probabilistic Neural Networks [4]) or local observations like flow cytometry datasets (Radial Basis Function Neural Network [5]).

To monitor phytoplankton dynamics, many marine instrumented stations, fixed buoys and ferry boxes, were implemented with High Frequency (HF) multi-sensor systems. Often, collected data are incomplete due to problems of sensor readings, communication failures and the lack of environmental information (taxa). Accordingly, unsupervised or semi-supervised machine learning approaches are suitable for phytoplankton dynamics monitoring.

This paper focuses on how to build a marine monitoring system based on HF multisensor signal collected from MAREL-Carnot station (IFREMER, Boulogne-sur-Mer, France) in an unsupervised context. This marine station measures physico-chemical and biological parameters every 20 minutes. A lack of information stops to set up a training database at high frequency. Indeed, no information are directly acquired by MAREL-Carnot station about phytoplankton taxonomic composition and local activities (e.g., dredging, opening dams). And, the resolution of complementary regional monitoring programmes in the area is too low (the objectives are differents).

Hidden Markov Model, noted HMM, is a well-adapted stochastic signal model to represent time series dynamics. The success of HMM in speech and handwriting recognition [6] leads to their application in marine monitoring. HMM approaches are based on static parameters defined by states and symbols, and dynamic parameters related to state transition and observation symbol probabilities. For instance, in speech recognition, a word is a sequence of phonemes (states) structured by transition probabilities where each phoneme is considered to be a spectral fingerprint (symbols) with some occurrence probabilities.

HMM building needs to estimate not only the number of states but also the characteristics of each of them. Commonly, HMM parameters are learned with labeled database or fixed with *a priori* information. Here, we address phytoplankton bloom

This work was supported in part by IFREMER, in part by the Artois-Picardie Water Agency with a Ph.D. Grant, and in part by the DYMAPHY Interreg IV A 2 Mers Seas and Zee en program. The IFREMERs Department REM and its unit Research and Technological Development have operational control for servicing the MAREL system.

K. Rousseeuw is both with the LISIC Lab (LISIC : Laboratoire dInformatique Signal et Image de la Cte dOpale address: ULCO/LISIC, BP 719, FR-62228 Calais cedex) and with French Research Institute for Exploitation of the Sea (IFREMER, address: IFREMER Centre MancheMer du Nord, BP 699, FR-62321 Boulogne-sur-Mer, France) (e-mail : kevin.rousseeuw@gmail.com).

É. Poisson Caillault and D. Hamad are with LISIC lab (e-mail: emilie.caillault@lisic.univ-littoral.fr; denis.hamad@lisic.univ-littoral.fr).

A. Lefebvre is with IFREMER - Centre MancheMer du Nord (e-mail: Alain.Lefebvre@ifremer.fr).

Digital Object Identifier 10.1109/JSTARS.2014.2341219

forecasting issue using a hybrid HMM. The specific objective of this work is to design a system able to model phytoplankton dynamics from large database and no prior knowledge. For this purpose, a fully unsupervised HMM is built using spectral clustering algorithm to generate HMM symbols and states.

The paper is organized as follows. Section II describes the monitoring system and the proposed hybrid HMM with three parts. Part II-A discusses about usual unsupervised techniques to build HMM, and spectral clustering approach is argued to estimate HMM static parameters (states and symbols). Part II-B details HMM symbol generation by a self tuning fast K-means proposed algorithm. Part II-C defines HMM state generation by spectral clustering algorithm. Section III describes protocol of experimentations and collected data used from the IFREMER MAREL-Carnot station that registers water characteristics at HF resolution. First a fixed 2-state HMM is built in order to assess symbol and state generations thanks to an artificial labeling. Thus, our algorithms are compared with other machine learning techniques. Then, in section IV experiment results of a fully unsupervised N-state HMM are presented, and are related to examine biological expectations.

II. Monitoring system based on hybrid HMM

Fig. 1 presents the proposed monitoring system architecture. Data collected at high frequency resolution from 2005 to 2008 are first pretreated. Then, the clustering step is applied in order to find environmental states. The final step relies on temporal information between these states to develop a phytoplakton dynamics model. The built model is used to predict a new or forthcoming phytoplankton bloom, or specific states (classification/alert box in Fig. 1).

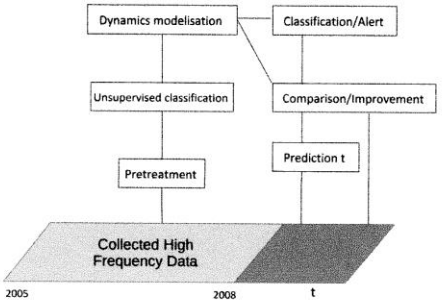


Fig. 1. General scheme of the monitoring system.

A. Hidden Markov Model

According to normal course of phytoplankton succession highlighted by Margalef [7] and Reynolds *et al.* [8] works, we assume that phytoplankton biomass is constrained by a high level of dependence among successive observations. Besides, it may be viewed as the result of a probabilistic walk along the environmental states. So let us see how to design one ergodic Hidden Markov Model to characterize the dynamics of phytoplankton blooms from physico-chemical and biological parameters in an unsupervised context.

A HMM noted $\lambda = \lambda(\mathbf{S}, \mathbf{V}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ is defined with 2 static sets $(\mathbf{S}, \mathbf{V})$ and 3 computed sets of probabilities $(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ that we recalled as follows [6]:

1) $\mathbf{S} = \{s_1, s_2, \ldots, s_N\}$ is the set of states with N the number of distinct states. For instance: non-productive period, pre-bloom, bloom, post-bloom and other rare events, like dam opening, factory or agricultural discharge. The number of states is generally set by expert people in relation with the applications, or automatically determined by penalized maximum likelihood criterion.
2) $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_M\}$ is the set of symbols and M is the number of distinct symbols. In the simpliest cases, observation symbols correspond to the system outputs. Environmental state is not characterized by a unique representative per state, and two system outputs can belong to several states. Thus, it is necessary to build a codebook of symbols by vector quantization [9]–[11]. So data space will be represented by this codebook $\mathbf{V}$ of M symbols.
3) $\boldsymbol{\pi} = \{\pi_i\}$ of size N, defines the initial probability distribution, $\pi_i = P(s(t=1) = s_i)\}$. There is no information about the state that will be predominant during data acquisition. *A priori* initial states are equiprobable.
4) $\mathbf{A} = \{a_{ij}\}$ of size $N \times N$, defines the transition matrix with $a_{ij} = P(s(t) = s_i | s(t-1) = s_j)$ the conditional probability. Therefore, the number of times that we move from a state s_i to state s_j is estimated, then $\mathbf{A}$ is normalized in row.
5) $\mathbf{B} = \{b_{ik}\}$ of size $N \times M$, defines the emission probability with $b_{ik} = P(\mathbf{v}(t) = \mathbf{v}_k | s(t) = s_i)$.

From a finite observation sequence without any labeled states, HMM symbols, transition and emission matrices [12] are adapted iteratively. Expectation-Maximisation approach (EM) is used with entropy criterion with Minimum Description Length

(MDL) constraint as Penalized Maximum Likelihood criterion [13]. Whatever the used criterion is (Bayesian Information Criterion and its derived), EM performance depends on initialization step, and can be time-consuming for large complex database. To avoid HMM iterative parameter estimate and the initialization step, we choose to use a spectral clustering approach to generate HMM state and symbol parameters from spatial information in one-pass algorithm.

Spectral clustering (SC) [14], [15] is a multi-cut method based on the eigen-decomposition of the Gram affinity matrix from the original dataset. Eigenvectors represent a new feature space where data are simply clustered by a K-means algorithm. It succeeds in clustering convex and non-convex distributed data. SC algorithm has been addressed for several applications: image segmentation, speech recognition, information retrieval, and so on [16]. Recently, algorithms have been developped to avoid their tuning requirements: to build the affinity function and to find the number of clusters. These steps are automatically completed using techniques, especially from [17], [18]. And, works [19], [20] allow to treat applications with a high volume of data.

The hybrid HMM building is schematically shown in Fig. 2. A step of vector quantization allows to extract HMM symbols. From these symbols, SC algorithm extracts HMM states. The HMM emission and transition probability matrices are then computed from the observed sequence. The transition matrix $\mathbf{A}$ is determined with the number of occurences moving from one state to another. $\mathbf{B}$ matrix corresponds to the number of times that observation $\mathbf{o}_t$ is both in a state s_i and in a symbol $\mathbf{v}_k$.

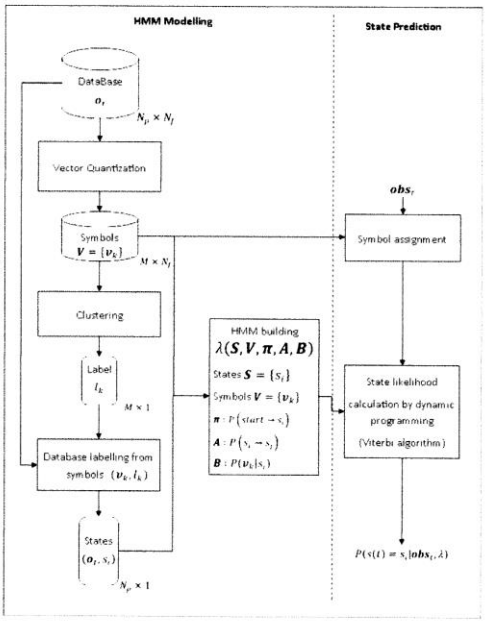


Fig. 2. Hybrid HMM building scheme. Dotted line separates HMM modelling on left side from state prediction of a new observation $\mathbf{obs}_t$ on right side.

When new data $\mathbf{o}_t$ is collected, it is associated with its nearest symbol. Viterbi algorithm [21], [22] is then applied to estimate its environmental state.

B. Symbol generation

MAREL-Carnot database consists of 26,280 $\times$ 19 parameters per year as from November 2004. To discover underlying states in this large database, instance selection is required. K-means algorithm is a well-adapted vector quantization method, and is popular for data clustering too [23]. The main idea is to build vector prototypes from a set of observations denoted $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_{N_p}\}$ of N_p data points preserving HF information. The fast K-means algorithm [24] is modified to obtain a self tuning K-means based on Hartigan-Wong algorithm [25], and on Elbow criterion: the number of clusters K is incremented until a fixed percentage of explained variance or a $Kmax$ number of retained prototypes (i.e. symbols) is met. The principle of this proposed algorithm, named Self Tuning Fast K-Means (STFKM), is described in Fig 3. The center initialization for large databases (number of points $Np > 20,000$) is very important here to speed up the process convergence. $Kmax$ is the maximum number of reduced points specified by the user or in the default case, the number of measures in the time series. $varExplained$ is the explained variance desired by the user, by default this number is set to 95 percent.

procedure STFKM($\mathbf{O}, Kmax, varExplained$)
  if $varExplained$ not defined **then**
    $varExplained$=0.95
  end if
  if $Kmax$ not defined **then** $Kmax$=nrow($\mathbf{O}$)
  end if
  Variable: k=1, vE=0;
  while $k < Kmax$ or $vE < varExplained$ **do**
    $k = k + 1$;
    Step 1: Initialization of k centers
    ..Cut Data in n subsamples of 20,000 points
    ..Compute K-means (K=k) on each subsample
    ..Select the k clusters centers from the best partition according $MS(within)/MS(between)$
    Step 2: Decide the class memberships of the N_p points by assigning them to the nearest center.
    Step 3: Re-estimate the k cluster centers, by assuming the new memberships found
    Step 4: If none of N_p points changed membership in the last iteration, Otherwise goto 2.
    Step 5: $vE = MS(between)/MS(total)$
  end while
  return k obtained centers
end procedure

Fig. 3. Outline of Self Tuning Fast K-Means algorithm noted STFKM. The variance, MS(.) for Mean Square defined the variance between or within groups, or the total.

C. State generation by spectral clustering

After STFKM procedure on the pretreated data, M symbols summarize the entire database. From these M symbols, N states are detected by unsupervised clustering. Each MAREL-Carnot physico-chemical parameter follows a stochatic, non-linear and non-stationary process (except sea-level), see section III. They have not-gaussian distributions, and environmental state characterisation is unknown. So, SC method is the best way to avoid some assumptions about data shape. SC is capable of classifying data which are connected, but which are not necessarily compact, or clustered within convex boundaries. Indeed, the key idea of SC is to transform the input data space into a new feature space where K-means clustering could be applied. The most typical method by Ng, Jordan *et al.* [14] is recalled in Fig. 4.

procedure SPECTRALCLUSTERING($\mathbf{O}$, K)
  Variable: $\mathbf{W}, \mathbf{D}, \mathbf{Lap}, \mathbf{X}, \mathbf{Y}, \mathbf{l}$
  Compute a Gram affinity matrix $\mathbf{W}_{M\times M}$ from $\mathbf{O}$
  $\mathbf{D} = diag(\mathbf{W})$; ▷ Degree matrix
  $\mathbf{Lap} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$; ▷ Laplacian matrix
  Select the K-largest eigenvectors $\mathbf{x}$ of $\mathbf{Lap}$;
  Form the matrix $\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_k] \in \mathbb{R}^M$ by stacking the eigenvectors in column
  Form the $\mathbf{Y}$ matrix from the row-normalisation of $\mathbf{X}$ thus $y_{ij} = w_{ij}/(\sum_j x_{ij}^2)^{1/2}$
  $\mathbf{l}$ =K-means($\mathbf{Y}$,K) ▷ each row of $\mathbf{Y}$ is a point
  Assign original Point $\mathbf{o}_i$ to the cluster l_i
  return label vector : $\mathbf{l}$
end procedure

Fig. 4. Spectral clustering algorithm

The number of clusters K in input of the SC algorithm and the way to build the Gram Affinity matrix $\mathbf{W}$ have both significant effects on the classification result. Gaussian kernel function is the most widely used function for constructing $\mathbf{W} = \{w_{ij}\}$ defined as:

$$w_{ij} = exp(-\frac{\|\mathbf{o}_i - \mathbf{o}_j\|^2}{2\sigma^2}) \quad (1)$$

The scaling parameter σ helps to sparse the matrix and tends to obtain an ideal case with a robust eigen-decomposition (i.e. in

TABLE I
LIST OF MAREL-CARNOT SIGNALS: ACRONYM, NAME AND MEASUREMENT FREQUENCY.

Acronym	Name	Frequency	RD	NC
E_TA	Air temperature	20 minutes	✓	
C_O21	Corrected dissolved oxygen	20 minutes	✓	✓
CSAL1	Salinity	20 minutes	✓	✓
CSAT1	Oxygen saturation percentage	20 minutes	✓	
E_CO1	Conductivity	20 minutes	✓	
E_LU1	P.A.R. Photosynthetically Available Radiation	20 minutes	✓	✓
E_O21	Non-corrected dissolved oxygen	20 minutes	✓	
E_PH1	pH	20 minutes		
E_TU1	Turbidity	20 minutes	✓	✓
E_VDM	Direction wind	20 minutes		
E_VVR	Gust wind speed	20 minutes	✓	✓
E_VVM	Average wind speed	20 minutes	✓	
ECHL1	Fluorescence	20 minutes		
EMAHH	Sea-level (measured)	20 minutes		
ETCO1	Water temperature	20 minutes	✓	✓
XMAHH	Sea-level (calculated)	20 minutes	✓	✓
C_PO1	Phosphate concentration	12 hours	✓	✓
C_NI1	Nitrate concentration	12 hours	✓	✓
C_SI1	Silicate concentration	12 hours	✓	✓

the ideal case, the first K eigenvalues are equal to one). However, a bad choice of σ brings an incorrect classification. Zelnik and Perona (ZP) [17], or Kong et al. [18] proposed a local scale parameter σ_i for each data $\mathbf{o}_i$ based on its neighborhood, instead of selecting a uniform scaling parameter σ. The ZP affinity matrix $\mathbf{W}$ is chosen with a z-neighborhood ($\mathbf{o}_{nz}$ the z^{th} neigborhood of the point $\mathbf{o}_i$):

$$w_{ij} = exp(-\frac{\|\mathbf{o}_i - \mathbf{o}_j\|^2}{2\sigma_i\sigma_j}) \text{ with } \sigma_i = \|\mathbf{o}_i - \mathbf{o}_{nz}\| \quad (2)$$

Many authors proposed to overcome the choice of the number of clusters K by analysing either eigenvalues magnitude (equal or nearest to one) or eigengap, or eigenvectors (see [18], [26], [27]). To select the number of states N for HMM topology, the eigengap method is used: it is the simpliest one to implement, and it has the lowest complexity.

From the spectral clustering of the M symbols $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_M\}$ issued from the STFKM step, we assign the observation data $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_{N_p}\}$ thanks to the label $s_i = l_k$ of their cluster center $\mathbf{v}_k$.

III. DATA AND FIXED 2-STATE HMM VALIDATION

One HMM is built according to the scheme described in the previous section from MAREL-Carnot multivariate marine signals in order to model the phytoplankton dynamics in the French Channel coast around Boulogne-sur-Mer. Data and their curves are available on the website (http://www.ifremer.fr/difMarelCarnot/) with authorisation request. These data are first presented, then the experiment validation protocol follows.

Without ground truth on the environmental states and in order to assess our system, we decide to create an automatic data labeling based on the monitoring sampling strategy for the EU Water Directive Framework (EU-WFD). Thus, data from March to October are labeled s_1, corresponding to the productive period (in terms of biomass production capacity), and the others s_2 for the non-productive period. Furthermore, this labeling will allow to compare our sytem with other machine learning algorithms.

A. Data presentation

MAREL-Carnot station registers 19 signals: 16 water characteristics every 20 minutes, and 3 nutrient levels every 12 hours. These signals are detailed in Table I. Collected data signals come from different sensors. They require pretreatments to respect sensor range, and sensor time alignment. In case of sensor failures, its measurements are not retained (all pH values and measured sea-level values are removed). The sensor ranges are adapted to Boulogne-sur-Mer ecosystem. Time alignment is obtained by a moving average technique on a small sampling rate (20 minutes). Nutrient parameters are duplicated to obtain the same time resolution. From this step, signal database for 2005-2008 contains 105,192 points in $\mathbb{R}^{19}$. Only half of the data

(48,157 points) have no missing values (due to sensor default). To reduce missing value, a moving average over one week (temporal scale of a bloom) is applied, leading to a completed data of N_p =84,614 points.

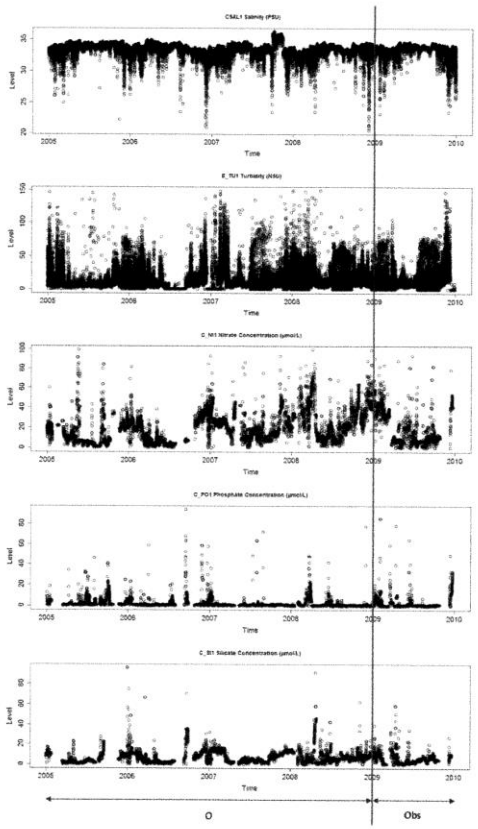


Fig. 5. Five completed MAREL-Carnot signals from 2005 to 2009 by mobile average: C_SAL1, E_TU1, C_NI1, C_PO1, C_SI1. Vertical line separates data into 2 subsets. Subset **O** corresponds to 2005-2008 period which is used for HMM building. Subset **Obs** relates to the year 2009 which is used for generalisation and accordingly does not participate in HMM building.

Fig. 5 illustrates 5 signals after this data completion: Salinity, Turbidity and nutrients (Nitrate, Phosphate and Silicate); residual missing measurements of nitrate concentrations, for instance, can be visualized. After a correlation analysis, N_f =10 physico-chemical parameters, not correlated, are retained and detailed in NC column of the Table I. Note that Fluorescence signal is not taken into account, but is used to validate clustering results since we have no ground truth: it is a presence indicator of phytoplankton cells.

When an observation data contains at least one missing value (among in $\mathbb{R}^{10}$), this point does not participate to the generation of symbols and states. Centering and standard deviation scalings are achieved on each parameter to avoid parameter range influence.

Data from 2005 to 2008 are considered to build HMM parameters. To test the time modeling, data from 2009 will be tested with the same pretreatment protocol.

B. Vector quantization validation

The number of symbols $\mathbf{V}$ required to characterize a state is first analyzed. Selection of the M symbols from the data is performed by 100 random drawings. A 1-Nearest Neighbor algorithm (1-NN) is used to evaluate the approximation of a mixture of components per state. 1-NN is first applied on each parameter, and then on the multidimensional matrix (from 2005 to 2008). K-means algorithm is also applied to build $K = M$ representatives per state.

Two scores are considered: Rate Recognition (RR) and the monthly Overlap defined by the following equation:

$$Overlap = \frac{\sum_i [|s_1(i)| + |s_2(i)| - max(|s_1(i)|, |s_2(i)|)]}{|s_1 \cup s_2|} \quad (3)$$

TABLE II
1-NN RR AND OVERLAP SCORES (MEAN AND STANDARD DEVIATION IN PERCENT) FOR M REPRESENTATIVES PER STATE ACCORDING EU-WFD LABELING : 2005-2008 DATABASE

M values	Random selection		K-means	
	RR	Overlap	RR	Overlap
1	68.1 (8.9)	18.4 (6.8)	82.7 (0.1)	11.3 (0.1)
10	79.4 (4.0)	18.5 (3.7)	89.8 (0.8)	10.0 (0.7)
100	87.6 (0.9)	12.2 (0.9)	95.1 (0.2)	4.9 (0.2)
1,000	94.7 (0.2)	5.2 (0.2)	**98.4 (0.1)**	**1.6 (0.1)**

TABLE III
STFKM INFLUENCE MEASURED WITH A SVM APPROACH: RR AND OVERLAP SCORES (IN PERCENT)

Sampling	Training		Test	
	RR	Overlap	RR	Overlap
No sampling-SVM	97.9	2.1	92.9	7.0
Random 1,000-SVM	93.3	6.7	92.2	6.7
STFKM-SVM	95.4	4.6	92.6	7.4

$|.|$ is the cardinal operator and $|s_1(i)|$ defines the number of points labeled s_1 during the i^{th} month. Phytoplankton productive and non-productive periods are expected to have no overlap according to EU-WFD.

For the monodimensional analysis, water temperature ETCO1 is the most discriminative parameter, with a recognition rate from 75.1% ($\pm$ 3.5) for one representative per state, to 77.8% ($\pm$ 0.4) for 1,000 representatives. For the multidimensional analysis, Table II summarizes the mean and standard deviation of the two scores, RR and Overlap, for different M-values. Approximating data distribution with one unique random representative gives poor recognition rate (68.1%) and often an important Overlap (18.4%) of the two desired environmental states. To decrease this Overlap around 10%, more than 100 random representatives are required.

K-means is a geometric approach adapted for linearly separable data sets. This algorithm requires to know the desired number of symbols (centers) M. Here with 10 symbols per state, the Overlap is around 10%. To reduce this Overlap around 5%, 100 representatives per state are required.

The proposed STFKM automatically searches the number of symbols that describes the data structure, and it is fast running. The impact of the STFKM selection is tested with a learning machine: a Support Vector Machine (SVM). A SVM model (radial basis kernel) was trained on 2005-2008 data and was tested on 2009 data with 10 cross-validation. Three experiments for the SVM training were led: on all training data, on 1,000 randomly selected representatives per state (M =2,000) of this data, and on symbols issued from STFKM vector quantization.

Table III summarizes SVM capacities of training and generalisation (test) for these 3 studies with RR and Overlap scores. In generalisation, STFKM algorithm allows to keep a similar recognition rate (92.6%) and Overlap score (7.4%) to no sampling-SVM. STFKM-SVM gives better training capacity than a random selection of M symbols. In this supervised context, we can conclude that the obtained vector quantization by STFKM is a relevant data reduction.

The stability of STFKM algorithm is then assessed according to the Rand Index (RI) of 10 achieved symbol generations. RI score [28] is a measure of similarity between two data clusterings $\mathbf{P}_1$ and $\mathbf{P}_2$ of a given set of n elements $\mathbf{E} = \{e_1, \ldots, e_n\}$. Note that the number of clusters in each partition can be different. RI is computed according to the following equation:

$$RI(\mathbf{P}_1, \mathbf{P}_2) = \frac{(a+b)}{\binom{n}{2}} \tag{4}$$

where a (resp. b) is defined by the number of pairs of elements in $\mathbf{E}$ that are in the same set in $\mathbf{P}_1$ (resp. in different sets) and in the same set in $\mathbf{P}_2$ (resp. in different sets) .

In a fully unsupervised context, STFKM approach allows to respect the high frequency information without losing data structure. Table IV shows that RI scores of the 10 obtained partitions and the number of retained symbols are quite similar. The RI score near to one shows that STFKM algorithm gives a robust vector quantization.

TABLE IV
RAND INDEX AND M-VALUE BOXPLOT VALUES FOR 10 SYMBOL GENERATIONS

Boxplot	Min	Q1	Median	Mean	Q3	Max
RI	0.99	0.99	0.99	0.99	0.99	0.99
M	2744	2749	2754	2759	2763	2790

TABLE V
RR AND OVERLAP SCORES (IN PERCENT) OF UNSUPERVISED CLASSIFICATION

Sampling	Building Database		Test Database	
	RR	Overlap	RR	Overlap
S-EM (C_SAL1)	70.1	14.7	77.0	11.9
STFKM-EM (EII)	83.6	13.4	91.2	4.3
STFKM-HC	66.9	0.6	66.9	0.5
STFKM-SC	79.1	11.9	84.1	5.7

TABLE VI
SIMILARITY SCORES FOR SIGNAL RECONSTRUCTION BASED ON STATE CLASSIFICATION FROM HMM MODEL.

Year	2005	2006	2007	2008	Mean	2009
sim	0.87	0.84	0.83	0.86	0.85	0.79

C. State generation validation

Then, the spectral clustering algorithm is compared to usual unsupervised approaches : EM, Hierarchical Clustering (HC) for a set number of states $N = 2$ with the same STFKM symbols. For experiments, a 7-neighborhood is considered for the scaling parameter of the similarity matrixin SC. We used EM and HC algorithms implemented in R-Gui (library Mclust and stats: http://www.r-project.org/). Only the best or runnable options are retained: Expectation-Maximisation (EII model), HC with complete linkage. EM is also performed on each of the 10 parameters, only results of the most discriminative parameter (salinity CSAL1) are presented. 1-NN algorithm is used to label data from the built model. The 2005-2008 period is named building database, and the year 2009 test database.

Table V presents RR and Overlap scores to analyse jointly. In spite of its very low Overlap score around 0.5%, hierarchical clustering (cutting obtained tree to 2 clusters) does not separate productive and non-productive periods. State 1 represents 84,118 of 84,614 points, 99% of the building database and state 2 concerns few points in August, September and November for the building database and February, April and June for the test database without any biological or sensor interpretation. EM approach offers the best RR results for the two databases. STFKM approach gives lower RR than EM one, but its Overlap for the largest database (building database) is reduced. STFKM-SC approach is a balanced one for EU-WFD labeling, and will be relevant for a number of states greater than 2. Indeed, we expect that our hybrid HMM system can detect more than 2 states like phytoplankton spring bloom or automnal bloom, rare events.

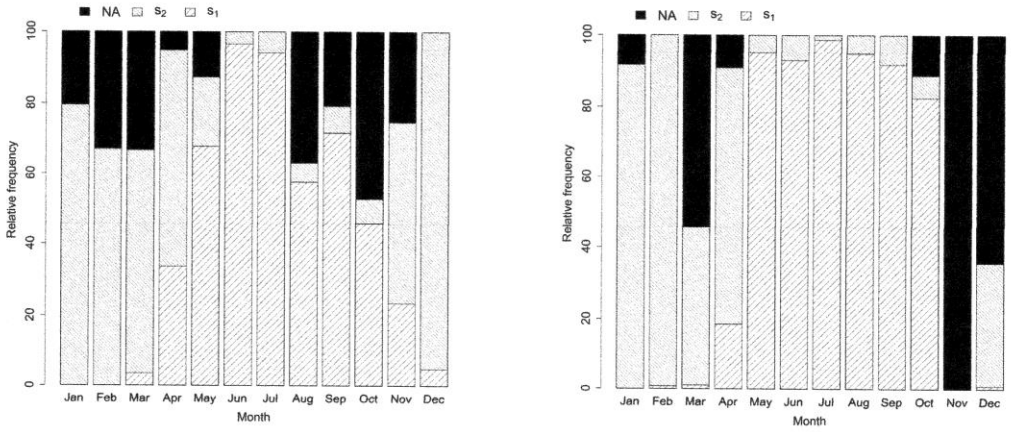
D. Time modeling validation, fixed 2-state HMM

We evaluate, through experiments, the reliability of our hybrid model: the entire procedure for building one HMM from clustering is repeated 10 times. The 10 partitions have a mean RI score of 0.95, so the whole STFKM-SC step (symbol and state generation) is robust. We keep the symbol and state partition with the smallest normalised multi-cut, MNCut [14] to build HMM. For experiments, the number of states is set: $N = 2$ and other parameters are automatically tuned: explained variance is fixed to 95 percent. According to the MNCut criterion, HMM built has $M = 2794$ symbols.

According to the EU-WFD labeling, 79.3 percent of building database is well recognized with an Overlap of 11.7%, and 82.1 percent of 2009 test database is well recognized with an Overlap of 6.7%. Fig. 6 illustrates the distribution of labeled states by HMM prediction for the building database and the 2009 test database. In 2005-2008, state s_1 in red color ties in with the period from March to December with a dominant April-November period whereas state s_2 in green color is dominant in the November-April period. Over the period 2009 state s_2 ties in with the period from April to October, whereas state s_2 in green color is dominant in the December-April period. Many data in March and August-November have no estimated state (noted NA in black color in Fig. 6) due to one least missing value in $\mathbb{R}^{10}$; that means the system forecasts confusions in transition periods. But states s_1 and s_2 match well with the two main environmental EU-WFD states: s_1 and s_2 characterize the phytoplankton dynamics, the productive and the non-productive periods.

To validate the relevancy of the built model from 2005-2008 and its symbol quantization in a fully unsupervised context, MAREL-Carnot signals $\widehat{\mathbf{Obs}}$ on the year 2009 are reconstructed and compared with the original data $\mathbf{Obs}$. For an observation $\mathbf{obs}_t$, the system estimates its state s_i with the higher likelihood. Then the most present symbols $\mathbf{v}_k$ in this state are retained, see Fig. 7. A similarity sim score is defined by the following equation 5:

$$sim(\mathbf{Obs}, \widehat{\mathbf{Obs}}) = \frac{1}{|\mathbf{Obs}|} \sum_{t=1}^{|\mathbf{Obs}|} \frac{1}{\left\| \mathbf{obs}_t - \widehat{\mathbf{obs}_t} \right\| + 1} \tag{5}$$



(a) Period 2005-2008, HMM building database.

(b) 2009, test database.

Fig. 6. State distribution per month for a typical seasonal cycle: state s_1 is represented in red color and s_2 in green color. The black color, named NA, concerns measures whose state is not estimated (one least missing values).

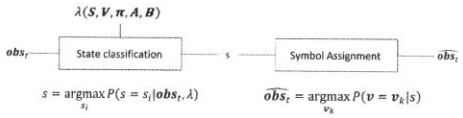


Fig. 7. Time series reconstruction from HMM classification. Signal $\widehat{\mathbf{Obs}}$ is reconstructed from original observation **Obs** thanks to the M symbols.

From 2005 to 2008, each year participates to HMM building, and reconstructed signals are compared to the original ones so as to assess the modeling power. Table VI shows the similarity scores, Eq.(5), between original and reconstructed signals for each year over the period 2005-2008 and their related mean. The last column corresponds to the year 2009. This similarity score is bounded by 0 and 1: 1-value implies that reconstructed signals are exactly the same as the original ones. The similarity scores and their mean are greater than 0.83 from 2005 to 2008. So, we conclude that reconstruted signals are very close to the original data, and that the vector quantization algorithm for HMM states is efficient. Therefore, the proposed system has an interesting generalisation power, for the year 2009, which does not participate in HMM building. Indeed, the time series are built with a high similarity score: 0.79 according to the choice of the most probable symbol of the state. Fig. 8 illustrates original and reconstructed signal of one parameter, dissolved oxygen concentration C_O21 for 2009.

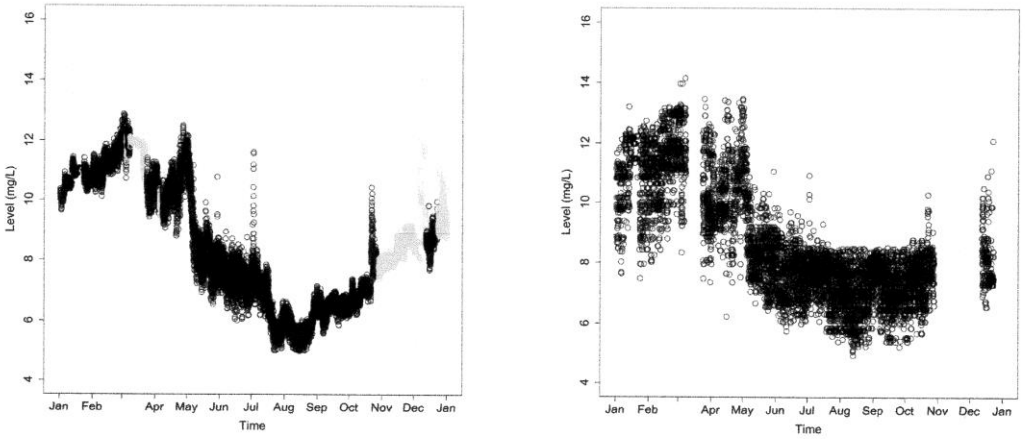
IV. N-STATE HMM FOR PHYTOPLANKTON DYNAMICS

Considering that HMM is now validated on a fixed 2-state biological dynamics, the next step is to increase the number of states to refine and to try to better understand the bloom determinism and its dynamics. A 7-state HMM with M =2,884 symbols is built from 2005-2008 database, $N = 7$ according to the eigengap technique. Time modeling validation is achieved according to the same protocol as the one of the 2-state HMM. Reconstructed 2009 signals have a similarity score above 0.8 with the original data.

Fig. 9 is the color-state projection of the estimated states over the period 2005-2008 on Fluorescence signal; black color denotes unlabelled observations due to missing value. The state distribution and sequencing are illustrated in Fig. 10 with the same color standard.

To interpret this model and to relate a ecological meaning, we analyze the state sequencing and the characterization of each state by a correlation analysis (Table VII).

State s_6 (yellow) clearly highlights high salinity values, ranging from 33.9 to 36.2 with a mean value of 35.4 (Fig. 5). These values are more representative of offshore waters. And then we can conclude in this coastal zone that mainly state s_6 corresponds to salinity anomalies (sensor failures). Nevertheless, s_6 may sometimes be explained when west winds persist and consequently bring more offshore waters to the coast. State s_3 (blue) is representative of the winter non-productive period, with high nutrient concentrations and low temperature (Table VII).



(a) C_O21 original signal. (b) $\widehat{C\_O21}$ estimated signal.

Fig. 8. 2009 - Dissolved oxygen concentration signals: original signal (a) and estimated signal (b) by HMM. In gray color, time measurements are not estimated by HMM due to at least one missing parameter at this time.

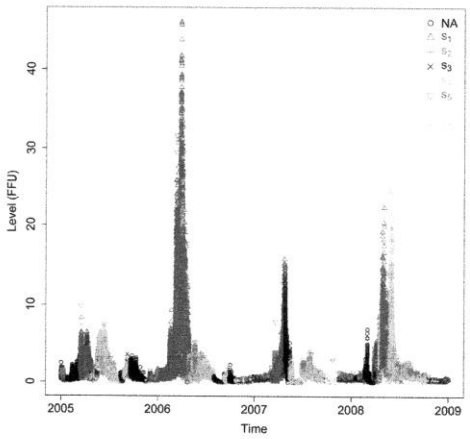
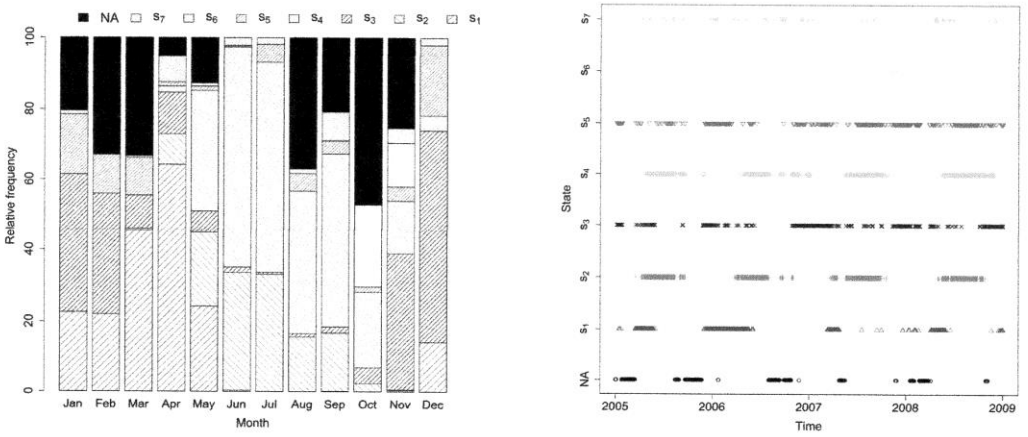


Fig. 9. Clustering results: Color-state sequencing projection on the fluorescence signal for 2005-2008. Black color, named NA concerns measures whose state is not estimated (one least missing value).

The initiation of the main phytoplankton bloom, and then the growing phytoplankton stage (between February and May: inter-annual variability of the bloom) are characterized by s_1 (red). High oxygen concentrations, explained by a high phytoplankton production (photosynthesis) are observed during this stage (Table VII). During state s_1, phytoplankton mainly uses the winter nutrients stock, and consequently this state corresponds to the new production period [29]. States s_2 and s_4 follow state s_1, and are identified as the regenerated production period when phytoplankton production is based on regenerated nutrients (transformation of the organic matter from the previous bloom - state s_1 - into new available nutrients).

States s_5 (pink) and s_7 (grey) correspond to rare or short events, respectively, with high turbidity during storm events, and high phosphate and silicate concentrations (C_PO1 and C_SI1 in Table VII). More investigations are required to better understand the main processes involved during these periods.

Fig. 11 illustrates the predicted states for the year 2009 with their sequencing. The 7-state HMM succeeds in predicting phytoplankton biomass dynamics. The state sequencing matches our assumption with a pre-bloom winter period (state s_3 mainly) followed by the main phytoplankton bloom based on external nutrient inputs (state s_1), and the regenerated bloom (states s_2 and s_4).



(a) State distribution per month for a typical seasonal cycle over the period 2005-2008.

(b) Measure sequencing in each state over the period 2005-2008.

Fig. 10. Clustering results: NA concerns measures whose state is not estimated (one least missing value).

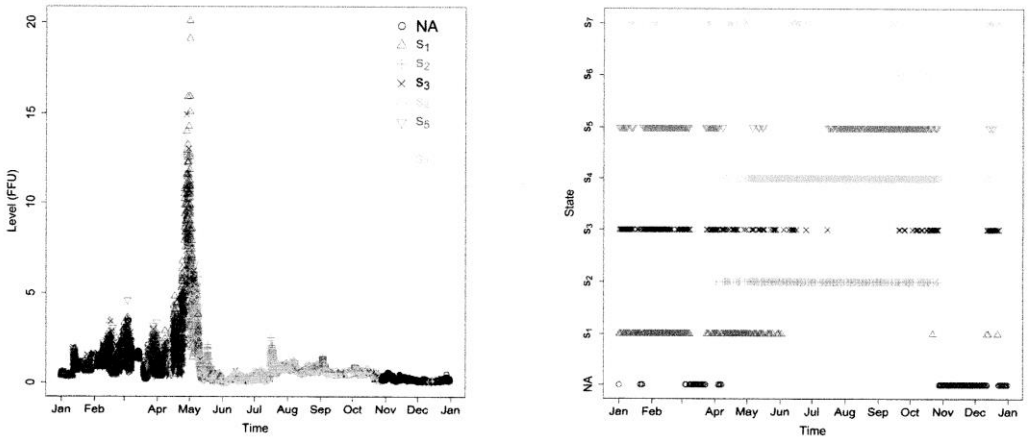
TABLE VII
CORRELATIONS BETWEEN PARAMETERS AND STATES (IN BOLD: THE HIGHEST CORRELATION COEFFICIENTS).

State	s_1	s_2	s_3	s_4	s_5	s_6	s_7
Color	red	green	blue	cyan	pink	yellow	grey
C_NI1	-0.14	-0.24	**0.54**	-0.24	0.12	0.02	-0.03
C_O21	**0.64**	-0.16	-0.03	**-0.38**	0.05	-0.17	0.01
C_PO1	-0.08	-0.03	-0.04	-0.07	-0.02	-0.03	**0.57**
C_SI1	-0.11	-0.16	0.20	-0.25	0.15	0.04	**0.47**
CSAL1	0.12	0.13	**-0.36**	0.10	-0.25	**0.42**	0.00
E_LU1	-0.08	**0.73**	-0.21	-0.24	-0.06	-0.06	0.00
E_TU1	-0.11	-0.14	-0.01	-0.22	**0.76**	-0.03	-0.05
E_VVR	-0.25	-0.03	0.16	-0.05	**0.31**	-0.06	-0.03
ETCO1	**-0.50**	**0.32**	**-0.37**	**0.56**	-0.14	0.07	0.05
XMAHH	-0.00	0.03	0.02	-0.03	0.01	0.01	0.00

V. CONCLUSIONS AND FUTURE WORKS

Two N-state HMM was built in order to forecast phytoplankton blooms near the French Channel coast from MAREL-Carnot signals (IFREMER, Boulogne-sur-Mer) without any biological knowledge. HMM building requires at least two parameters: a number of states, a number of symbols that characterize states. These parameters are commonly estimated iteratively by Expectation-Maximisation. We propose a one-pass process to estimate HMM symbols and states in a fully unsupervised context. A proposed Self Tuning Fast K-Means STFKM algorithm extracts symbols from observation data. From this vector quantization, a spectral clustering approach, with no tuning too, generates HMM states that allows to treat non-convex data. A signal reconstruction approach is proposed to assess HMM prediction.

Result analyses from the MAREL-Carnot buoy data first demonstrate interests and the stability of each used algorithm (state and symbol generation) throughout the monitoring chain. The high resolution information is preserved. Built 2-state HMM permits to detect the main productive and non-productive periods, as used for the purposes of the EU Water Framework Directive to assess good environmental status. A 7-state HMM was proposed to refine knowledge about phytoplankton bloom dynamics in a temperate ecosystem, temporarily dominated by a harmful algae (*Phaeocsytis globosa*). The obtained state sequencing coincides with dynamics described using measurements from low resolution system near the MAREL-Carnot (Rephy/SRN data [30]). The proposed HMM system succeeds in characterizing phytoplankton dynamics from new incoming data (in near real-time approach). Using the main statistical characteristics of the parameters underlying the definition of a given state, the system will allow to further increase knowledge about the main controlling or forcing parameters (*i.e.*, nutrient pressure, light availability, turbidity), the environmental status (*e.g.*, phytoplankton biomass), and the direct and/or indirect effects of such blooms (*e.g.*, oxygen concentration).



(a) Color-state projection on Fluorescence signal.

(b) State sequencing.

Fig. 11. Predicted state results in 2009 by HMM: Predict state results in 2009 by HMM: (a) Color-state projection on Fluorescence signal and (b) state sequencing. NA (black color) corresponds to measures with no estimated state.

The main limiting step in the monitoring chain is the removing samples with missing values. Indeed, the latter affects the state estimation and characterization (symbol process). Some phytoplankton blooms were not taken into account for HMM building.

Several environmental monitoring and research programmes could benefit from the proposed method to avoid the critical expert labeling step when modelling. It could help to process large multivariate time series as generated by high resolution (in time and/or space) platforms, more and more frequently implemented for the integrated observation of pelagic ecosystems and biogeochemical cycles in the oceans. Moreover, the possibility of identifying environmental states (characterized by a combination of several parameters) is a clear opportunity to better understand what a good environmental status is, as defined and used for the needs of the WFD, the MSFD or other regional sea convention (as OSPAR).

REFERENCES

[1] *Directive 2000/60/EC of the European Parliament and of the Council. Establishing a framework for Community action in the field of water policy. Official Journal of the European Communities L 327/1.*, 2000.

[2] *Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive)*, 2008.

[3] B. Gokaraju, S. Durbha, R. King, and N. Younan, "A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the gulf of mexico," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 4, no. 3, pp. 710–720, Sept 2011.

[4] ——, "Ensemble methodology using multistage learning for improved detection of harmful algal blooms," *Geoscience and Remote Sensing Letters, IEEE*, vol. 9, no. 5, pp. 827–831, Sept 2012.

[5] G. Pereira and N. Ebecken, "Combining in situ flow cytometry and artificial neural networks for aquatic systems monitoring." *Expert Systems with Applications: An International Journal*, vol. 38, no. 8, pp. 9626–9632, 2011.

[6] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[7] R. Margalef, "Life-forms of phytoplankton as survival alternatives in an unstable environment," *Oceanologica acta*, vol. 1, pp. 493–509, 1978.

[8] C. S. Reynolds, V. Huszar, C. Kruk, L. Naselli-Flores, and S. Melo, "Towards a functional classification of the freshwater phytoplankton," *Journal of Plankton Research*, vol. 24, no. 5, pp. 417–428, May 2002.

[9] J. M. Koo, H. Lee, and C. Un, "An improved vq codebook design algorithm for hmm," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1, 1992, pp. 357–360 vol.1.

[10] A. H.-R. Ko, R. Sabourin, and A. de Souza Britto Jr., "A new hmm training and testing scheme." in *ICPR*. IEEE, 2008, pp. 1–4.

[11] M. Debyeche, J. P. Haton, and A. Houacine, "Improved vector quantization approach for discrete hmm speech recognition system," *Int. Arab J. Inf. Technol.*, vol. 4, no. 4, pp. 338–344, 2007.

[12] X. Liao, P. Runkle, and L. Carin, "Identification of ground targets from sequential high-range-resolution radar signatures," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 4, 2002.

[13] M. A. T. Figueiredo, S. Member, and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 381–396, 2002.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 849–856.

[15] U. von Luxburg, "A tutorial on spectral clustering," *CoRR*, vol. abs/0711.0189, 2007.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[17] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 1601–1608.

[18] W. Kong, S. Hu, J. Zhang, and G. Dai, "Robust and smart spectral culstering from normalized cut," *Neural Computing and Applications*, vol. 23, no. 5, pp. 1503–1512, October 2013.
[19] D. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in *15th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD)*, Paris, France, 2009, pp. 907–916.
[20] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation." in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, W. Burgard and D. Roth, Eds. AAAI Press, 2011, pp. 313–318.
[21] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, 1967.
[22] G. J. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.
[23] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.
[24] M. Shindler, A. Wong, and A. Meyerson, "Fast and accurate k-means for large datasets." in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. Granada, Spain: Curran Associates, Inc., 2011, pp. 2375–2383.
[25] J. Hatrigan and M. Wong, "A k-means clustering algorithm," *Journal of Royal Statistical Society. Series C (Applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
[26] G. Sanguinetti, J. Laidler, and N. D. Lawrence, "Automatic determination of the number of clusters using spectral algorithms.in," in *IEEE Machine Learning for Signal Processing. 28-30 Sept 2005*, 2005, pp. 28–33.
[27] T. Xiang and S. Gong, "Spectral clustering with eigenvector selection," *Pattern Recogn.*, vol. 41, no. 3, pp. 1012–1029, Mar. 2008.
[28] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
[29] V. Gentilhomme and F. Lizon, "Seasonal cycle of nitrogen and phytoplankton biomass in a well-mixed coastal system (eastern english channel)," *Hydrobiologia*, vol. 361, pp. 191–199, 1997.
[30] A. Lefebvre, N. Guiselin, F. Barbet, and F. L. Artigas, "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," *ICES Journal of Marine Science*, vol. 68, no. 10, pp. 2029–2043, Sep. 2011.

Liste des figures

Chapitre 1

Chapitre 2

Chapitre 3

Chapitre 4

Chapitre 5

Liste des tableaux

Chapitre 1

Chapitre 2

Chapitre 4

Chapitre 5

Liste des algorithmes

Chapitre 2

Chapitre 3

Liste des publications et valorisations liées à la thèse

1. Articles dans des revues d'audience internationale

K. Rousseeuw, E. Poisson-Caillault, A. Lefebvre, D. Hamad. Hybrid Hidden Markov Model for Marine Environment Monitoring. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS 2014)*. DOI: 10.1109/JSTARS.2014.2341219.

2. Conférences internationales avec actes et comité de lecture

K. Rousseeuw, A.Lefebvre, E. Poisson Caillault, A.R. Nzigou, 2014. Detection and estimation of environmental states by unsupervised dynamics modelling. Application to FerryBox data. In *6th FerryBox Workshop*. 8-9 September 2014, Marine Systems Institute at Tallinn University of Technology.

K. Rousseeuw, É. Caillault Poisson, A. Lefebvre, D. Hamad, 2013. Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2013)*, pp. 3962-3965, Melbourne, 21-26 July 2013. http://dx.doi.org/10.1109/IGARSS.2013.6723700

K. Rousseeuw, A. Lefebvre, É. Caillault Poisson, D. Hamad, 2013. Detection of contrasted physico-chemical and biological environmental status using unsupervised classification tools. In *5Th FerryBox Workshop*, Helsinki, Finland, 24-25 April 2013

3. Workshop / Conférence nationales

K. Rousseeuw, E. Poisson Caillault, A. Lefebvre, D. Hamad, 2014. Modèle de Markov caché hybride pour la surveillance de l'environnement marin. In *Colloque Instrumentation haute fréquence pour l'observation et la surveillance de l'environnement marin, 10 ans MAREL Carnot*. 12 et 13 Juin 2014, Boulogne-sur-Mer – France.

A. Lefebvre, E. Poisson Caillault, K. Rousseeuw, D. Hamad, D. Soudant, F. Gohin, J. Facq, M. Répécaud, 2014. La station instrumentée MAREL Carnot : retours d'expériences de 10 ans l'observation à haute fréquence d'une zone côtière sous influence anthropique. In *Colloque Instrumentation haute fréquence pour l'observation et la surveillance de l'environnement marin, 10 ans MAREL Carnot*. 12 et 13 Juin 2014, Boulogne-sur-Mer – France.

E. Poisson Caillault, K. Rousseeuw, A. Lefebvre, B. Fassimut Mombo, 2014. Complétion de séries temporelles en utilisant l'appariement élastique ; application aux données de la station MAREL Carnot. In *Colloque Instrumentation haute fréquence pour*

l'observation et la surveillance de l'environnement marin, 10 ans MAREL Carnot. 12 et 13 Juin 2014, Boulogne-sur-Mer – France.

L.F. Artigas, S. Alvain, Z. Ben Mustapha, S. Bonato, M. Broutin, M. Courcot, V. Cornille, J. Chicheportiche, V. Creach, N. Degros, F. Gòmez, N. Guiselin, P.A. Hébert, D. Hamad, E. Houliez, E. Lecuyer, A. Lefebvre, F. Lizon, X. Mériaux, É. Poisson-Caillaut, K. Owen, M. Rijkeboer, K. Rousseeuw, T. Rutten, F. Schmitt, M. Thyssen, A. Veen, G. Wacquet, S. Zongo, 2013. Le projet DYMAPHY: Vers le développement d'un système DYnamique d'observation pour l'évaluation de la qualité des eaux Marines basée sur l'analyse du PHYtoplancton à haute résolution. In *Colloque Instrumentation haute fréquence pour l'observation et la surveillance de l'environnement marin, 10 ans MAREL Carnot*. 12 et 13 Juin 2014, Boulogne-sur-Mer – France.

K. Rousseeuw, A. Lefebvre, E. Poisson Caillault, A.R. Nzigou, 2014. Détection et estimation d'états environnementaux par modélisation dynamique non supervisée ; application aux données Ferry Box. In *Colloque Instrumentation haute fréquence pour l'observation et la surveillance de l'environnement marin, 10 ans MAREL Carnot*, 10 ans MAREL Carnot. 12 et 13 Juin 2014, Boulogne-sur-Mer – France.

K. Rousseeuw, A. Lefebvre, É. Caillault Poisson, D. Hamad, 2013. Unsupervised Hidden Markov Model building for high frequency data. DYMAPHY, project Interreg IVa entre 2 mers, final event: Presentation. 3 décembre 2013, CCI, Boulogne-sur- Mer – France.

4. Doctoriales

E. Poisson Caillault, K. Rousseeuw, A. Lefebvre, B. Fassimut Mombo, 2014. Complétion de séries temporelles en utilisant l'appariement élastique. Application aux données de la station MAREL-Carnot, *2ème Doctoriales de la Mer, Campus de la Mer*, 9 octobre 2014, Boulogne sur Mer – France.

K. Rousseeuw, A. Lefebvre, É. Caillault Poisson, D. Hamad, 2013. Système de surveillance utilisant une classification non-supervisée et une modélisation dynamique. *Doctoriales de la Mer, Campus de la Mer*, 10 octobre 2013, Boulogne sur Mer – France.

K. Rousseeuw, 2013. Complétion de séries temporelles en utilisant l'appariement élastique. Application aux données de la station MAREL-Carnot. *Doctoriales Lille Nord de France*, 2-7 juin 2013 Marcq en Baroeul – France.

K. Rousseeuw, E. Poisson Caillault, A. Lefebvre, D. Hamad, 2012. Détection et caractérisation des efflorescences phytoplanctoniques à partir des données haute fréquence. *EDSPI, Université Lille nord de France*, 6 décembre 2012, Lille – France.

5. Rapports et présentations techniques

A. Lefebvre, K. Rousseeuw, 2014. MAREL Carnot : Rapport n° 8 : Bilan d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan de l'année 2013. Rapport *Ifremer/RST.LER.BL/14.02*, 28 pages

A. Lefebvre, K. Rousseeuw, 2013. MAREL Carnot : Rapport n° 7 : Bilan d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer). Bilan de l'année 2012. *Rapport Ifremer/RST.LER.BL/13.09,* 27 pages.

A. Lefebvre, E. Caillault, K. Rousseeuw, D. Hamad, 2013. Implementation of high frequency approaches to characterize the phytoplankton community and the physico-chemical supporting parameters: Pocket FerryBox / Algae Online Analyser (Lagrangian approach), MAREL Carnot (Eulerian approach): Presentation DYMAPHY, SCM6 – 22 March 2013, Calais.

A. Lefebvre, K. Rousseeuw, 2012. MAREL Carnot : Variabilités mensuelle et interannuelle de la fluorescence, de la salinité, de la turbidité et de l'oxygène. *Rapport Ifremer/RST.LERBL/12.10*, 20 pages

A. Lefebvre, K. Rousseeuw, E. Caillault, 2012. MAREL Carnot : Rapport n°6 : Valorisation des données d'une surveillance à haute fréquence en zone côtière sous influence anthropique (Boulogne-sur-Mer) Bilan de l'année 2011. *Rapport Ifremer/RST/LER.BL/12.05*, 36 pages.

6. Séminaires

Ifremer, Brest, Présentation d'une modélisation des efflorescences à partir d'un modèle hybride HMM-Kmeans. Contribution aux réflexions pour le montage du projet H2020 JERICO-Next. 16 septembre 2013

Rijkswaterstaat, Middelbourg, Pays-Bas. Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling, DYMAPHY, 10 juillet 2013.

IFREMER, Boulogne, Présentation de mes travaux de deuxième année de thèse, 13 juin 2013

LISIC, Laboratoire Informatique, Signal et Image de la Côte d'Opale, Calais. Présentation de mes travaux de deuxième année de thèse. 23 mai 2013

Ifremer, Boulogne, Présentation de mes travaux de première année de thèse. 19 juin 2012.

RESUME

La prise de conscience des problèmes d'environnement et des effets directs et indirects des activités humaines a conduit à renforcer la surveillance haute fréquence des écosystèmes marins par l'installation de stations de mesures multicapteurs autonomes. Les capteurs, installés dans des milieux hostiles, sont sujets à des périodes de calibration, d'entretien voire des pannes et sont donc susceptibles de générer des données bruitées, manquantes voire aberrantes qu'il est nécessaire de filtrer et compléter avant toute exploitation ultérieure. Dans ce contexte, l'objectif du travail est de concevoir un système numérique automatisé robuste capable de traiter de tel volume de données afin d'améliorer les connaissances sur la qualité des systèmes aquatiques, et plus particulièrement en considérant le déterminisme et la dynamique des efflorescences du phytoplancton. L'étape cruciale est le développement méthodologique de modèles de prédiction des efflorescences du phytoplancton permettant aux utilisateurs de disposer de protocoles adéquats. Nous proposons pour cela l'emploi du modèle de Markov caché hybridé pour la détection et la prédiction des états de l'environnement (caractérisation des phases clefs de la dynamique et des caractéristiques hydrologiques associées). L'originalité du travail est l'hybridation du modèle de Markov par un algorithme de classification spectrale permettant un apprentissage non supervisé conjoint de la structure, sa caractérisation et la dynamique associée. Cette approche a été appliquée sur trois bases de données réelles : la première issue de la station marine instrumentée MAREL Carnot (Ifremer) (2005-2009), la seconde d'un système de type Ferry Box mis en œuvre en Manche orientale en 2012 et la troisième d'une station de mesures fixe, installée le long de la rivière Deûle en 2009 (Agence de l'Eau Artois Picardie - AEAP).
Le travail s'inscrit dans le cadre d'une collaboration étroite entre l'IFREMER, le LISIC/ULCO et l'AEAP afin de développer des systèmes optimisés pour l'étude de l'effet des activités anthropiques sur le fonctionnement des écosystèmes aquatiques et plus particulièrement dans le contexte des efflorescences de l'algue nuisible, *Phaeocystis globosa*.

Mots clés : Apprentissage non supervisé - Modèle de Markov Caché - Classification spectrale - Mesures hautes fréquences - Efflorescences phytoplanctoniques nuisibles - Qualité des milieux aquatiques.

ABSTRACT

Because of the growing interest for environmental issues and to identify direct and indirect effects of anthropogenic activities on ecosystems, environmental monitoring programs have recourse more and more frequently to high resolution, autonomous and multi-sensor instrumented stations. These systems are implemented in harsh environment and there is a need to stop measurements for calibration, service purposes or just because of sensors failure. Consequently, data could be noisy, missing or out of range and required some pre-processing or filtering steps to complete and validate raw data before any further investigations. In this context, the objective of this work is to design an automatic numeric system able to manage such amount of data in order to further knowledge on water quality and more precisely with consideration about phytoplankton determinism and dynamics. Main phase is the methodological development of phytoplankton bloom forecasting models giving the opportunity to end-user to handle well-adapted protocols. We propose to use hybrid Hidden Markov Model to detect and forecast environment states (identification of the main phytoplankton bloom steps and associated hydrological conditions). The added-value of our approach is to hybrid our model with a spectral clustering algorithm. Thus all HMM parameters (states, characterisation and dynamics of these states) are built by unsupervised learning. This approach was applied on three data bases: first one from the marine instrumented station MAREL Carnot (Ifremer) (2005-2009), second one from a Ferry Box system implemented in the eastern English Channel en 2012 and third one from a freshwater fixed station in the river Deûle in 2009 (Artois Picardie Water Agency).
These works fall within the scope of a collaboration between IFREMER, LISIC/ULCO and Artois Picardie Water Agency in order to develop optimised systems to study effects of anthropogenic activities on aquatic systems functioning in a regional context of massive blooms of the harmful algae, *Phaeocystis globosa*.

Keywords: Unsupervised classification - Hidden Markov Model - Spectral clustering - High resolution measurements - Harmful algal blooms - Aquatic ecosystems quality.