الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR ANNABA

جامعة باجي مختار- عنابة

Année 2016

Faculté des sciences de l'ingéniorat
Département d'électronique

THÈSE

Présentée en vue de l'obtention du diplôme de DOCTORAT DE TROISIÈME CYCLE

LA TÉLÉSURVEILLANCE DE PHÉNOMÈNES OU D'OBJETS DANS DES ENVIRONNEMENTS DIVERS

Option : Instrumentation et traitement de l'information

Par :

Youssouf DIAF

Directeur de Thèse:
Dr. Mohamed KADDECHE — **Université d'Annaba**

Devant Jury:

Président:			
Pr.	**Larbi**	**ALLAL**	**Université d'Annaba**
Examinateurs:			
Pr.	**Abdelkrim**	**MOUSSAOUI**	**Université de Guelma**
Pr.	**Layachi**	**BENNACER**	**Université de Guelma**
Pr.	**Rafik**	**DJEMILI**	**Université de Skikda**
Pr.	**Ali**	**TAHAR**	**Université d'Annaba**

REMERCIEMENTS

Je remercie, avant tout, Dieu, le Tout-Puissant, de m'avoir accordé parmi Ses innombrables Grâces, santé et courage pour accomplir ce travail.

Je remercie Allal LARBI, Professeur de la Faculté des Sciences de l'Ingéniorat de l'Université d'Annaba d'avoir accepté d'être le président du jury.

Je suis très reconnaissant à Abdelkrim MOUSSAOUI, Professeur à l'Université de Guelma, pour avoir accepté d'être examinateur de cette thèse et pour les critiques constructives qui m'ont permis d'améliorer ce manuscrit.

Je remercie vivement Rafik DJEMILI, Professeur à l'Université de Skikda, qui m'a fait l'honneur de participer à ce jury en tant qu'examinateur.

Merci beaucoup au Professeur Layachi BENNACER, professeur à l'Université de Guelma,, qui a accepté d'être examinateur de cette thèse. Je me souviendrais spécialement de son remarquable enthousiasme scientifique, de sa grande disponibilité et de son soutien moral constant.

Je remercie beaucoup Ali TAHAR, Professeur à l'Université d'Annaba département de biologie, qui m'a fait l'honneur de participer à ce jury en tant qu'expert en la matière.

Je tiens tout particulièrement à remercier Mohamed KADDECHE, Docteur à l'université Badji Mokhtar Annaba et directeur de cette thèse, qui a accepté de passer toutes ces années à mes côtés. Je lui suis très reconnaissant de m'avoir accordé sa confiance et il m'est difficile de traduire en mots son soutien, sa disponibilité permanente et aussi son enthousiasme. Qu'il trouve ici le témoignage de ma sincère reconnaissance et de ma sympathie.

Je remercie le Professeur Salah TOUMI, Professeur à l'Université d'Annaba et Directeur du laboratoire LERICA pour ces excellents conseils dans la manière d'aborder mon travail et pour son aide plus que précieuse dans les moments difficiles. Je n'oublierai jamais sa simplicité, sa gentillesse et surtout sa remarquable qualité humaine.

Je souhaite témoigner également mes profonds remerciements à tous mes frères et amis surtout les membres et les doctorants du LERICA qui ont eu la patience et le courage de s'accrocher à notre amitié.

Merci à toutes les personnes que j'aurais oubliées…

Merci et encore merci à mes parents qui ont toujours su me soutenir et me donner les forces nécessaires pour arriver à la fin de cette thèse. Je n'aurais jamais écrit chacune de ces pages sans leur inestimable aide et amour.

Merci à ma conjointe de m'avoir soutenue dans les moments difficiles que j'ai eu à vivre, de m'avoir écouté et pour sa générosité. Merci mon Ange d'être toujours là pour moi...

Dédicaces

A la mémoire de ma grande mère mama fatima,
Puisse Dieu, le tout puissant, l'accueillir dans son infinie miséricorde,

A mes très chers parents,
A toi Rym pour ce que tu m'apportes, pour ta confiance et ton réconfort,
A mes frères : Ilyes, Amine et Khalil,
A toute ma famille,
A tous mes professeurs que j'ai eu tout au long de mon parcourt d'études,
A tous mes amis et ceux qui me tiennent à cœur,

Pour le pire et le meilleur... Je vous dédie cette thèse.

الملخص

المراقبة عن بعد و التعرف السريع على المناطق البحرية الملوثة بالظواهر الضارة الناجمة عن وجود الطحالب السامة هو ضرورة اليوم، نظرا للجوانب الصحية والبيئية الهامة المتعلقة بالبيئات المائية .تصميم نظام التعرف أو التنبؤ يجب أن يخلف الأسلوب القديم الذي يكلف الكثير من الوقت والمال.

نظم الرصد الموجودة أو قيد التطوير متعددة، ولكن لديها كلها صعوبات في التصنيف أو التنبؤ وأكثر من ذلك فالنتائج هي ذات صلة ببيئات معينة أو قواعد بيانات محددة. الهدف من عملنا هو البحث في كلا الاحتمالين: الأول هو التنبؤ بالكثافة العددية للطحالب من خلال تحليل بيانات البيئة، والثاني هو التعرف الاوتوماتيكي على الطحالب عن طريق تحليل عينات المياه باستخدام أجهزة مثل تدفق الكريات الذي يوفر بيانات عن البنية الداخلية والمورفولوجية لكل خلية.

طرق معالجة المعلومات المقترحة في هذا العمل هي آلات التعلم التالية: الشبكة العصبية، أشجار القرار و خليط نماذج قوس. لقد تم تجريبها في كلا الاحتمالين و النتائج المتحصل عليها واعدة.

Abstract

The telemonitoring and the rapid identification of contaminated aquatic areas by the appearance of harmful events caused by toxic algae presence phenomena is a necessity today, due to the significant health and environmental issues related to aquatic environments. The design of a recognition or prediction system must be made instead of old method that costs a lot of time and money.

Monitoring systems already in place or under development are multiple, but they have all classification or prediction difficulties and more than that the results are relevance to specific environments or database.

The purpose of our work to exploit both possibilities: the first is the prediction of intense algal blooms by analyzing the environments data, the second is the automatically recognition of the alga by analyzing water samples using devices such as the flow cytometer which provide data on the internal structure and morphology of each cell. Treatment methods proposed in this work are machines learning: the neuronal network, decision trees and gaussian mixture model. They are tested for the two approaches and the results are promising.

Résumé

La télésurveillance et l'identification rapide des zones aquatiques contaminées par l'apparition des événements nuisibles causée par le phénomène de la présence des algues toxique est une nécessité aujourd'hui due à l'important enjeu sanitaire et environnemental lié aux milieux aquatiques. La conception d'un système de reconnaissance ou de prédiction doit être faite à la place des méthodes anciennes qui coutent beaucoup de temps et d'argent.

Les systèmes de surveillance mis en place déjà ou en cours de développement sont multiples, mais ils trouvent tous des difficultés de classification ou de prédiction. Plus que ça la pertinence des résultats à des milieux précis ou à des bases de données déterminées, rend les modèles non généralistes.

L'objet de notre travail est d'exploité deux possibilités : la première et la prédiction des blooms d'algale intense par analyse des données issues de milieux de vie naturelle d'algue. La deuxième vise à reconnaitre de façon automatique l'algue par analyse des échantillons d'eau en utilisant des appareils comme le cytomètre en flux qui fournissent des données sur la structure interne et la morphologie de chaque cellule. Les méthodes de traitement proposé dans ce travail sont des machines à apprentissage : les réseaux de neurones, les arbres de décision et les modèles de mélange gaussiens. Ils sont testés dans les deux approches et les résultats sont prometteurs.

Table Des Matières

Chapitre 3 :

Reconnaissance et identification des espèces phytoplanctoniques

Chapitre 4 :

Résultats et discussions

Liste des figures

Chapitre 1

Chapitre 2

Chapitre 3

Chapitre 4

Liste des tableaux

Chapitre 2

Chapitre 3

Chapitre 4

Liste des acronymes

Acronymes	**Définition**
AD	Arbres de Décision
Algorithme EM	Expectation–maximization algorithm
ANN	Artificial Neuronal Network
APPL	l'Agence pour la Protection et la Promotion du Littoral
BP	backpropagation
CBR	Case-Based Reasoning
CDPH	California department of public health U.S.A
CF	cytométrie en flux
CTD	Conductivity, Temperature, Depth
D. Acuminata	Dinophysis Acuminata
FLO	Forward-scattered Light Orange
FLR	Forward-scattered Light Red
FLY	Forward-scattered Light Yellow
FWS	Forward Scatter
GMM	Gaussian mixture modele
GRNN	Generalized Regression Neuronal Network
HAB	Harmful Agal Blooms ou blooms nuisibles d'algues
HS	High Sensitivity
IFREMER	Institut français de Recherche pour l'Exploitation de la Mer
In situ	locution latine qui signifie "dans son milieu naturel"
ISE	Ion Selective Electrode
kNN	k Nearest Neighbor
LS	Low Sensitivity

ML	Machine Learning
MLP	MultiLayer Perceptron
PCA	Principal Component Analyze
RBF	Radial Basis Function
REMI	Le réseau microbiologique
REPHY	Réseau de surveillance du phytoplancton et des phycotoxines
RNO	Réseau National d'Observation de la qualité du milieu marin
ROCCH	Réseau d'Observation de la Contamination Chimique du littoral
SVM	Support Vector Machine
SVM	Support Vector Machine
SWS	Sideward Scatter

Introduction générale

Introduction générale

Ce thèse de recherche cible la télésurveillance. Un sujet qui est très vaste avec plusieurs disciplines dont le domaine de la pollution.

La pollution est un phénomène qui se propage chaque jour dans les milieux terrestre, atmosphérique et aquatique. Elle présente aujourd'hui un risque majeur pour l'existence humaine.

La pollution aquatique occupe une grande partie des recherches aujourd'hui, due à l'importance de l'eau qui constitue 75 % de notre planète. Cette eau est essentielle pour la survie des êtres vivants. La pollution aquatique prend plusieurs formes ; l'une de ses formes est la pollution biologique. Ce type de pollution est défini par la présence d'êtres vivants capables de détruire leurs environnements.

Ces derniers sont des microorganismes appelés les phytoplanctons qui sont de nature végétale et de taille de l'ordre du µm. Ces algues peuvent transporter des toxines absorbées par les produits de mer à forte consommation tels que les poissons, les crustacés etc.

La détection de la prolifération de ces micro-organismes doit être faite d'une façon rapide (temps réel) ou prédictive. Le problème rencontré avec les méthodes traditionnelles est le temps énorme consacré au comptage et la discrimination de cellules afin de prendre une décision par un personnel qualifié et professionnel.

La surveillance automatique est donc la meilleure façon d'accélérer la tâche de reconnaissance et de prédiction de ces phénomènes nocifs. Dans cette recherche nous abordons ces deux tâches par deux applications différentes. A cet effet, on utilise des techniques appropriées à des modèles intelligents.

Dans un premier temps nous étudions un modèle de prédiction de bloom d'algue toxique Dinophysis Acuminata dans le littoral du Havre France. La modélisation utilise les réseaux de neurones comme approche analytique intelligente. Ce modèle sert comme un système de surveillance côtier.

Dans la deuxième partie on a exploité la voie de reconnaissance et d'identification automatique en temps réel. Nous présentons plusieurs modèles déjà étudiés, ainsi que l'appareil de mesure utilisé dans ce genre de travaux de recherche. L'analyse est effectuée sur une base de données très riche.

L'étude qui suit est organisée autour de quatre chapitres. Dans le premier chapitre, les notions de la pollution biologique aquatique, des phytoplanctons, des différents phénomènes et les paramètres liés à l'apparition de ces microorganismes sont illustrés.

Nous allons, dans le deuxième chapitre, détailler l'analyse prédictive des phénomènes d'apparition algale. Nous présentons le modèle développé, à base de réseaux de neurones, pour la prédiction des blooms d'algues toxiques.

Le troisième chapitre est basé sur la biotechnologie appliquée au système aquatique. Nous allons détailler la technique et l'instrument d'analyse des microorganismes à l'échelle microscopique : le cytomètre en flux. Ensuite, nous présentons deux modèles pour l'analyse des signaux issus de l'instrument pour la reconnaissance. Les deux méthodes utilisées sont : les modèles de mélange gaussien et les arbres de décision.

Nous représentons dans le quatrième chapitre tous les résultats obtenus par les deux approches : celle de la prédiction et celle de la reconnaissance. Enfin, nous conclurons par une discussion détaillée de ces résultats.

Le logiciel utilisé au long de ce chapitre pour les simulations des modèles est le logiciel MATLAB. Ce logiciel contient plusieurs fonctions et toollbox qui permettent l'analyse d'information sous différents modèles de façon simple et avec des interfaces bien contrôlées.

Je terminerai ce travail par une conclusion et les perspectives qui sous-tendent notre travail ainsi que les horizons de recherche future dans ce domaine.

Chapitre 1

Problématique

1.1. Introduction

Le milieu aquatique est indispensable pour la vie humaine; il fournit le premier élément de la vie qui est l'eau ainsi qu'une variété de ressources vivantes.

Les milieux aquatiques sont généralement menacés par les problèmes d'aménagement, pollution et exploitation irrationnelle des ressources.

Dans ce chapitre, nous allons répondre aux questions suivantes: le type de phénomène aquatique surveillé, pourquoi cette surveillance? Quels sont aujourd'hui les dispositifs, les programmes et les organisations assurant ce type de surveillance ?

Nous définirons ainsi quelques notions de biologie qui nous aideront à mieux comprendre les modèles développés dans cette recherche.

1.2. Pollution biologique aquatique

La pollution d'eau est due aux agents organiques (les matières biodégradables... etc.), physiques (matières en suspension, chaleur, radioactivité... etc.), chimiques (substances indésirables ou dangereuses... etc.), et biologiques. Ces polluants rendent impropres l'eau et par conséquent son utilisation par l'homme et perturbent aussi les écosystèmes aquatiques.

1.2.1. La pollution biologique par phytoplancton

L'objet central d'un bioprocessus est la cellule. La cellule vivante est un système complexe qui est souvent défini comme la plus petite unité autonome biologique.

La cellule est capable de construire ses propres composants et fournir sa propre énergie à travers des processus physiques et chimiques qui constituent le métabolisme cellulaire. Mais dans des conditions bien définies, cette vie microscopique peut faire de grands dégâts dans son environnement.

La pollution biologique est définie comme la présence et la contamination des eaux par des microorganismes toxiques ou non toxiques qui détruisent la faune et rendent l'eau et ses produits aquacultures impropres à la consommation.

L'un de ces microorganismes est le phytoplancton ou l'algue. Les phytoplanctons sont des microorganismes unicellulaires autotrophes, capables de réaliser la photosynthèse et de produire tous les composants nécessaires à la cellule en utilisant l'énergie lumineuse et les

nutriments inorganiques. Elles sont invisibles à l'œil et très variables en taille (allant du picomètre au micromètre).

Plus de 50 % de l'oxygène produit sur notre terre est d'origine phytoplanctonique. Ils jouent un rôle essentiel dans le rétrocontrôle du climat global et ils constituent 50 % de la production primaire de la chaine alimentaire à l'échelle mondiale.

Quelques espèces de phytoplanctons peuvent produire deux problèmes majeurs à savoir:

1.2.1.1. Les blooms

La figure 1.1 présente quelques exemples de bloom, qui est essentiellement l'eutrophisation ou l'efflorescence des algues. Ces efflorescences phytoplanctoniques sont des événements de production rapide et d'accumulation de biomasse qui sont des réponses à l'enrichissement de l'eau en éléments nutritifs. Ces derniers sont notamment composés d'azote et de phosphore qui accélèrent le développement des algues et provoquent une perturbation de l'équilibre des organismes présents dans l'eau.

Le phytoplancton colonise alors le milieu et provoque des dommages temporaires et parfois sérieux ; surtout dans les eaux douces. L'efflorescence est jugée dangereuse si le phytoplancton en question est toxique [1,2].

Figure 1.1 Quelques exemples des blooms.

1.2.1.2. Les phycotoxines

Les phycotoxines sont des toxines produites par quelques espèces phytoplanctoniques. Certaines de ces toxines sont dangereuses pour les consommateurs, car elles s'accumulent dans les coquillages ou dans l'eau (toxines diarrhéiques, paralysantes, amnésiantes... etc.). D'autres menacent la faune marine (poissons et coquillages par colmatage direct des branchies ou par les sécrétions des algues...). La figure 1.2 montre quelques exemples des effets des phycotoxines sur la vie aquatique.

Figure 1.2 Quelques exemples des effets des phycotoxines

1.2.2. Surveillance des phytoplanctons

La définition des groupes phytoplanctoniques ne correspond pas nécessairement à une classification taxonomique des organismes, mais il s'agit de déterminer les groupes d'espèces en fonction de leurs physiologies, morphologies et les facteurs qui répondent aux variations des conditions environnementales [3].

D'autre part, le suivi du phytoplancton et l'apparition des événements nuisibles toxiques sont des enjeux importants que ce soit au niveau sanitaire ou environnemental [4]. D'après la « directive européenne Cadre sur l'eau » : ce suivi est devenu un critère de

classement pour la qualité de milieux aquatiques. Des programmes de surveillance sont mis en place dans différents systèmes côtiers à travers le monde [5].

L'une des plus grandes organisations de la surveillance aquatique est l'IFREMER (Institut français de Recherche pour l'Exploitation de la Mer). Elle est fonctionnelle depuis 1974 et structurée en 3 réseaux principaux : le réseau de contrôle microbiologique (REMI), le réseau de surveillance du phytoplancton et des phycotoxines (REPHY) et le réseau national d'observation de la qualité du milieu marin (RNO/ROCCH).

Le réseau de surveillance du phytoplancton et des phycotoxines (REPHY) a été créé par l'IFREMER en 1984, suite à l'observation d'intoxications de type diarrhéique chez des consommateurs de coquillages sur les côtes bretonnes. Ces intoxications avaient pour origine le développement dans le milieu d'algues toxiques appartenant au genre Dinophysis et produisant des toxines diarrhéiques.

Le REPHY est un réseau national ayant pour objectifs d'observer l'ensemble des espèces phytoplanctoniques dans les eaux côtières et de surveiller plus particulièrement les espèces produisant des toxines dangereuses pour l'homme. La surveillance régulière de l'ensemble des espèces phytoplanctoniques permet la détection des espèces toxiques et nuisibles connues, mais également d'espèces potentiellement toxiques. C'est la présence d'espèces toxiques dans l'eau qui déclenche la surveillance des toxines dans les coquillages. Le REPHY assure la surveillance des coquillages dans leurs milieux naturels (parcs, gisements).

Un autre programme est celui de la surveillance et la protection des biotoxines marines avant la récolte des mollusques (*Preharvest Shellfish Protection and Marine Biotoxin Monitoring Program*) [6], qui est un programme de la Direction de la gestion de l'environnement dans la Division de CDPH (*California department of public health U.S.A*). Il a pour objectif principal la surveillance d'eau et la gestion de l'environnement d'eau potable. Le programme enquête, classifie et surveille également de nombreux points le long de la côte californienne de ces biotoxines phytoplanctoniques.

On peut citer un nombre important de programmes et organisation mondiale, soit gouvernementale, ou attachée à des unités de recherche qui s'occupe de la surveillance des phytoplanctons toxiques ou des blooms.

L'Algérie s'est investie dans ce domaine de surveillance des milieux aquatiques et dispose d'un seul réseau de surveillance du phytoplancton toxique au niveau de wilaya d'Alger. Il a été mis en place par l'Agence pour la Protection et la Promotion du Littoral algérois (APPL) en 2007 [7] et il a pour objectifs :

- Observer l'ensemble des espèces phytoplanctoniques environ 52, dont les espèces toxiques et nuisibles (environ 22), et recenser les événements tels que les eaux colorées.
- Surveiller plus particulièrement les espèces produisant des toxines dangereuses.

A cet effet 20 stations ont été sélectionnées entre la commune de Zéralda et celle de Réghaïa.

1.3. Stratégie et méthodes utilisées pour la surveillance

La stratégie d'étude dépend de l'objectif visé. Il s'agit de trouver un lien entre l'observation d'un phénomène et la méthodologie appropriée pour l'identifier. Face au degré de complexité des réponses des phytoplanctons, les méthodes d'échantillonnage et d'analyse manuelle sont souvent insuffisantes. D'autre part, deux facteurs sont importants pour comprendre et définir au mieux le degré de menace au sein de l'écosystème : le type et la biomasse du phytoplancton.

La "chlorophylle *a*" est la principale forme de chlorophylle présente chez les organismes qui mettent en œuvre la photosynthèse. La détermination de dosage de la "chlorophylle *a*" est alors la méthode standard la plus utilisée pour déterminer la biomasse phytoplanctonique. Cette mesure est obtenue par : la spectrophotométrie (fluorimétrie) [8] ou la chromatographie en phase liquide à haute performance (HPLC) qui est la méthode la plus précise [9].

Les méthodes déterminant la concentration en "chlorophylle *a*", permettent d'obtenir des informations sur les principaux groupes phytoplanctoniques [10]. Ces méthodes pigmentaires sont sûres et rapides dans le cas de surveillance des blooms, mais ils ne permettent pas d'évaluer directement les espèces et leurs classes.

L'analyse taxonomique du phytoplancton est effectuée manuellement par microscopie optique. Elle permet une reconnaissance et une quantification en détail des espèces existantes dans les échantillons d'eaux [11]. Cependant, cette méthode est très coûteuse en temps et

demande un opérateur qualifié et expert dans ce domaine. L'inconvénient est qu'elle ne permet pas d'observer les cellules de petite taille et qu'elle nécessite également plusieurs manipulations telles que le stockage et l'utilisation de fixateurs chimiques avant l'analyse.

Dans ce contexte, le cytomètre en flux offre un compromis entre la rapidité des méthodes pigmentaires et la précision à l'échelle individuelle des méthodes microscopiques. L'application de la cytomètre en flux dédiée à l'analyse du phytoplancton est récente [12]. La méthode repose sur les propriétés auto fluorescentes des cellules phytoplanctoniques. Elle permet également d'obtenir des informations sur la morphologie et la structure des cellules [13,14], et de les quantifier sans passer par les traitements chimiques des échantillons.

Après collection d'information sur le bloom ou le phytoplancton en lui-même, on doit déterminer le meilleur moyen d'analyser ces données, pour nous permettre une surveillance rapide, efficace et à moindre coût.

1.4. Conclusion

Notre travail de recherche s'inscrit dans le domaine de la biotechnologie, qui est définie comme étant un champ de recherche multidisciplinaire où travaillent les biologistes, les médecins, les informaticiens, les mathématiciens, les physiciens et les électroniciens afin de résoudre un problème scientifique posé par la biologie.

Un exemple d'organisation qui utilise la bio-informatique dans le monitorage des microorganismes aquatique est le laboratoire d'Écologie numérique des milieux aquatiques [15]. Il utilise les nouveaux systèmes électroniques et informatiques, afin de réaliser des outils automatiques (logiciels d'écologie : le Zoo/Phytoimage [15], et de biostatistiques: le SciViews [15]) aidons les biologistes à accélérer leurs tâches de surveillance et de traitement.

L'objectif principal de notre travail est de réaliser des programmes et des équipements capables à :

- Résoudre les problèmes de détection des blooms : notre contribution consiste à l'application d'un modèle intelligent à une base de données réduite afin de prédire l'évolution des biomasses phytoplanctoniques.
- Améliorer l'identification des phytoplanctons toxiques : nous proposons un programme d'analyse des données cytométriques utilisant la méthode de mélange gaussien.

- Accélérer l'identification des phytoplanctons toxiques : à cet effet nous avons utilisé un modèle d'arbre de décision.

Dans les deux chapitres qui suivent, nous allons détailler les améliorations introduites.

Chapitre 2

Modélisation et prédiction des blooms

2.1 Introduction

La compréhension complète des mécanismes physico-chimiques et biologiques mis en œuvre dans le processus d'efflorescence phytoplanctonique est pratiquement impossible. L'insuffisance des études scientifiques sur le processus de croissance des phytoplanctons et la diversité des espèces nous contraint à s'inspirer des processus de croissance d'autres microorganismes utilisés dans divers domaines tels que: la biotechnologie, les industries de fermentation, l'écologie, la biologie, la microbiologie, l'enzymologie ... etc.

L'objectif principal de ce chapitre est de construire un modèle capable de décrire le processus d'évolution d'un phytoplancton toxique, ce qui nous permettra de prédire une prolifération et de prendre avec bon escient les mesures nécessaires à savoir:

- Donner l'alerte sur la prolifération.
- Déclencher un contrôle de toxicité dans les coquillages et éventuellement la fermeture des centres de conchyliculture et l'interdiction de la commercialisation des coquillages.

Le modèle devra aussi fournir une compréhension la plus complète possible et par conséquent contribuer à l'amélioration de la connaissance de ce processus complexe et mal défini.

La connaissance des facteurs physiques, chimiques et biologiques élémentaires qui déterminent le comportement dynamique d'un tel écosystème est l'un des moyens possibles et efficaces pour la modélisation. Comme pour tout processus biologique complexe, le modèle devra tenir compte des influences simultanées d'un grand nombre de facteurs de diverses natures.

2.2 La prédiction des blooms

Les développements de phytoplancton sont à l'origine de manifestations d'eaux colorées dans le monde entier. Les eaux colorées les plus préoccupantes sont liées à la présence de quelques espèces de dinoflagellés.

Ces dernières années, l'apparition et le développement de ces micro-organismes dans les eaux côtières ont eu d'importantes retombées économiques et sanitaires: fermeture ponctuelle des centres de conchyliculture et des mesures d'interdiction de la commercialisation de certains coquillages.

Le phytoplancton se définit comme le plancton de nature végétale. Il est constitué par des algues microscopiques qui sont des cellules isolées ou réunies et enchainées, mesurant de quelques microns à quelques centaines de microns. Le phytoplancton est la source de la production de la matière organique dans les milieux aquatiques. L'augmentation massive de la biomasse algale et les changements de la composition en espèces phytoplanctoniques dans des nombreux milieux aquatiques a provoqué des manifestations d'efflorescence phytoplanctoniques nocives à travers le monde.

Les blooms peuvent être naturels comme les efflorescences printanières ou le plus souvent d'origine humaine. Dans ce dernier cas, les blooms déséquilibrent l'écosystème aquatique par leurs tailles qui peut s'étaler sur de vastes zones (22000 km² en 2007 au large de l'estuaire du Mississippi). Dans les cas extrêmes d'expansion de ces microorganismes, des effets néfastes affecteront la chaine alimentaire et l'écologie aquatique (pollution organique, émissions de gaz à effet de serre, mortalité de poissons et crustacés).

Ces blooms peuvent devenir un danger dans le cas où le phytoplancton sécrète des phycotoxines, ce qui cause des intoxications pour la faune aquatique ainsi que pour les consommateurs de poissons et de mollusques.

Une prédiction d'un tel phénomène peut sauver des vies humaines et éviter des pertes économiques.

2.2.1. Présentation du processus de bloom

Le processus d'apparition et de développement des phytoplanctons reste mal défini. Les conditions et les facteurs qui favorisent leurs apparitions sont mal connus. Les phytoplanctons se caractérisent par une importante diversité, des taux de croissance variables et des réponses rapides aux changements environnementaux [16]. Donc, les facteurs hydro-climatiques, physico-chimiques et biologiques sont des paramètres qui peuvent avoir une influence sur ce phénomène.

Sous certaines conditions météorologiques favorables (éclairement, température, mer calme .... etc.), les phytoplanctons élaborent toutes les substances organiques nécessaires à leur croissance par photosynthèse à partir de carbone, d'eau et des sels minéraux.

Les matières organiques (phosphore, azote et carbone) ainsi élaborées sont décomposées par les bactéries, entre autres, en matières minérales et éléments nutritifs qui contribuent à la croissance et au développement de ces micro-organismes marins.

L'apport en substances terrigènes véhiculées par de fortes précipitations constitue un renouvellement des sels nutritifs, ce qui assure le maintien en vie du phytoplancton. On peut penser que l'homme accélère considérablement le processus de croissance et de développement du phytoplancton en déversant dans les eaux côtières des grandes quantités de matières organiques ou minérales provenant des déchets urbains riches en phosphates et nitrates.

Évidemment, les micro-organismes ne sont pas seuls dans le milieu et par conséquent, ils peuvent être directement soumis à l'intervention de prédateurs tels que le Zooplancton. Tout ceci nous conduit au schéma fonctionnel très simplifié de la figure 2.1.

Les phytoplanctons, dans un milieu riche en éléments nutritifs, sont capables de stocker les substances nutritives en excès par rapport à leurs besoins métaboliques immédiats et de les utiliser lorsque le milieu extérieur en manque ce qui leur permettra de poursuivre la croissance, et certaines espèces sont capables de se diviser plusieurs fois par jour [17]. Il s'agit alors du phénomène de bloom ou d'efflorescence qui correspond à un état de prolifération à un point tel qu'il en résulte des nuisances pour l'écosystème [1,18].

L'efflorescence ou le bloom est défini comme étant l'enrichissement d'un plan d'eau par des éléments nutritifs utiles à la croissance des plantes ou autres producteurs primaires [19]. Elle est donc l'augmentation massive de la production biologique associée à des changements majeurs des paramètres physico-chimiques du milieu. Les sources d'enrichissement sont principalement l'érosion du bassin versant, les eaux usées des zones urbanisées et l'engrais d'origine agricole [20]. Les éléments principaux contribuant à cet enrichissement sont le phosphore et l'azote [21].

L'eutrophisation des lacs et des rivières est un problème croissant à l'échelle mondiale et affecte de plus en plus les communautés [22,23].

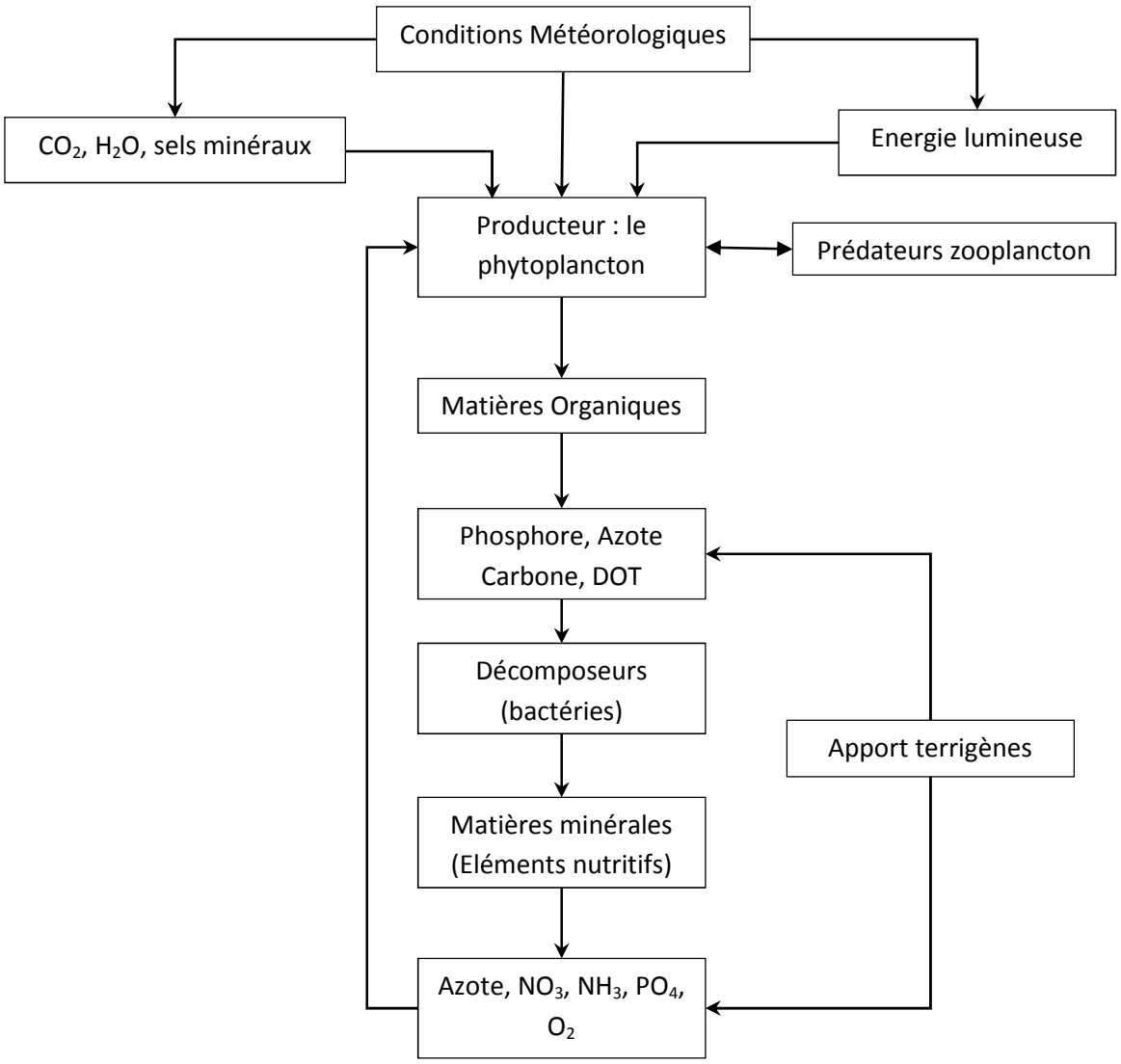

Figure 2.1. Schéma fonctionnel simplifié du processus de production organique phytoplanctonique.

2.2.2. État d'art de modélisation et prédiction de l'efflorescence phytoplanctonique

On ne peut pas présenter tous les travaux faits sur la modélisation et la prédiction des algues toxiques ou les blooms nuisibles d'algues HAB (Harmful Agal Blooms), à cause de leur nombre important, de la diversité des méthodes et des données utilisées. Tous les travaux ciblent l'utilisation des outils qui mesure soit les paramètres de développement d'algues pour la prédiction ou le taux d'efflorescence pour la détection et l'identification, avec l'association de plusieurs méthodes de traitement.

L'une des premières études sur les paramètres prépondérants qui affectent les blooms a été développée par Wyatt et Horwood, 1973 [24]. Depuis, d'autres études en vue le jour, les paramètres utilisés sont ceux liés à la croissance du plancton dans un environnement homogène contrôlé [25], ou à l'évolution des conditions physiques comme les changements d'éléments nutritifs [26], ou aux conditions d'écoulement des rivières et océans [27], ou même à la chaine déterministe des prédateurs-proies [28,29].

De même, les méthodes de traitement et les outils de mesure sont divers. Elles sont utilisées pour caractériser les HAB par la détection, l'identification, la classification et la prédiction.

Plusieurs techniques pour la prédiction de HAB existent : la méthode manuelle [30,31] et le raisonnement par cas CBR (Case-Based Reasoning) [32,33].

Dans leur travail, Kim et Li [31, 32] analysent manuellement les algues en question pour prévoir le HAB.

Song et al. [32] ont proposé un système de surveillance de HAB en utilisant le raisonnement par cas où la méthode utilisée est la classification par les k plus proches voisins kNN (k Nearest Neighbor).

Fdez-Riverola [33] a proposé un système de prévision pour prédire les HAB, sa méthode utilise un modèle de CBR et un modèle flou pour fournir une prédiction de l'événement.

Alors que les deux approches précédentes (manuelle et CBR) peuvent souffrir de la non-linéarité et l'incertitude des connexions entre paramètres. Les techniques ML (Machine Learning) peuvent gérer la relation non linéaire entre l'efflorescence du phytoplancton et les divers paramètres de l'eau. Ces dernières années, les ML sont des outils très utilisés pour prédire les HAB. Ces techniques comprennent le réseau de neurones artificiels (ANN) [34-45], la machine à support de vecteurs (SVM) [44, 46, 47], les arbres de classification [48-51], la théorie floue [52,53], le traitement d'image [54-57] et les méthodes hybrides [58-60]...etc.

2.2.2.1. Réseaux de neurones

Il y a plusieurs dizaines d'articles concernant l'utilisation des réseaux de neurones pour la prédiction des HAB, nous citons chronologiquement plusieurs articles intéressants.

Z. Rong et al [38], effectuent des recherches sur la prédiction des HAB basée sur un réseau de neurones flou.

Z. Liu et al [35], propose une modélisation à court terme pour la prédiction des HAB basé sur un réseau de neurones RBF, où plusieurs architectures de ce réseau ont été testées.

L. O. Teles et al [36], ont utilisé un réseau de neurones de régression généralisée (GRNN) pour prédire l'apparition de HAB dans Crestuma réservoir ; qui est une source d'eau potable importante pour la région de Porto, Portugal. Ces modèles peuvent potentiellement être utilisés pour fournir aux opérateurs des usines de traitement de l'eau un avertissement précoce pour le développement des HAB. Les paramètres biologiques, physiques et chimiques collectés ont été divisés en trois séries de temps indépendant et dont chacune a une périodicité bimensuelle.

W. Xiaoyi et al [37], proposent une méthode intelligente pour la prédiction à court terme des blooms par réseau neuronal de type Perceptron multicouche (MLP), basé sur les données traitées par la théorie des ensembles rugueux et l'analyse en ondelettes. Cette méthode analyse les facteurs qui affectent le déclenchement de bloom. Ces facteurs ont été alors traités par la méthode des ensembles rugueux pour réduction d'espèce d'entrées, ainsi les entrées principales obtenues sont analysées par ondelettes multi résolution, pour éliminer les facteurs de brouillage. Le réseau MLP établit la relation non linéaire entre les facteurs d'entrée et donne comme résultat la prédiction de bloom.

W. Xiaoyi et al [38], proposaient une amélioration par la méthode de prédiction des blooms à base de réseau neuronal gris MLP. Un système de surveillance à distance de l'environnement et d'alerte précoce de bloom basé sur la technologie de communication sans fil GPRS est construit. Le système peut obtenir en temps réel et automatiquement des informations de changement de la qualité de l'eau et prédire l'efflorescence. Il présente un de système efficace et pratique pour le contrôle de l'environnement de l'eau.

S. Zhu et al [39], proposent une méthode de prévision par réseau neuronal artificiel gris MLP. La théorie grise a été utilisée pour obtenir des prévisions préliminaires de blooms,

combinée avec un réseau de neurones pour mettre en œuvre la compensation d'erreur pour le résultat de prévision. Comparée avec MLP, cette méthode peut détecter les changements de la chlorophylle avec plus de précision, améliorer considérablement la précision et prolonger la période de prédiction. Elle fournit une nouvelle méthode efficace pour la prévision à moyen terme des blooms.

Dans leur travail, S. Cho et al [40], ont pour objectif la détermination des facteurs qui influent sur l'efflorescence d'algues et prédire les niveaux de la chlorophylle *a* dans un barrage à l'aide d'un ANN.

L. Zaiwen et al [41], sélectionnent une méthode pour confirmer et identifier les variables dominantes à court terme. Après, une méthode de détection souple de HAB basée sur réseau de neurones RBF a été proposée. La capacité du réseau neuronal est discutée en fonction de différents nœuds de la couche cachée du réseau de neurones RBF. Les résultats montrent que le modèle de détection basé sur le réseau de neurones RBF possède des grandes capacités de prévision et de précision ; c'est un nouvel outil de recherche pour la prévision des HAB dans les rivières et les lacs.

L. Siying [42] traite le problème de la prédiction de HAB en court terme. Les facteurs importants des HAB sont étudiés et un modèle de prédiction à court terme basé sur le réseau neuronal d'Elman est présenté. L'algorithme d'Elman est d'abord amélioré, puis le modèle de prédiction est formé, testé et comparé avec le modèle MLP. Les résultats expérimentaux montrent que le changement à court terme de la chlorophylle peut être mieux prédit par le modèle de prédiction d'Elman.

Le travail de H. Luo et al [43], consiste à établir un modèle prédictif de prolifération des algues dans le lac de Three-Gorge Dam en chine par le réseau de neurones. Une collection des données physiques et chimiques a été réalisée.

Xiu Li et al [44] réalisent une étude sur l'efficacité de quelques méthodes ML. Ils utilisent alors les données issues de Tolo Harbour (Hong Kong, chine) pour former plusieurs méthodes d'apprentissage comme des modèles de prédiction de la prolifération d'algues. Trois différents types de modèles sont conçus : le réseau de neurones MLP, le réseau neuronal à régression généralisée GRNN et la machine à support de vecteurs SVM . Les résultats expérimentaux montrent que l'algorithme MLP amélioré et SVM fonctionnent mieux que les méthodes GRNN.

J. Kuo et al [45], utilisent un réseau neuronal MLP pour relier les facteurs clés qui influent sur la prolifération algale dans un réservoir dans le centre de Taiwan. Les résultats de l'étude montrent que le réseau neuronal est capable de prédire l'efflorescence avec une précision raisonnable.

2.2.2.2. Machine à vecteurs de support

La méthode de machine à vecteurs de support est une autre méthode souvent utilisée pour l'analyse et les prédictions des HAB.

En général les machines à vecteurs de support SVM sont des méthodes et des techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les intérêts majeurs des SVM sont leurs capacités de travailler avec des données de grandes dimensions, leurs identifications mathématiques (une généralisation des classificateurs linéaires), et leurs résultats en pratique. Dans notre travail, les SVM ne sont pas utilisés.

Parmi les travaux réalisés pour la prédiction des HAB, Y. Xu et al [46] présentent une méthode de détection et de prédiction des HAB par SVM. Il propose une estimation basée sur les paramètres d'environnement pour identifier le bloom. Cette étude a démontré que les variables les plus importantes pour prédire la prolifération d'algues dans l'eau de mer sont les éléments nutritifs, la température et l'oxygène dissous.

Z. Liu et al [47] proposent un modèle de prédiction intelligent des blooms dans les rivières et les lacs basés sur LSSVM (Least Squares Support Vector Machine). Les données étaient traitées par la méthode des ensembles rugueux pour la réduction d'espace d'entrées. Puis, ce modèle est comparé avec le modèle d'un réseau neuronal artificiel de prédiction.

Le travail de X. Li et al [44] montre que les modèles basés sur SVM présentent des meilleures performances que les réseaux de neurones. Les résultats indiquent que l'utilisation des données à court terme peut simuler la tendance générale de la biomasse d'algues mais ne sont pas idéales pour des prédictions exactes. L'utilisation des données de plus hautes fréquences peut améliorer la précision des prévisions.

2.2.2.3. L'arbre de décision

L'utilisation des arbres de classification est récente dans le demain de la prédiction et l'identification des HAB. l'arbre de classification ou l'arbre de décision est une méthode couramment utilisée afin de créer un modèle qui prédit une cible (classe prédite) en fonction de plusieurs variables d'entrée. L'arbre de décision est une représentation simple pour classer de nouveaux exemples. Cette méthode sera détaillée ultérieurement dans le chapitre 3 section 3.3.3.

Plusieurs recherches utilisent ces classificateurs pour la prédiction et l'identification des HAB. Parmi eux, le travail d'A. Peretyatko et al [48], où une classification des données environnementales pour prédire les blooms avec le choix des facteurs les plus importants liés à la présence de blooms est présentée. Ils critiquent les méthodes statistiques classiques basées sur des relations linéaires qui se reflètent négativement sur la capacité de prévision. Par contre, les arbres de classification sont conçus pour le traitement des données complexes [49] et sont les plus appropriés pour la prédiction de HAB. Les principaux objectifs sont d'utiliser les arbres de classification pour:

(1) Déterminer l'importance relative des facteurs environnementaux mesurés pour le contrôle de la prolifération.

(2) Quantifier le risque de bloom correspondant à des conditions environnementales déterminées par les facteurs ayant le plus fort impact sur les algues dans les bassins étudiés.

(3) Vérifier si les résultats produits par les arbres de classification sont compatibles avec ceux produits par l'approche probabiliste présentée dans [50].

L'article de S. Park et al [51] présente une méthode de prédiction de Bloom utilisant un arbre de décision. La méthode proposée permet d'améliorer la précision de prédiction, car le classificateur d'arbre de décision est renforcé par une base de données importante.

2.2.2.4. Logique floue

La logique floue est une méthode de classification mondialement reconnue par ces résultats de classification des données dans tous les domaines. Plusieurs études utilisent ces classificateurs pour le suivi des HAB.

Parmi ces travaux, celui de J. Laanemets et al [52], où un modèle de logique floue qui décrit l'évolution saisonnière des blooms de Nodularia spumigena dans le golfe de Finlande a

été construit et calibré sur la base des données de surveillance. Le test du modèle flou pour la prédiction de la biomasse maximale de N. Spumigena prédit l'apparition de blooms environ un mois auparavant.

S. Park et al [53], proposent un modèle de prévision des HAB sur la côte de la Corée du Sud en utilisant un modèle flou combiné avec un modèle d'arbre de décision.

2.2.2.5. Traitement des images

La méthode basée sur le traitement des images attire aussi l'attention des chercheurs, suite au développement des technologies des capteurs (miniaturisation) et de nouvelles méthodes de traitements d'images (minimisation de la consommation d'énergie). Parmi les travaux réalisés, J.K. Choi et al [54], ont étudié l'applicabilité de GOCI (Geostationary Ocean Color Imager) pour le suivi de la distribution et du mouvement temporel d'un HAB. GOCI est un nouveau satellite géostationnaire d'imagerie couleur pour l'océan. Il recueille des images horaires prises durant la journée, ce qui lui permet le suivi de la variabilité temporelle des algues dans l'océan.

B. Zhang et al [55] ont étudié la combinaison entre les données terrestres et l'image satellitaire pour l'identification de la concentration en chlorophylle a pour la surveillance d'un lac d'eau douce en chine.

W. Qingyu et al [56] ont étudié aussi la combinaison entre les données terrestres et l'utilisation d'image satellitaire pour l'identification de la concentration en chlorophylle a. Ils ont abouti à la création d'un modèle de prédiction de HAB en se basant sur les conditions météorologiques et hydrologiques. Avec ces deux modèles, un système d'alerte et de surveillance dynamique et automatique des HABs dans le lac Taihu (chine) a été développé.

Pettersson et al [57], ont utilisé des images de SeaWiFS pour la prédiction des HAB dans les eaux Norvégiennes.

2.2.2.6. Méthodes hybrides

Afin d'avoir de meilleurs résultats pour le suivie des HAB, plusieurs méthodes hybrides ont apparu. Il est impossible passer en revue toutes les combinaisons faites par les différents chercheurs. Nous allons vous présenter quelques-unes.

Q. Chen et al [58] utilisent un modèle d'arbre de décision pour prédire qualitativement l'instant du bloom. Puis, un modèle de régression non linéaire par morceaux pour prédire

quantitativement l'intensité de HAB (concentrations cellulaires). L'algue en question est le *Phaeocystis globosa* présente dans les eaux côtières néerlandaises en mer du Nord. La recherche démontre que les arbres de décision et la régression nonlinéaire par morceaux sont des techniques alternatives très prometteuses dans la modélisation de HAB.

Rousseeuw et al [59] propose un système de surveillance combinant la méthode K-means et un modèle de Markov caché afin de comprendre la dynamique des phytoplanctons et prévoir la prolifération. Les états du modèle de Markov cachés et les codifications des symboles sont obtenus grâce à l'algorithme K-Means.

Z. Wang [60] présente une méthode de modélisation capteur/logiciel dite AlgaeSense. Les données utilisées sont recueillies par des capteurs simples afin de prédire un bloom. L'information prédite est utilisée pour surveiller la prolifération des algues en temps réel et déclencher des alertes. Les images de surface d'eau issue d'une caméra sont utilisées comme entrées du modèle de prédiction avec les données physico-chimiques de l'eau. La taille du bloom est calculée à partir de ces images. Une modélisation par la méthode de régression gaussienne est proposée puis une comparaison est faite avec trois autres méthodes : les réseaux de neurones MLP, BNN (Bayesian Neural Network) et la régression linéaire multiple LMR (linear Multiple Regression).

Pour combler ces lacunes et améliorer la performance des ANN, les chercheurs ont proposé des méthodes améliorées [61, 62].

Les types d'algues et les paramètres du milieu sont des facteurs qui varient d'une zone à une autre. Un modèle de surveillance d'un type d'algue peut fonctionner dans des zones et pas dans d'autres, même si le type d'algue ne change pas. Donc, chaque étude donne un modèle spécifique du milieu où les données sont prélevées.

Dans notre travail, nous avons utilisé un réseau de neurones de type MLP. Notre choix a été fixé, vu les bons résultats donnés dans toutes les études écologiques faites [63-65] pour le suivi des HAB.

La modélisation de suivi des HAB est relative à un type d'algue. Dans notre cas, nous allons d'abord introduire l'algue étudiée en question qui est le Dinophysis Acuminata.

2.2.3. L'efflorescence de D. Acuminata

2.2.3.1. Dinophysis Acuminata

Dinophysis Acuminata est une espèce de plancton marin appartenant au dinoflagellé phylum que l'on trouve dans les eaux côtières des océans atlantique nord et du pacifique. Elle mesure de 30 à 35 µm de longueur et de 38 à 58 µm de diamètre. Son corps est de couleur brun-rougeâtre et elle est recouverte d'une armure appelée thèque [66] ; la figure 2.2 présente la vue microscopique de D. Acuminata.

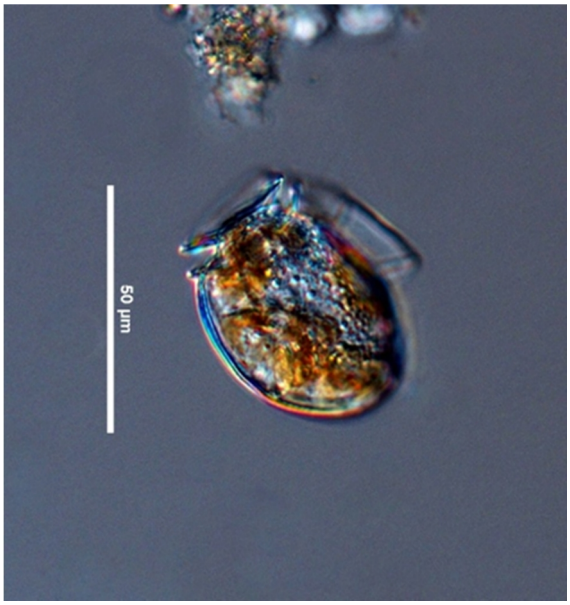

Figure 2.2. Vue microscopique du D. Acuminata.

D. Acuminata est l'une de plusieurs espèces de Dinophysis phototrophiques classées comme toxiques, car elle produit de l'acide okadaïque. La consommation de ces espèces provoque une intoxication diarrhéique sévère DSP (Diarrheic Shellfish Poisoning).

Les proliférations de D. Acuminata sont des menaces constantes. Ce dinoflagellé est responsable de plusieurs blooms toxiques de type DSP. La commercialisation des coquillages des lagunes infectées est interdite pendant plusieurs semaines. Les pertes économiques et sanitaires ont poussé les conchyliculteurs à réagir à cette menace. Ils ont sollicité leurs partenaires scientifiques et techniques pour trouver des solutions à ce phénomène [67-69].

Beaucoup d'incertitudes scientifiques planent sur les facteurs responsables des efflorescences de ce dinoflagellé. La première réussite de la culture du D. Acuminata en milieu artificiel était réalisée en 2006, mais pas suffisamment longtemps pour étudier son cycle biologique et ses conditions optimales de développement [70].

Des travaux menés sur les côtes bretonnes ont révélé que certaines conditions hydrologiques, telles que la stratification des masses d'eaux en couches de température et la salinité différente, favorisent le développement du D. Acuminata [71].

2.2.3.2. Analyse d'influence des différents facteurs

La croissance du phytoplancton est liée à deux types de facteurs : les facteurs physiques (comme la lumière, la température, la turbulence, la turbidité... etc.) et les facteurs nutritionnels (les sels azotés, phosphates, silicates, d'origine naturelle ou anthropique [72]).

L'étude de l'influence de ces paramètres aide fortement à la création de modèles plus compréhensibles. Les facteurs les plus influents sur la croissance cellulaire du D. Acuminata sont les suivants :

- **La température de l'eau:**

Des températures élevées assurent une prolifération optimale et favorisent le processus de développement [73].

- **L'ensoleillement:**

Il peut intervenir comme un facteur stimulant. Il présente la source d'énergie qui permet au phytoplancton d'élaborer toutes les substances organiques pour sa croissance. L'ensoleillement ou l'éclaircissement solaire agit sur la division des cellules. La chlorophylle et autres pigments absorbent surtout l'énergie contenue dans les longueurs d'onde comprises entre 0.4 µm 0.5 µm ou entre 0.61 µm 0.69 µm[73,74].

- **Les précipitations:**

Des précipitations abondantes peuvent favoriser l'enrichissement des eaux en substances nutritives par le lessivage des sols ce qui contribue à la croissance et au maintien en vie du phytoplancton [74,75].

- **Les sels nutritifs:**

Les nitrates, les phosphates... etc., proviennent de la décomposition de matières organiques. L'azote existe en mer sous les formes minérales suivantes: Ammonium NH_4^+, nitrite NO_2^- et nitrate NO_3^-. L'ammonium et surtout les nitrates sont abondants dans les conditions naturelles [76]. Le phosphore est un élément constitutif nécessaire, il existe en quantité réduite dans le milieu marin [76].

- **La salinité:**

C'est un facteur conditionnant la distribution et la croissance de 20 à 40 % d'espèces phytoplanctoniques. Par expérience, un pH variant entre les valeurs 7.4 et 8.8, n'influe pas sur la croissance de la culture du phytoplancton. Cette expérience couvre largement les variations de pH observées en mer libre [76].

- **La concentration de gaz carbonique :**

Selon Martin Y. [77] la concentration de gaz carbonique CO2 affecte la croissance cellulaire des cyanobactéries.

- **Les vents, les marées et les courants:**

Ce sont des facteurs favorisant la dispersion et le transport du phytoplancton.

2.2.4. Modélisation de l'efflorescence de D. Acuminata

La modélisation des processus biologiques est caractérisée par:

- Le grand nombre de paramètres décrivant le processus allant jusqu'aux mesures du métabolisme cellulaire (Une grande base de données) [78].
- La description détaillée des processus cellulaires exige des mesures obtenues avec des instruments très chers et des temps d'analyse trop longs [79].

Les modèles les plus efficaces sont ceux qui sont limités aux variables les plus influentes (prépondérants). Le modèle que nous proposons utilise un nombre limité de paramètres qui affectent la croissance cellulaire et qui sont mesurés directement sur le site : modèle de la figure2.3.

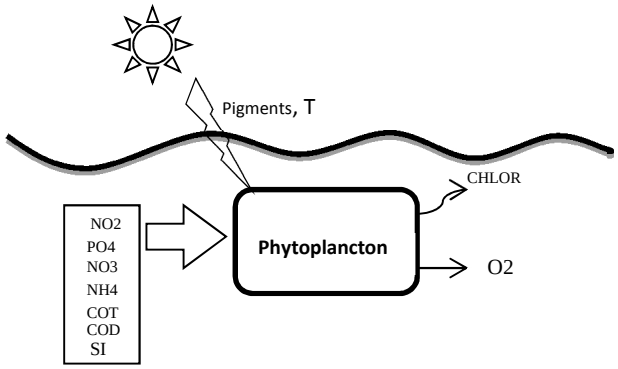

Figure 2.3. Paramètres qui affectent la croissance cellulaire.

Le modèle est réalisé avec une base de données composée de 37 échantillons prélevés de l'eau de mer du littoral français (Havre France) durant la période de juillet, août et septembre de l'année 1985. Les paramètres environnementaux sont représentés dans le Tableau 2.1.

Tableau 2.1. Paramètres environnementaux mesurés.

Variable	Paramètre	Unité de mesure
T	Température	°C
O2	Oxygène	mg/l
COD	Carbone Organique Dissous	ppcm
COT	Carbone Organique Total	ppcm
PO_4	Phosphate	µatg/l
NH_4	Ammonium	µatg/l
NO_2	Nitrite	µatg/l
NO_3	Nitrate	µatg/l
SI	Silice	µatg/l
CHLOR	Chlorophylle	mg/m3
Pigm	Pigments	mg/m3

Les données présentent l'évolution de la concentration cellulaire du D. Acuminata en fonction de paramètres physico-chimiques. Les propriétés de cette base de données sont:

- ✓ La durée courte de la série temporaire qui permette une modélisation sur des courts termes. Le modèle que nous développons utilisera les données mensuelle ou hebdomadaire pour générer des prédictions à court terme. Dans ce travail une prédiction de 10 jours est générée après un apprentissage basé sur des données bimensuelles.
- ✓ Les mesures effectuées en 1985 sont devenues maintenons les plus basic, c'est-à-dire les moins chers. De nos jours, un seul instrument peut effectuer deux ou trois mesures de cette base.
- ✓ Le type de phytoplancton D. Acuminata qui nous permet de modéliser un HAB.

Parmi ces variables d'environnement, certaines affectent de manière directe la croissance cellulaire algale (la température, les pigments solaires, les sels nutritifs, les carbones organiques et la salinité) et d'autres sont des résultats de la photosynthèse de cellule (chlorophylle et oxygène).

À partir de la base de données nous allons créer un modèle de type boite noire qui a pour entrées les paramètres physico-chimiques et qui peut prédire la concentration cellulaire de phytoplancton.

À partir de la section (2.2.2.), nous avons jugé que l'utilisation d'un réseau de neurones artificiel est la plus adaptée pour notre système. Nous allons détailler chaque phase de traitement de notre système. Le schéma bloc de la figure 2.4 montre toutes ces phases.

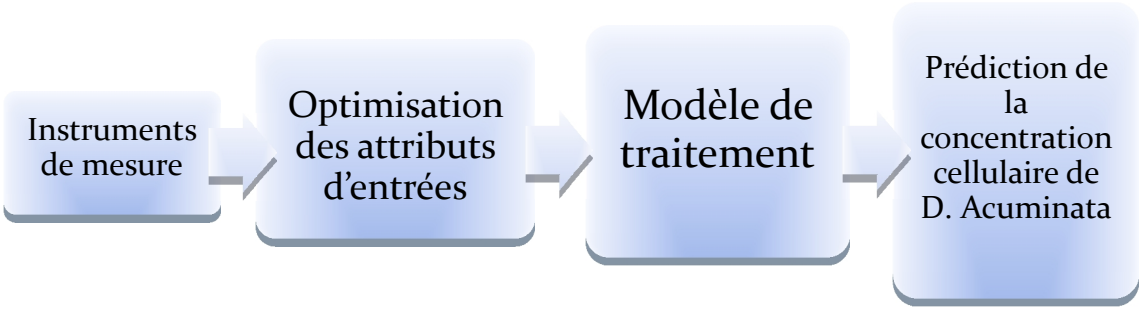

Figure 2.4. Chaine de traitement du système de prédiction proposé

2.3 Système de prédiction de D. Acuminata

Dans cette section nous allons détailler chaque étape de notre système de prédiction du D. Acumunata.

2.3.1. Acquisition des données et instruments de mesure

2.3.1.1. Données physiques

Les prélèvements de température, salinité et pigment lumineux sont réalisés par plusieurs types de capteurs. Ces capteurs sont des sondes numériques multiparamétriques qui réalisent une ou plusieurs mesures à la fois. On parle alors des capteurs intelligents qui intègrent les fonctionnalités de mesure, traitement et transmission.

Les instruments largement utilisés sur site sont les sondes CTD (Conductivity, Temperature, Depth). Un exemple est présenté à la figure 2.5. Les CTD sont conçus

principalement pour mesurer les paramètres physiques de l'eau : température, conductivité, salinité et pression. Leurs conceptions modulaires rendent facile la configuration sur le champ pour une large gamme de capteurs auxiliaires et optionnels tels que : les capteurs d'oxygène dissous, de pH, de fluorescence, de transmissivitée et de rétrodiffusion optique. Les sondes CTD fonctionnent avec des batteries et les données ainsi captées peuvent être enregistrées en mémoire interne ou transmises par port série RS-232 optoisolée.

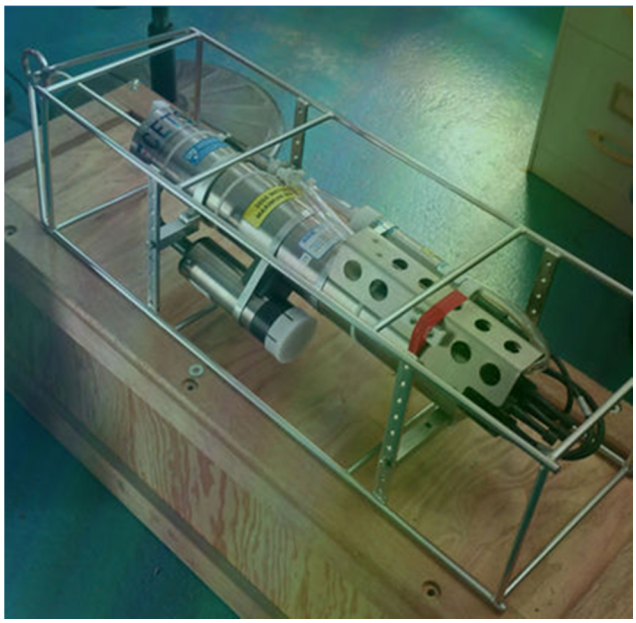

Figure 2.5. Exemple d'une sonde CTD [80].

La température est mesurée avec deux techniques [81] :

- **Les thermomètres à résistance de platine.**

La mesure de température est couramment réalisée par un thermomètre à résistance de platine auquel on adjoint une thermistance. Un tel montage permet d'avoir un temps de réponse suffisamment rapide et avec une bonne précision. L'effet de la pression sur le fil de platine diminue sa résistance, donc le capteur est protégé par un tube inoxydable aussi fin que possible.

- **Thermistance et chaîne de thermistances.**

Une thermistance est constituée d'un mélange d'oxydes métalliques semi-conducteurs (oxyde de manganèse, nickel, cobalt) qui se présente sous forme de poudre agglomérée par frittage à haute température et haute pression. La résistance d'une thermistance varie en sens inverse de la température. Les chaînes de thermistances ou chaînes bathythermiques sont constituées d'une série de thermistances disposées régulièrement dans un tube protecteur

rempli d'huile. Elles sont utilisées pour le suivi régulier du profil de la température sur des couches d'eau de quelques dizaines de mètres.

D'autres instruments sont aussi utilisés pour les analyses des larges zones comme les Gliders ou les glisseurs. Ils sont des planeurs sous-marins autonomes réutilisables. Plusieurs types des Gliders sont présentés à la figure 2.6. Récemment développés aux États-Unis et en France. Ces instruments sont conçus pour glisser sous l'eau dans une direction donnée de la surface des océans jusqu'à une profondeur prédéterminée et ensuite remonter en surface comme illustre la figure 2.7. Ils mesurent les paramètres physiques (température, salinité, O2, chlorophylle, rétrodiffusion optique…) le long de la trajectoire destinée réalisant ainsi un très bon échantillonnage à travers l'océan. Les données sont enregistrées dans l'appareil ou transmises par ondes hertziennes.

Figure 2.6. Exemples des Gliders [82].

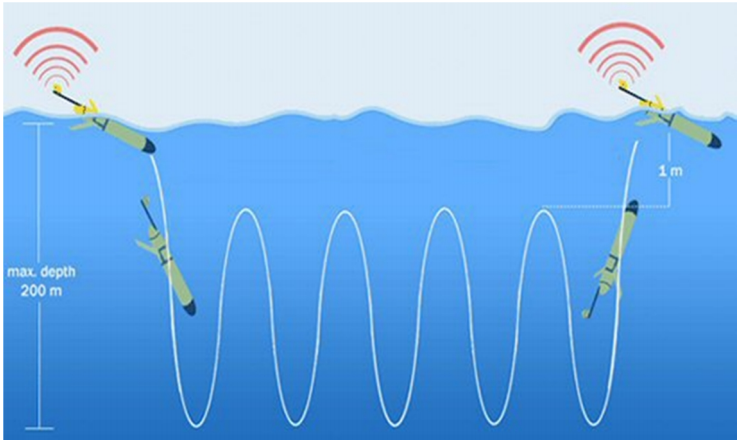

Figure 2.7. Mouvements des Gliders [83].

Dans tous les cas, les prélèvements présentent des profils verticaux à une vitesse de descente uniforme, cette vitesse influence la qualité des mesures. Après les données correspondant aux prélèvements de surface telle que la température et la salinité sont moyennées. D'autres paramètres comme les pigments lumineux doivent être traités ou calculer.

2.3.1.2. Données chimiques

2.3.1.2.1. Oxygène dissous [81]

La concentration en oxygène dissous est mesurée à l'aide d'un capteur polarographique dont son schéma de principe est illustré sur la figure 2.8. Deux électrodes, une en argent et l'autre en or, baignent dans un électrolyte constitué de chlorure de potassium isolé du milieu extérieur par une membrane en téflon. Un potentiel constant est appliqué au niveau des électrodes.

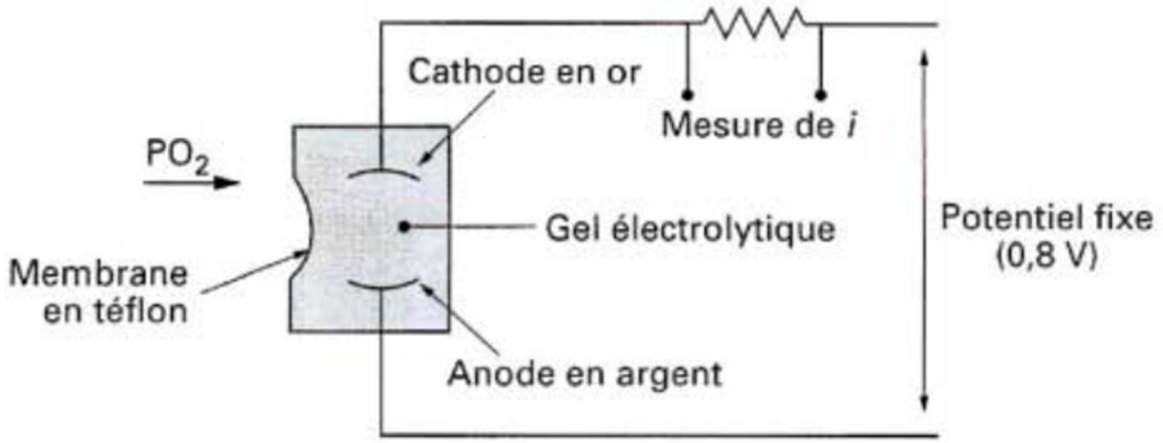

Figure 2.8. Schéma de principe de la mesure de l'oxygène dissous par capteur polarographique.

L'oxygène dissous dans l'eau de mer diffusé à travers la membrane est décomposé au niveau de la cathode suivant une réaction chimique. L'intensité du courant est directement proportionnelle à la pression partielle d'oxygène dissous.

La mesure oxygène de la sonde est relativement instable, ce qui oblige à faire une moyenne. Pour calibrer la sonde, des prélèvements à différentes profondeurs sont effectués.

2.3.1.2.2. Sels nutritifs

On présente ici deux façons pour mesurer les dosages des nutriments:

a) Méthode colorimétrique, à l'aide de bandelettes à réactif coloré:

Les méthodes colorimétriques sont les plus utilisées, ils fournissent des mesures extrêmement précises. Le dosage par colorimétrie est couramment utilisé pour quantifier les ions nitrite, nitrate, ammonium, phosphate.

Le dosage colorimétrique repose sur la quantification de produits colorés, issus d'une réaction chimique. Elle est possible lorsque l'intensité de la coloration est proportionnelle à la concentration de l'élément à doser. Les dosages colorimétriques s'appuient sur la loi de Lambert-Beer et les protocoles de prélèvement. La mesure, l'étalonnage des échantillons et le dosage colorimétrique changent par rapport aux standards de référence d'un élément chimique à l'autre. Par exemple pour l'ammoniac la méthode de Koroleff [84], pour les nitrites la méthode de Bendshneider et Robinson [85], pour les nitrates la méthode de Woods [86], et pour les phosphates la méthode de Murphy et Riley [87]. Après, le dosage colorimétrique est analysé par spectrophotométrie d'absorption, pour déterminer la concentration des ions dans la solution.

Les laboratoires ou les navires scientifiques utilisent aujourd'hui des colorimètres en flux continu. L'appareil le plus couramment utilisé est l'autoanalyseur Technicon, régulièrement embarqué sur les navires océanographiques. C'est un appareil de laboratoire qui opère à partir d'échantillons. À l'aide de ce système d'analyseur automatique, il est possible d'effectuent des mesures automatiques de manière cyclique par rapport aux standards de référence.

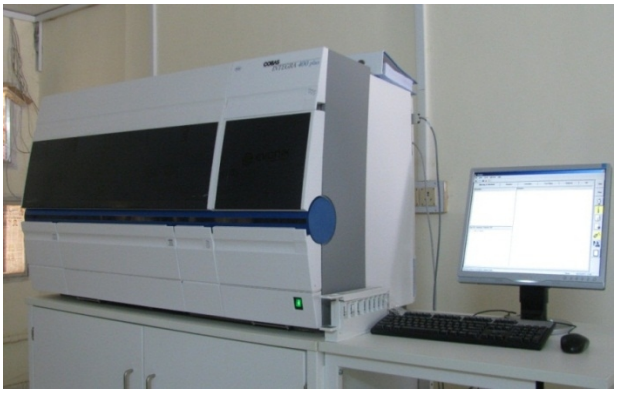

Figure 2.9. Exemple d'un autoanalyseur colorimétrique [88].

b) Méthode électrochimique à l'aide d'une sonde à capteur chimique :

Les capteurs électrochimiques permettent de mesurer directement sur site les différents paramètres nutritifs par la surveillance de la quantité des espèces ioniques en milieu liquide, ce qui fournit une haute résolution à l'échelle temporaire. Ils permettent la détection des pics rapidement. Il existe trois modes de détection : potentiométrique, ampérométrique et conductimétrique. Ils sont tous basés sur des principes de détection électrochimique [89].

✓ **Capteurs potentiométriques [89]**

Une différence de potentiel est mesurée entre une électrode de référence dont le potentiel reste invariable quelle que soit la composition ionique de la solution dans laquelle elle est plongée et une électrode redox constituée de matériaux conducteurs électroniques permettant des échanges d'électrons avec tous ou certains ions contenus dans la solution. La concentration des ions est obtenue par mesure de différence de potentielle entre les électrodes.

Pour la détection des espèces ioniques des sels nutritifs, les électrodes spécifiques ISE (Ion Selective Electrode) sont utilisées. La figure 2.10 présente un exemple de ces capteurs. Ils mettent en jeu des équilibres électrochimiques aux interfaces entre des électrolytes liquides et des membranes constituées de matériaux conducteurs ioniques (électrolytes solides, polymères conducteurs ioniques, membranes liquides).

La nature de l'ion échangé est déterminée par la composition du matériau constituant la membrane. On choisit le matériau de membrane pour que l'échange ne soit possible que pour un ion déterminé, alors la mesure est spécifique pour cet ion.

Figure 2.10. Exemple capteur potentiométrique de type ISE [90].

- ✓ **Capteurs ampérométriques [89,91]**

Un capteur ampérométrique s'intéresse au courant qui traverse une solution. L'intensité est en fonction de la tension imposée et des espèces chargées présentes dans la solution. La détermination de la concentration de certains éléments est possible après un étalonnage si l'on connait la plupart des autres éléments présents dans la solution et leur participation à l'électrolyse.

La mesure s'effectue avec une électrode redox et une électrode de référence. La tension est fixée telle que l'ensemble de l'élément sera sous forme réduite et l'intensité traversant la solution est alors proportionnelle à la concentration du dit élément. La figure 2.11 illustre le principe de fonctionnement et un exemple de capteur.

Généralement, les courants mesurés varient entre le pico ampère et le milliampère. Ils dépendent de la tension appliquée, de la solution, de la température et des électrodes ainsi que leur état de surface.

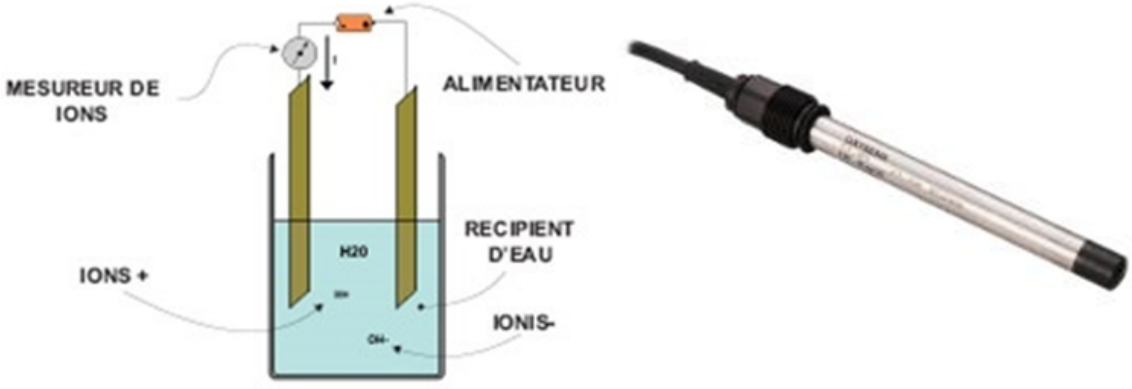

Figure 2.11. Principe et exemple d'un capteur ampérométriques [92].

- ✓ **Capteurs conductimétriques [91]**

Ces types de capteurs s'intéressent à la conductivité des matériaux. La mesure de la conductance d'une solution électrolytique s'effectue en immergeant dans la solution une cellule de mesure comportant deux électrodes dont la surface et la distance sont déterminées par étalonnage dans une solution de conductivité connue.

La conductivité dépend des espèces présentes dans la solution, on l'appelle aussi conductance spécifique. La mesure de conductance permet de déterminer la conductivité.

Cette mesure ne s'effectue pas en courant continu, car les électrodes se polariseraient comme en potentiométrie et cela fausserait les résultats. Un courant alternatif est utilisé entre 50 Hz à quelques kHz. Pour avoir de bons résultats il faut adapter la fréquence à la solution à analyser (plus la conductivité est élevée plus la fréquence doit l'être aussi).

Pour des solutions à conductivité faible, on choisit une surface grande par rapport à la distance interélectrodes et l'inverse pour des conductivités élevées. Il existe beaucoup de capteurs différents, mais cette différence repose quasi uniquement sur leur géométrie, leur choix est donc orienté en fonction du milieu à mesurer. Un exemple de ces capteurs est présenté à la figure 2.12.

Figure 2.12. Exemple de capteur conductimétrique.

2.3.1.2.3. Carbone

Principalement, il existe trois principes de mesure du taux de Carbone Organique dissous et total qui peuvent répondre à des applications différentes :

a) Oxydation à chaud [93]:

C’est la plus efficace pour des mesures en ligne, elle consiste à oxyder toutes les matières organiques en dioxyde de carbone (CO2) puis à mesurer la quantité de CO2 ainsi produite. L’oxydation à chaud est la méthode qui reproduit le plus exactement la méthode de référence en laboratoire (selon la norme NF EN 1484). Il s’agit d’une oxydation thermique à haute température entre 800-950 °C, allons jusqu’à 1200 °C, selon les analyseurs. Quelle que soit la composition de la matrice organique ou inorganique d’échantillons, toutes les molécules sont cassées. La mesure du CO2 est réalisée la plupart du temps par sonde capteur infrarouge non dispersive (NDIR). La figure 2.13 présente un exemple d’un analyseur en ligne à oxydation chaude.

La principale contrainte provient alors du four, qui s'encrasse rapidement à cause de la présence de sels minéraux formés au cours de la dégradation qui peuvent être aussi à l'origine de fortes corrosions ; ce qui exige une longue maintenance à cause du temps de refroidissement du four. L'oxydation à chaud reste réservée à des procédés chargés avec des molécules organiques complexes.

Figure 2.13. Un exemple d'un analyseur en ligne à oxydation chaude

b) L'oxydation à froid [93]:

Afin de s'affranchir des fours, les fournisseurs d'analyseur en ligne ont mis au point une oxydation à froid. Il s'agit classiquement de casser les molécules en couplant deux actions oxydantes : un traitement par lampe aux ultraviolets et une oxydation chimique (généralement du persulfate de calcium). La figure 2.14 présente un exemple d'un analyseur en ligne à oxydation froide.

Par rapport à une oxydation thermique, le traitement à froid est un peu moins efficace (notamment sur les molécules complexes ou les longues chaînes de carbone). Elle est un peu plus lente aussi (entre 5 à 8 minutes pour un cycle de mesure), mais elle est beaucoup moins salissante. Ces dernières années, des améliorations ont été apportées sur l'optimisation de cette oxydation pour en simplifier le processus, améliorer son rendement et diminuer la consommation de réactifs.

Figure 2.14. Un exemple d'un analyseur en ligne à oxydation froide

c) Méthode optique :

La méthode optique est radicalement différente des deux précédentes. Elle s'appuie sur la capacité des composés organiques à absorber la lumière UV. Cette méthode est très séduisante, car elle n'oblige pratiquement aucune maintenance qu'un petit nettoyage de l'optique de temps en temps. Cette méthode utilise deux types d'équipements : les analyseurs extractifs avec prélèvement et les sondes immergeables directement dans le processus. La mesure peut se faire à une seule longueur d'onde (c'est alors la longueur d'onde à 254 nm qui est choisi, car elle correspond à la réponse la plus importante des matières organiques). D'autres analyseurs balaient une bande spectrale plus large dans l'ultraviolet pour prendre en compte un plus grand nombre de composés organiques. La figure 2.15 présente un exemple d'une sonde UV.

Il existe pour chaque substance un facteur de corrélation différent, lié à sa teneur en carbone. La mesure sera donc plus juste en prenant en compte l'absorption sur une plage de longueurs d'onde plutôt que sur une seule ; surtout dans les eaux qui contiennent de nombreuses substances avec des propriétés optiques différentes ou lorsque la composition, la

couleur, la teneur en matières solides varient. Généralement, les sondes immergeables travaillent à une seule longueur d'onde. Les constructeurs proposent cependant des sondes immergeables intégrant une mesure spectrale dans la plage UV-visible de 200 à 750 nm. Il reste néanmoins que certains composés n'absorbent pas du tout les UV et le principe ne peut évidemment pas être utilisé, et c'est le cas notamment des sucres ou des alcools.

Figure 2.15. Un exemple d'une sonde UV

2.3.2. Optimisation des attributs d'entrées

Pour n'importe quel modèle, la sélection des variables d'entrée est très importante. Cependant, pour les réseaux de neurones, une attention un peu moindre est demandée [95]. L'utilisation de beaucoup d'entrées peut affecter la rapidité du modèle et causer des problèmes de redondance entre les différentes variables [96].

La sélection des entrées est faite à base des connaissances à priori du processus. Les paramètres des blooms du D. Acumunata restent mal saisis. La sélection des attributs d'entrée du modèle peut être faite avec l'analyse en composantes principales (PCA).

2.3.2.1. L'analyse en composantes principales

L'extraction des caractéristiques principales de la base de données est une étape très importante pour la chaine de traitement. Elle permet de réduire considérablement le nombre de dimensions des vecteurs d'entrée d'un système. Cette action diminue aussi la taille du système et le nombre d'opérations de calcul. Il existe des méthodes qui permettent de réaliser

la réduction du nombre de paramètres de manière optimale sans perte d'informations significatives. La régression linéaire multiple, l'analyse factorielle des correspondances, l'analyse discriminante et le réseau de neurones de Kohonen sont quelques exemples de ces méthodes.

Dans notre travail nous avons choisi la méthode d'analyse par composantes principales PCA qui est une technique d'usage très courante pour l'élimination de toutes les corrélations entre les entrées.

La méthode PCA permet de connaître les rapports existant entre les différentes dimensions d'un ensemble de données. Cette technique permet de diminuer la quantité de données initiales en abaissant le nombre de dimensions avec conservation d'information. Les axes contenant les informations dominantes sont pris en considération tandis que ceux présentant une moindre influence sur le processus sont éliminés. Elle semble très prometteuse pour les applications de classification effectuées par les réseaux de neurones.

En effet, plus les données sont optimisées à l'entrée d'un réseau, plus rapidement s'effectuera la classification et la décision du système. Une des limitations de cette technique est son approche linéaire. Si les données sont entachées de bruit, la méthode PCA présente quelques problèmes au niveau de la recherche de la variance maximale.

2.3.2.2. Théorie

Fondamentalement, la méthode PCA est une technique de réduction du nombre de dimensions de l'espace des paramètres d'entrée d'un système. L'espace à plusieurs dimensions initiales est projeté dans un espace de dimension inférieure. L'orientation de cette projection révèle toutes les relations existantes entre chacun de ces points, sans changer leur orientation dans l'espace.

Cette méthode se compare à une forme d'apprentissage non supervisé. Notre objectif est de transformer des vecteurs X_n faisant partie d'un espace de dimension d $(x_1, x_2 \dots x_d)$ en vecteurs Y_n faisant partie d'un espace de dimension M $(y_1, y_2 \dots y_M)$ où $M < d$.

L'équation (2.1) représente le vecteur X comme une combinaison linéaire d'un ensemble d de vecteurs orthonormaux u_i:

$$X = \sum_{i=1}^{d} y_i u_i \qquad (2.1)$$

Où les vecteurs u_i sont calculés par l'équation (2.2)

$$u_i^T u_j = \delta_{ij} \tag{2.2}$$

δ_{ij} Représente le delta de Kronecker qui prend la valeur 1 si $i = j$ et zéro autrement. Les vecteurs u_i sont dans la direction des vecteurs propres de la matrice de covariance. Ces vecteurs sont ordonnés selon l'ordre décroissant des valeurs propres correspondantes. Comme la matrice de covariance est réelle et symétrique, les vecteurs propres sont orthogonaux. Ils forment une base naturelle de représentation des données.

On utilise l'équation (2.3) pour déterminer les coefficients y_i. L'équation (2.3) est une simple rotation du système de coordonnées des paramètres x en un système de coordonnées représentées par les paramètres y.

$$y_i = u_i^T X \tag{2.3}$$

Si nous retenions un sous-ensemble avec seulement M coefficients y_i, et remplacions les coefficients restants par des constantes b_i, alors chaque vecteur X sera approché par l'équation (2.4).

$$\tilde{X} = \sum_{i=1}^{M} y_i u_i + \sum_{i=M+1}^{d} b_i u_i \tag{2.4}$$

Dans l'équation (2.4), les composantes du vecteur X sont représentées par deux composantes dont l'une est de dimension d et l'autre de dimension M. Les valeurs des y_i sont des variables qui varient pour chaque échantillon alors que les b_i sont des constantes.

Considérons maintenant une base de données représentée par N vecteurs X^n où $n = 1, 2, \ldots N$. Notre objectif est de choisir une base de vecteurs u_i et des constantes b_i telles que l'approximation donnée par l'équation (2.4) avec les valeurs de y_i déterminées par l'équation (2.3) puissent donner la meilleure approximation du vecteur X.

L'erreur dans l'approximation de X^n généré par la réduction de dimension est donnée par l'équation (2.5).

$$X^n - \tilde{X}^n = \sum_{i=M+1}^{d} (y_i^n - b_i) u_i \tag{2.5}$$

La meilleure approximation sur l'ensemble des vecteurs est obtenue en minimisant la somme des erreurs au carré sur toute la base de données avec l'équation (2.6).

$$E_M = \frac{1}{2}\sum_{n=1}^{N}\left\|X^n - \tilde{X}^n\right\|^2 = \frac{1}{2}\sum_{n=1}^{N}\sum_{i=M+1}^{d}(y_i^n - b_i)^2 \tag{2.6}$$

En égalant la dérivée de E_M par rapport à b_i, nous trouvons l'équation (2.7).

$$b_i = \frac{1}{N}\sum_{n=1}^{N} y_i^n = u_i^T \bar{X} \tag{2.7}$$

Où $\bar{X}$ représente la moyenne des vecteurs X.

L'équation (2.3) et l'équation (2.7), nous permettent d'écrire la somme des erreurs au carré de l'équation (2.8).

$$E_M = \frac{1}{2}\sum_{i=M+1}^{d}\sum_{n=1}^{N}[u_i^T(y_i^n - b_i)]^2 = \frac{1}{N}\sum_{i=M+1}^{d} u_i^T \Sigma\, u_i \tag{2.8}$$

Où Σ est la matrice de covariance.

Lors de l'étape expérimentale, nous utilisons la méthode PCA pour déterminer le plus grand nombre de paramètres orthogonaux qui caractérisent la base de données. À partir de la méthode PCA, nous découvrirons les entrées de plus faible corrélation. Ensuite, nous construisons quelques regroupements de paramètres.

2.3.3. Modèle de traitement

Dans ce travail, nous avons opté pour la modélisation par réseau de neurones artificiel de type MLP. On s'intéresse uniquement au perceptron multicouche à rétropropagation d'erreur comme modèle de traitement. Pour plus de détails sur les réseaux de neurones, il faut se référer à l'ouvrage [97] est très utile.

2.3.3.1. Perceptron multicouche (MLP)

L'appellation 'le perceptron' a été introduit par Rosenblatt [98]. Les premiers réseaux de types perceptron sont les plus simples. Un exemple d'un perceptron simple est représenté à la figure 2.16. Il est constitué de neurones binaires (fonction d'activation seuil). Il s'agit d'une seule couche de *S* neurones connectés et un vecteur *P* de *R* entrés.

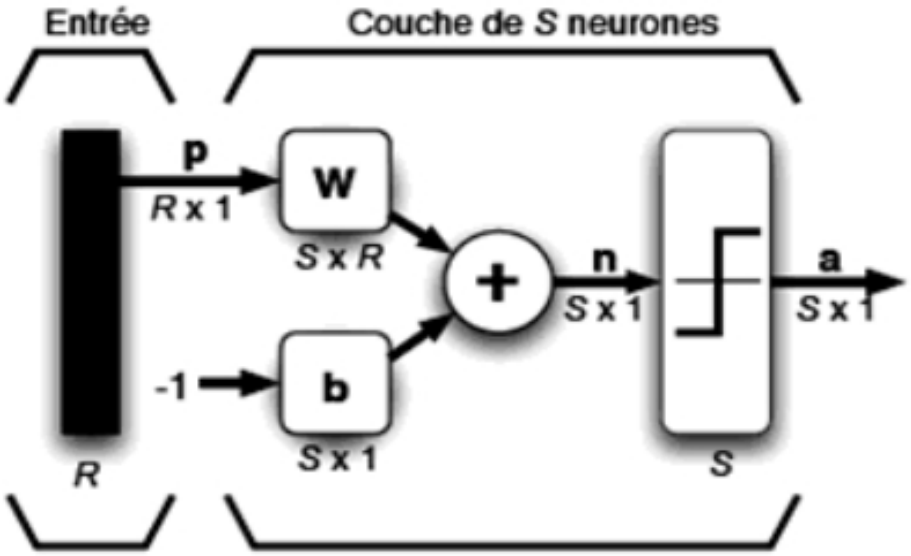

Figure 2.16. Exemple d'un perceptron simple.

La sortie du perceptron peut prendre seulement deux valeurs distinctes selon le niveau d'activation du neurone : '0' s'il est strictement inférieur à 0, sinon c'est un '1'.

La même architecture de réseau à une seule couche peut générer d'autres types de réseaux. Il suffit de changer la fonction de transfert. Le réseau ADALINE (ADAptive LInear NEuron), présenté dans la figure 2.17, est un exemple qui utilise une fonction de transfert linéaire.

Généralement le modèle perceptron ne peut résoudre que des problèmes linéaires. Pour remédier à cet inconvénient, les chercheurs ont proposé une nouvelle architecture qui est le perceptron multicouche.

Un réseau de neurones de type perceptron multicouche ou MLP n'est autre qu'un assemblage de couches en cascade. Les sorties d'une couche sont prises comme les entrées de la couche suivante. Un exemple d'un MLP est représenté à la figure 2.18. Un réseau MLP utile doit toujours posséder des neurones avec fonction de transfert non linéaire sur ses couches cachées et/ou sur sa couche de sortie. Cette condition rendra le réseau utilisable dans le cas de problèmes non linéaires.

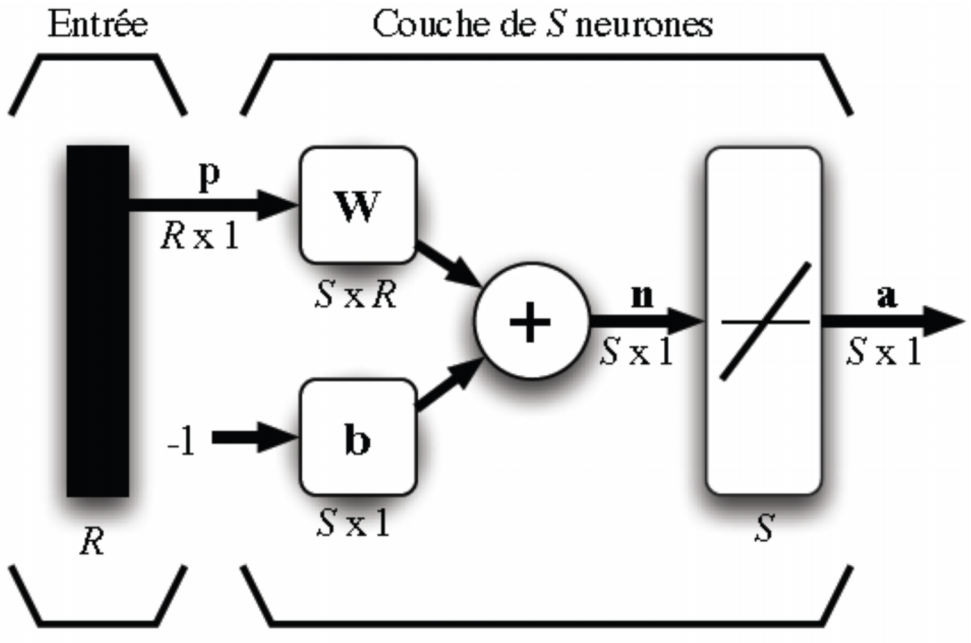

Figure 2.17. Exemple est le réseau ADALINE.

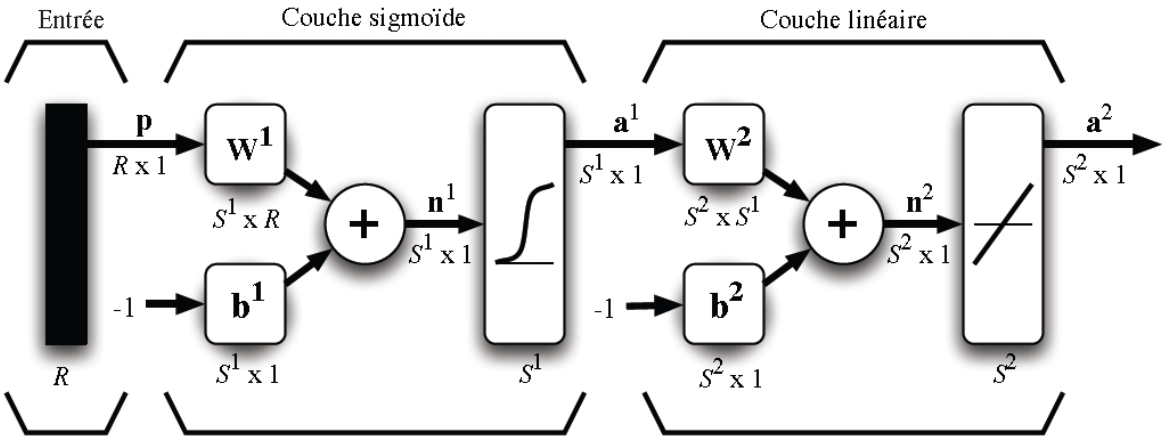

Figure 2.18. Exemple d'un réseau MLP

2.3.3.2. Algorithme d'apprentissage de perceptron

L'algorithme d'apprentissage le plus utilisé pour le perceptron est la règle de Widrow-Hoff qui est basé sur la minimisation de l'erreur quadratique moyenne LMS (Least Mean Square). Il y a aussi la règle de perceptron, mais avec un résultat qui est parfois sensible au bruit [97]. La règle d'apprentissage de Widrow-Hoff permet d'ajuster les poids d'un réseau de neurones pour diminuer à chaque étape l'erreur commise par le réseau (à condition que le facteur d'apprentissage soit bien choisi).

Soient "*p*" et "*y*"**,** respectivement les vecteurs d'entrée et sortie cible qui sont utilisés pour l'apprentissage du réseau et "*a*" comme étant la réponse du réseau.

L'objectif est de minimiser la fonction coût *F*, c'est-à-dire l'erreur quadratique moyenne entre la sortie cible et réponse du réseau. Le coût F est défini par la formule (2.11) :

$$F = \frac{1}{Q}\sum_{k=1}^{Q}[y(k) - a(k)]^2 = \frac{1}{Q}\sum_{k=1}^{Q}[e(k)]^2 \qquad (2.11)$$

Q étant le nombre d'échantillons, "*a*" est la sortie de réseau et "*y*" la sortie désirée. Cette minimisation se fait selon la formule (2.12) qui est représentée par une règle delta:

$$\Delta w = -\alpha \frac{\partial F}{\partial w} \qquad (2.12)$$

Le coefficient α sert à diminuer les changements afin d'éviter qu'ils deviennent trop grands, ce qui peut entraîner des oscillations du poids.

2.3.3.3. Algorithme d'apprentissage de réseau MLP

L'apprentissage d'un MLP est réalisé avec l'algorithme de rétropropagation d'erreur BP (Backpropagation). Il a été créé en généralisant la loi d'apprentissage de Widrow-Hoff pour le perceptron à des réseaux de neurones multicouches constitués de fonction de transfert non linéaire.

Initialement tous les poids peuvent avoir des valeurs aléatoires, qui sont normalement très petites avant de commencer l'apprentissage. La procédure d'apprentissage se décompose en deux étapes : Dans la première étape, les valeurs d'entrées sont introduites sur la couche d'entrée, le réseau propage ensuite ces valeurs jusqu'à la couche de sortie et donne ainsi la réponse du réseau.

À la deuxième étape, les sorties souhaiter sont présentées aux neurones de la couche de sortie, qui calculent la fonction coût ou fonction d'écart ***F***, modifient les poids de la couche de sortie et rétropropagent l'erreur jusqu'à la couche d'entrée pour modifier les poids des neurones cachés.

Le BP standard est un algorithme de descente du gradient dans lequel les poids du réseau sont ajustés dans le sens du gradient négatif de la fonction ***F***. Dans la formule (2.12), la difficulté réside toujours dans le calcul de $\frac{\partial F}{\partial W}$, où de nombreuses techniques, plus ou moins rapides, existent.

Nous ne pouvons pas détailler tous les algorithmes d'optimisation utilisés pour l'apprentissage des réseaux MLP. Nous allons s'attarder sur celui utilisé dans notre travail, en l'occurrence l'algorithme de Levenberg-Marquardt, qui est un algorithme très utilisé et très rapide.

2.3.3.4. Algorithme de Levenberg-Marquardt

L'algorithme de Levenberg-Marquardt, représenté dans la figure 2.19 [97], permet d'obtenir une solution numérique au problème de minimisation d'une fonction qui est souvent non linéaire et dépend de plusieurs variables. L'algorithme interpole l'algorithme de Newton et la méthode de descente de gradient. Plus stable que celui de Newton et du gradient conjugué, l'algorithme Levenberg-Marquardt aboutit à une solution rapide, même avec une initialisation éloignée du minimum.

Cependant, pour certaines fonctions très régulières, l'algorithme Levenberg-Marquardt peut converger légèrement moins vite. Dans le cas du réseau MLP, l'algorithme ajuste les poids de façon à minimiser l'erreur quadratique moyenne MSE (mean square error).

2.1.1.1. Applications

Les deux plus grandes applications des réseaux MLP sont l'approximation des fonctions et la classification.

2.1.1.1.1. Approximation des fonctions

Dans ce travail, nous avons réalisé un modèle d'un processus dynamique nonlinéaire. Les réseaux MLP sont capables d'approximer n'importe qu'elle fonction possédant un nombre fini de discontinuités. Dans la figure 2.20 nous avons présenté un exemple de ces fonctions.

Pour l'approximation d'une fonction, on peut montrer qu'un réseau MLP avec une seule couche cachée de neurones sigmoïdes et une couche de sortie avec des neurones linéaires permet d'approximer n'importe quelle fonction d'intérêt avec une précision arbitraire ; à la seule condition de disposer de suffisamment de neurones sur la couche cachée. Comme pour les séries de Fourier qui utilisent des sinus et cosinus, un réseau de neurones peut approximer n'importe quelle fonction d'intérêt par une combinaison linéaire de sigmoïdes.

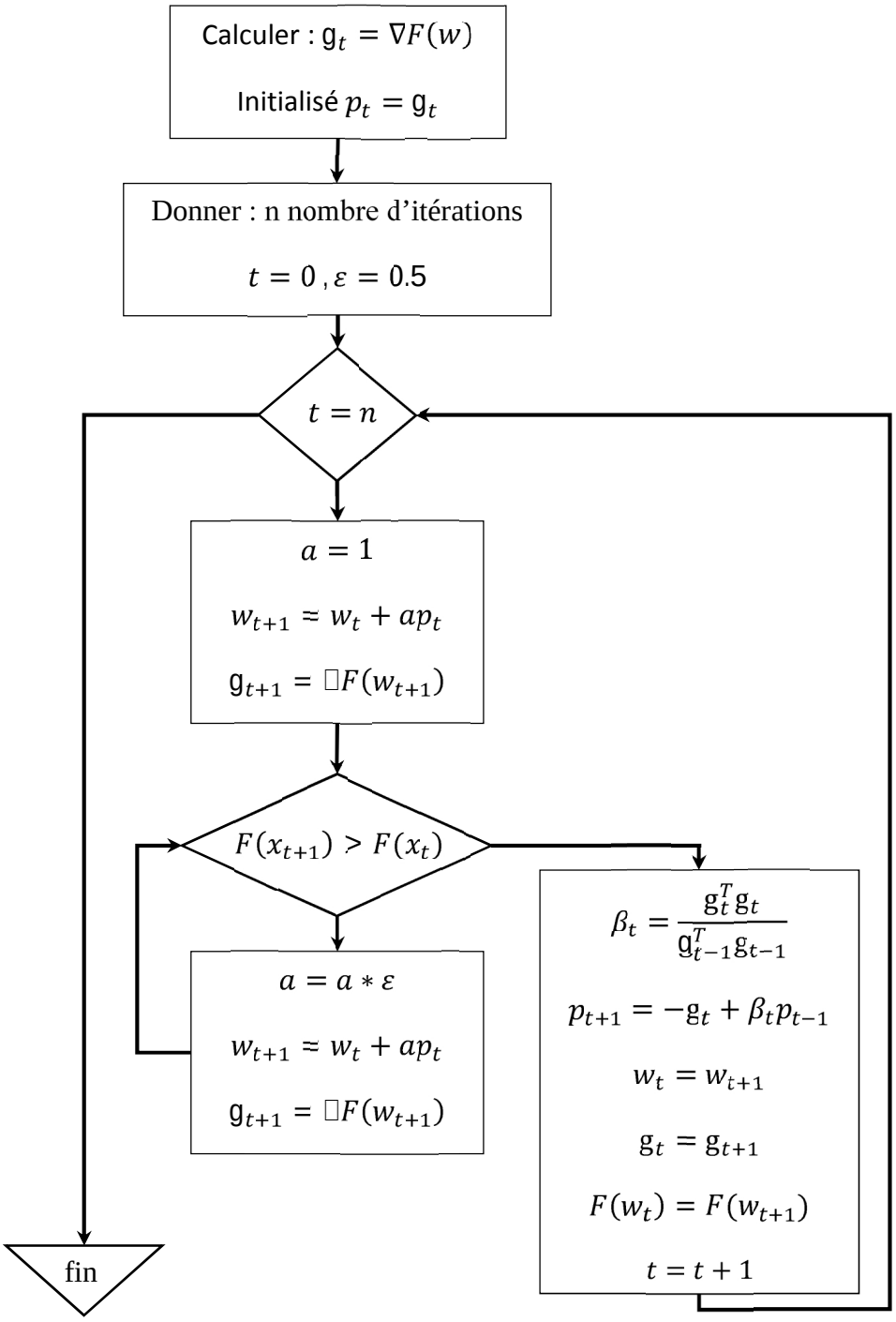

Figure 2.19. Organigramme de l'algorithme de Levenberg-Marquardt [97].

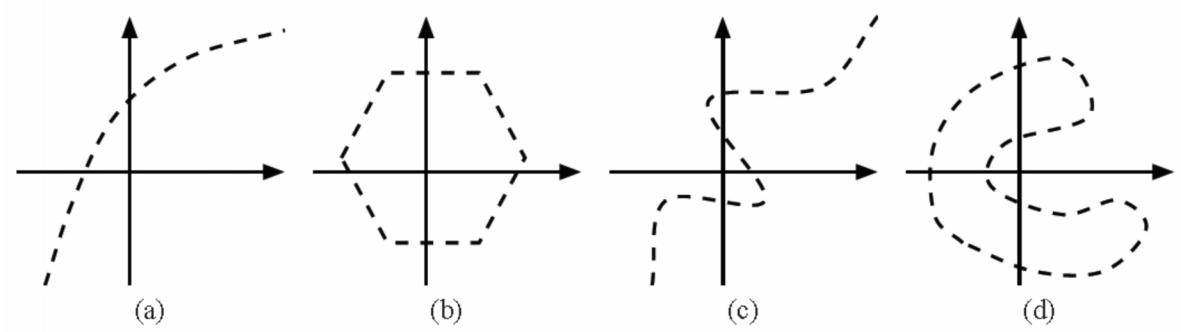

Figure 2.20. Exemples des fonctions approximés par un réseau MLP [97] : (a) convexe ouverte (b) convexe fermée (c) concave ouverte (d) concave fermée.

2.1.1.1.2. Classification

Pour la classification, on utilisera des réseaux à une couche cachée de sigmoïdes, cela suffira pour engendrer des frontières de décision convexe, ouvertes ou fermées, de complexité arbitraire, schématisée sur la figure 2.21. Avec deux couches cachées, le réseau MLP permet de créer des frontières de décision concaves ou convexes, ouvertes ou fermées.

La première couche cachée du réseau servira à découper l'espace d'entrée à l'aide de frontières de décision linéaires. La deuxième couche servira à assembler des frontières de décision non linéaires convexes. La couche de sortie permettra d'assembler des frontières de décision concaves.

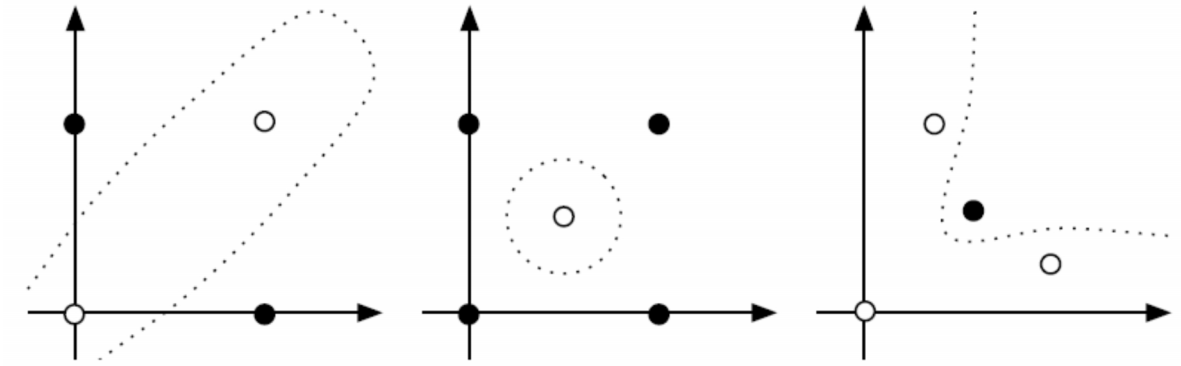

Figure 2.21. Exemples des frontières de décision

2.1.1.2. Procédure de développement d'un réseau MLP

Le cycle classique de développement d'un réseau MLP peut être décomposé en sept étapes [97] :

a. La collecte des données : L'objectif de cette étape est de recueillir des données suffisantes pour constituer une base représentative du problème ou de l'application posée.

b. L'analyse des données : Il est souvent préférable d'effectuer une analyse des données pour déterminer les degrés d'importance et de redondance des attributs. Cette analyse aura des conséquences sur la taille du réseau, ses performances et sur le temps de développement (temps d'apprentissage).

c. La séparation des bases de données : Il est nécessaire de disposer de deux bases de données : une base pour effectuer l'apprentissage et une autre pour tester le réseau obtenu et déterminer ses performances (validation).

d. Le choix d'architecture de réseau : Il existe un grand nombre d'outils permettant le choix convenable d'architecture du réseau dont chacun a ses avantages et ses inconvénients.

e. La mise en forme des données : De manière générale, les bases de données doivent subir un prétraitement afin d'être adaptées aux entrées du réseau.

f. L'apprentissage : Plusieurs types d'apprentissage peuvent être adoptés et le critère de choix, qui est le plus souvent utilisé, est la rapidité de convergence.

g. La validation : Après l'étape d'apprentissage, il faut tester son comportement en utilisant la base de validation. Il faut que ce réseau prédise par exemple l'évolution du système, pour des données hors de la base d'apprentissage.

2.1.1.3. Application du réseau MLP comme modèle de HAB

Les modèles de types boite noire ont prouvé leur utilité sur les processus théoriquement mal connus ou qui ont une forte non-linéarité tels que les processus biologiques de croissance cellulaires. Un réseau MLP peut approximer n'importe quelle fonction linéaire ou non linéaire. Cette propriété d'approximation trouve une grande application aux domaines de décision et de prédiction.

Une approche neuronale a été appliquée afin de modéliser le processus d'apparition des HAB et de réaliser un modèle de prédiction et d'alerte de bloom d'algues toxiques.

Comme nous avons vu à la section 2.2, la question critique de tels processus est qu'il est complexe, non linéaire et dynamique.

En suivant les étapes de développement d'un réseau MLP, nous avons effectué un apprentissage à partir d'une base de données, présenté dans la section 2.2.4. Un bon apprentissage d'un réseau de neurones bien structuré peut prédire l'évolution d'algues toxiques à partir de données nouvelles issues des différents capteurs.

2.2 Conclusion

Dans ce chapitre on n'a pas essayé de revoir un sujet déjà présenté dans d'autres travaux. Pour réaliser un système de surveillance permanent, on a présenté l'état d'art de la modélisation du processus de prolifération d'algues nocive dite HAB. Une proposition d'un système de prédiction de bloom spécifique pour l'espèce D. Acuminata a été élaborée. Ce système est conçu à partir de données facilement mesurables utilisant des sondes de moindre coût.

Des méthodes intelligentes sont souvent utilisées en biotechnologie pour la modélisation des processus biologiques. La difficulté dans ce cas de modélisation est la forte non-linéarité et le manque de connaissance à priori de ces processus biologiques.

La réalisation des modèles de prolifération algale spécifique aux espèces est très pratique par rapport aux modèles dits généraux (plusieurs espèces). D'autre part la précision de la décision et souvent erronée à cause de la ressemblance entre les conditions de prolifération.

Pour pouvoir lancer une alerte, il faut être sûr de la présence de l'espèce surveillée par analyse manuelle au laboratoire. Ces analyses sont laborieuses et prennent beaucoup de temps. Dans le chapitre suivant, et pour compléter notre travail de recherche, nous développerons un système automatisé pour la reconnaissance des phytoplanctons.

Chapitre 3

Reconnaissance et identification des espèces phytoplanctoniques

3.1. Introduction

La prédiction des HAB, présentée dans le chapitre précédent est une technique très avantageuse en termes d'utilité et de coût, mais sa précision n'est pas assurée. Dans le cas des systèmes de prédiction, pour avoir une décision définitive sur la présence d'algue nocive au moment d'alerte, un système d'identification doit être associé.

Le système d'identification des microorganismes trouve plusieurs applications dans les laboratoires et les systèmes de surveillance *in situ*. Le développement d'un tel outil nous conduit à une précision élevée que celle d'un système de prédiction.

Ce chapitre présente un système de reconnaissance des phytoplanctons. Nous discutons les instruments utilisés et les méthodes de traitement. Puis, nous détaillerons les deux méthodes de traitement que nous avons opté pour notre système de reconnaissance.

3.2. Instruments

Avant de parler d'identification d'algue, il faut tout d'abord la quantifier sous forme d'information. Pour les biologistes, la reconnaissance d'algues est concrétisée par l'étude d'images qui est l'information la plus déterminante. Il existe d'autres informations utiles liées à la structure interne qui sont extraites au laboratoire. Le microscope électronique et l'analyse taxonomique restent les moyens les plus sûrs pour l'identification. Les inconvénients de telles mesures sont le temps (collection des échantillons, traitement chimique, calibrage d'appareils) et l'exigence d'un personnel qualifié.

Un instrument qui pallie aux inconvénients cités est le cytomètre en flux. Il utilise les signatures optiques reflétées des cellules pour les caractériser. L'utilisation de cet instrument dans les études marines a considérablement augmenté au cours des dernières années.

La cytométrie en flux (CF) est une technique qui mesure la diffusion de la lumière et les caractéristiques de fluorescence de chaque cellule ou particule [99]. C'est une technique capable de mesurer directement les propriétés des microorganismes et fournir des informations qualitatives et quantitatives sur un phytoplancton. Elle est devenue rapidement un puissant outil de caractérisation des communautés phytoplanctonique dans les milieux aquatiques. En utilisant des différentes techniques de classification, il est maintenant possible de discriminer les principaux groupes de phytoplancton.

Les signaux optiques détectables peuvent refléter naturellement à partir d'algues, ce qui permette de savoir la taille de cellule [100], la structure [101] et les pigments endogènes [102] ou peuvent être générés par des tâches fluorescentes [103].

La CF a vu le jour, pour combler l'écart entre l'analyse microscopique traditionnelle laborieuse qui prend beaucoup de temps et les besoins d'une grande fréquence d'acquisition des paramètres physiologique des cellules [104,105].

3.2.1. La cytométrie en flux

La cytométrie en flux est une technique développée pour l'analyse individuelle des particules (cellules) en suspension dans un milieu liquide. Les particules en suspension doivent être séparées les unes des autres pour être analysées individuellement.

Chaque particule passe devant un faisceau laser, d'où il en résulte une diffusion de lumière accompagnée éventuellement d'une émission de fluorescence suite à l'excitation d'un pigment photosynthétique ou d'un marqueur fluorescent.

Le principe de la CF est général, mais chaque instrument fabriqué est caractérisé par ses paramètres et ses équipements. Il est recommandé de détailler le type utilisé directement sans passer par la description générale de la méthode.

Les cytomètres utilisés dans cette étude des phytoplanctons sont : le CytoSense Benchtop (CytoBuoy BV, Pays-Bas) et FLOWCAM.

Les instruments utilisés sont des cytomètres en flux dédiés à l'analyse des cellules phytoplanctoniques [106, 107]. Ces instruments effectuent des analyses automatisées et ils sont facilement transportables par rapport aux instruments traditionnels. L'échantillon est alors prélevé en milieu naturel sans subir des procédés de conservation et l'automatisation permet de réaliser des analyses à haute fréquence.

3.2.2. Le cytomètre en flux CytoSense

Le CytoSense se compose de trois sous-systèmes : fluidique, optique et électronique qui sont mentionnées sur la figure 3.1.

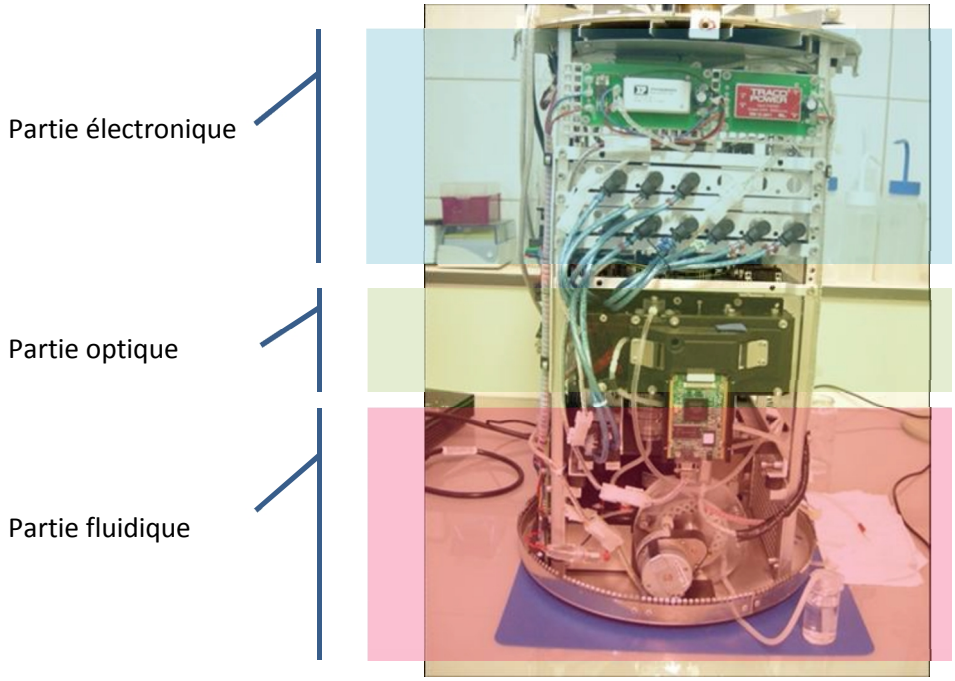

Figure 3.1. Les différentes parties de CytoSense [108].

- **Le système fluidique :**

L'une des principales particularités du CytoSense par rapport aux cytomètres en flux classiques, réside dans la taille des particules qui peuvent être prélevées. Un tube de diamètre intérieur de 800µm est relié à une pompe péristaltique qui amène l'échantillon dans une chambre appelée « injecteur » visualisé sur la figure 3.2.

Le débit de prélèvement est contrôlable et varie entre 0,49 et 9,7 µL/s. Ce débit doit être préalablement calibré selon le type de particules à analyser pour obtenir un comptage le plus précis possible [106].

Dans l'injecteur, un liquide gaine est injecté à une vitesse supérieure à celle de l'échantillon (2 m/s), garantissant la création d'un flux laminaire et une séparation des deux types de fluides. L'étirement de l'échantillon induit une séparation des particules qui circulent en file indienne au travers d'une cuve en quartz de 1 mm de diamètre où a eu lieu l'analyse optique. Cette technique est appelée la focalisation hydrodynamique résumée dans la figure 3.3. Ainsi toute particule pouvant entrer dans le tube de 800 µm peut passer à travers le système. La nature du liquide gaine dépend de celle de l'échantillon à analyser pour ne pas modifier l'indice de réfraction.

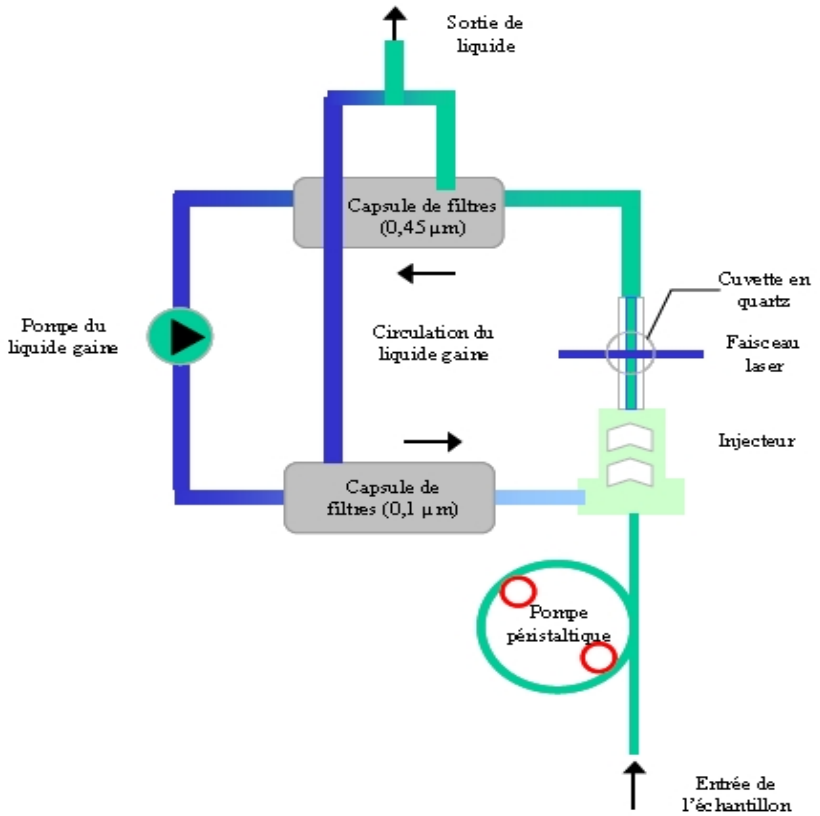

Figure 3.2. Schéma de système fluidique de CytoSense.

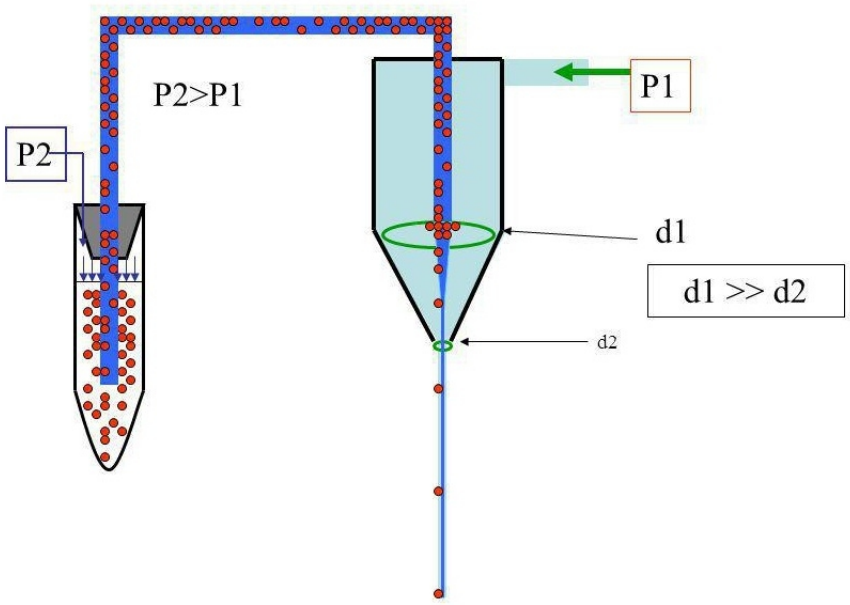

Figure 3.3. Le principe de focalisation hydrodynamique [109].

- **Le système optique :**

Dans une chambre d'analyse en quartz, les particules pénètrent un faisceau laser bleu (Coherent solid-state sapphire, 488 nm, 20mW) aplati de longueur 300 µm et de largeur 5 µm. Les particules alors diffusent la lumière et dans le cas de phytoplancton, émettent de la

fluorescence dont l'intensité et la longueur d'onde d'émission dépendront du type de pigments photosynthétiques. Tout le système est présenté à la figure 3.4.

Tout d'abord, dans la même direction que celle du faisceau laser, la diffusion vers l'avant FWS (Forward Scatter) est détectée par une photodiode PIN (positive intrinsic negative diode).

La diffusion à 90° SWS (Sideward Scatter), et la fluorescence sont collectées orthogonalement à l'aide de photomultiplicateurs (PMT). La photodiode est beaucoup moins sensible que le photomultiplicateur, mais plus rapide en temps de réaction. Une autre partie de la lumière diffusée est collectée et amplifiée par un miroir placé de l'autre côté de la cuvette. Un réseau holographique concave permet de séparer les différentes longueurs d'onde (avec un pas de 33 nm) et le reste de la lumière est collecté par une fibre optique représentant le SWS. Les différents secteurs de lumière sont amplifiés dans un photomultiplicateur.

Cinq propriétés optiques sont alors détectées pour chaque particule :

- ✓ La diffusion vers l'avant ou aux petits angles FWS (0-15°) permet de détecter la taille de la particule.
- ✓ La diffusion à 90° ou aux grands angles SWS (45-135°) révèle la structure interne et externe des particules (forme et constituants cellulaires).
- ✓ La fluorescence rouge FLR (668-734 nm) est la principale bande d'émission de la chlorophylle a.
- ✓ La fluorescence orange FLO (601-668 nm) permet de distinguer les particules contenant des pigments accessoires tels que les phycobilines (phycocyanines) contenus dans les cyanobactéries et les cryptophytes.
- ✓ La fluorescence jaune/verte FLY (536-601 nm) permet de distinguer les particules contenant des phycoérythrines également contenues dans les cyanobactéries et les cryptophytes. Ce type de fluorescence est utile pour le marquage artificiel des composants cellulaires (ADN).

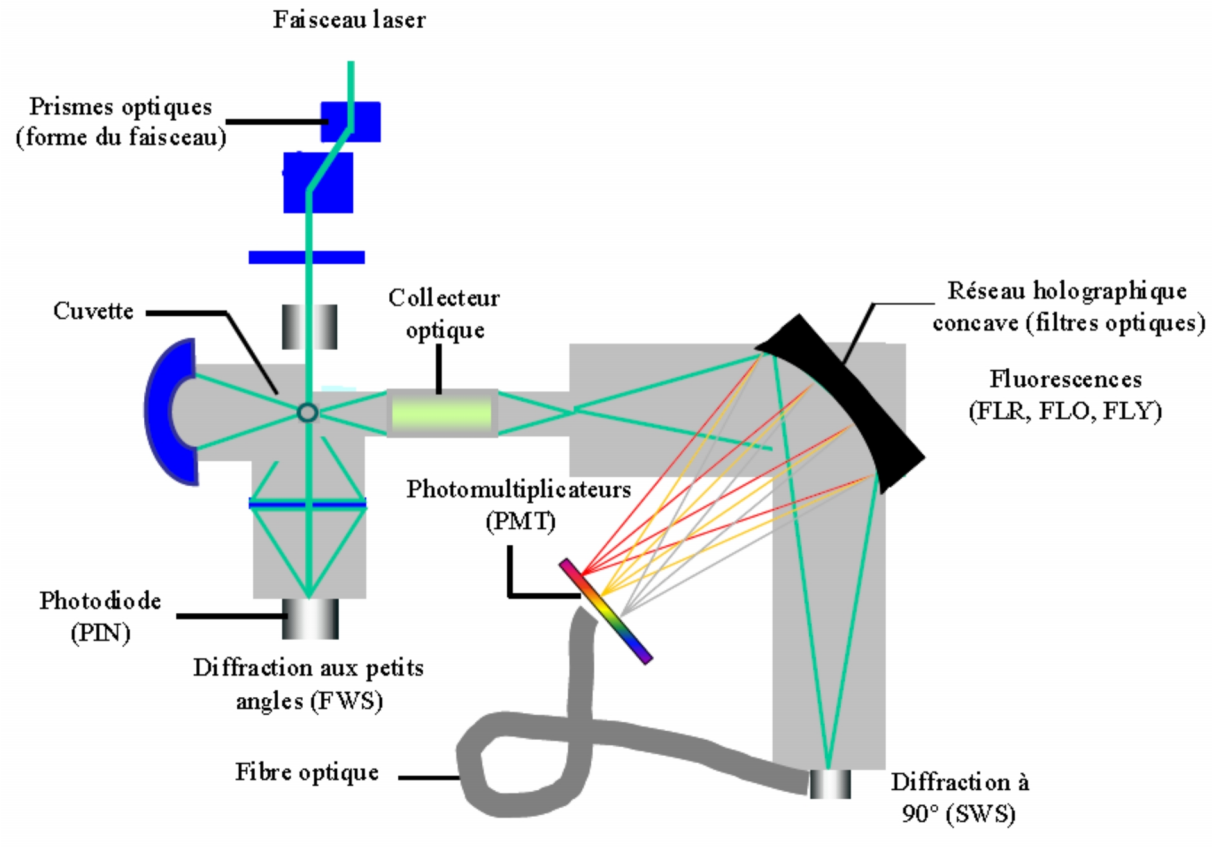

Figure 3.4. Le système optique de CytoSense.

L'instrument est également équipé de deux modules supplémentaires :

- ✓ Un module de détection de la courbure des particules (Curvature) consiste en un polariseur disposé sur l'unité FWS et permet de détecter les cellules à forme spiralée, courbée et les colonies.
- ✓ Un set de petits détecteurs de faible sensibilité (Low sensitivity, LS) est prévu pour la détection des grosses particules et un set de détecteurs plus gros est utilisé à haute sensibilité (High Sensitivity, HS) pour la détection de particules plus petites (picoplancton). Ce module de détection (HS, LS) du picoplancton ne comprend que les détecteurs de SWS et de fluorescence FLR et FLO.

➢ **Le système électronique**

Chaque variable collectée est transmise vers un disque dur associé placé au-dessus de la chambre optique, figure 3.1. Cette interface électronique convertit les signaux électriques bruts en données numériques. Chaque disque dur peut contenir 64 kb de données (équivalent à l'enregistrement d'une particule de 30 mm de long).

La largeur du faisceau laser à 50% du maximum de son intensité est de 3 µm, on peut considérer que les particules plus larges que 6 µm auront des informations valides sur leur organisation interne et leur forme. Contrairement aux cytomètres en flux conventionnels, le pulse obtenu n'est pas intégré, mais décomposé en plusieurs données à une vitesse de 4 MHz (4 données par µs, soit 1 donnée par 0,5 µm) lorsque la particule passe à travers le faisceau laser à une vitesse de 2 $m.s^{-1}$.

Sur la figure 3.4, le flux de photons émis par le passage d'une particule devant le laser est converti en impulsions électriques par les différents capteurs optiques, à leur tour converties en signaux numériques. Le principe est illustré dans la figure 3.5. Les signaux générés, représentés dans la figure 3.6, sont stockés sur des disques durs.

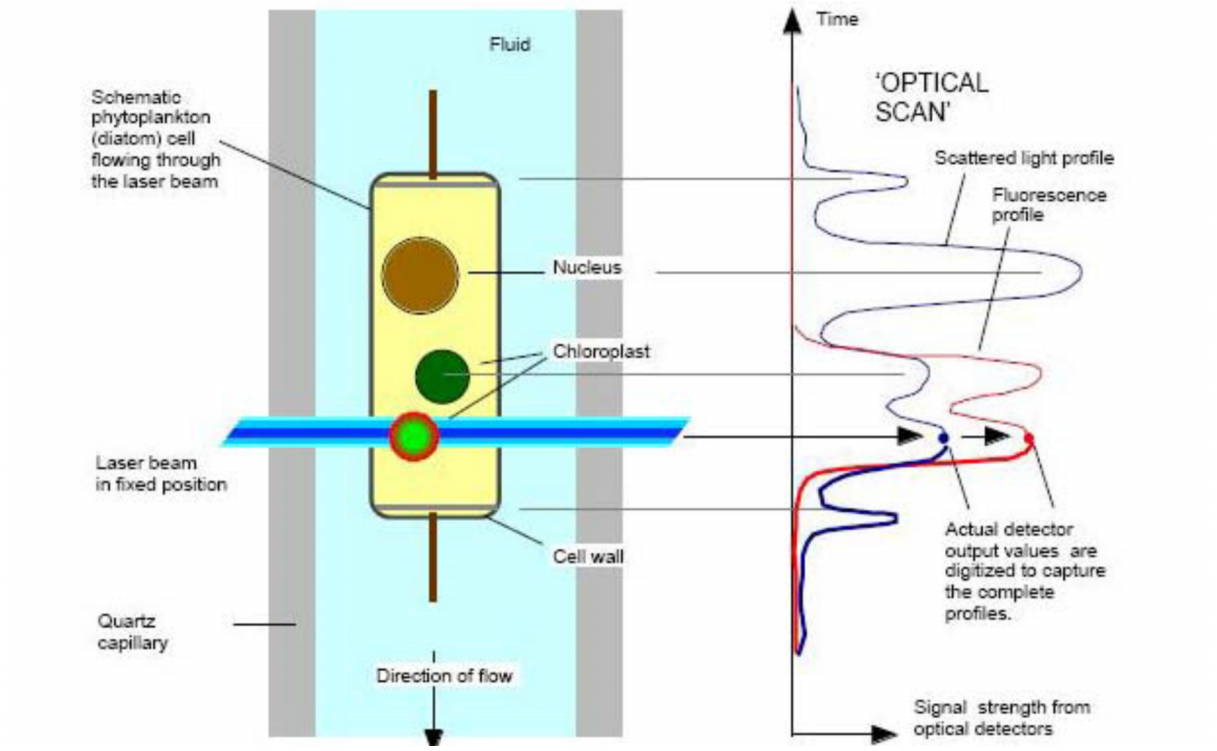

Figure 3.5. Principe de conversion optique en signaux électriques de de CytoSense.

Le seuil de déclenchement de l'enregistrement des signaux est fixé par l'opérateur en fonction du type de particule recherché. Il s'agit d'éviter de saturer la mémoire avec du bruit électronique et des particules parasites inertes. Pour cela, le seuil requis est basé sur des signaux qui ont le meilleur rapport signal/bruit. De même, comme la taille de particule est proportionnelle à son temps de passage devant le laser, il convient de choisir une vitesse

appropriée à la taille des particules à étudier. Par exemple, l'opérateur choisira une vitesse faible pour l'énumération des particules de petite taille. Une vitesse faible permet de bien séparer les particules au moment de la création du flux laminaire, limitant ainsi les phénomènes de coïncidence rencontrés lorsque les concentrations cellulaires sont élevées.

Le contrôle des paramètres de vitesse et de seuil de déclenchement se fait au moyen de logiciel CytoUSB installé dans un ordinateur externe, relié au CytoSense par port USB. Plusieurs modes d'opération de l'instrument sont possibles via le CytoUSB : automatique, interactif ou programmé. Pour plus d'information sur les modes et les paramètres de cette appareille consulter le site web www.cytobuoy.com.

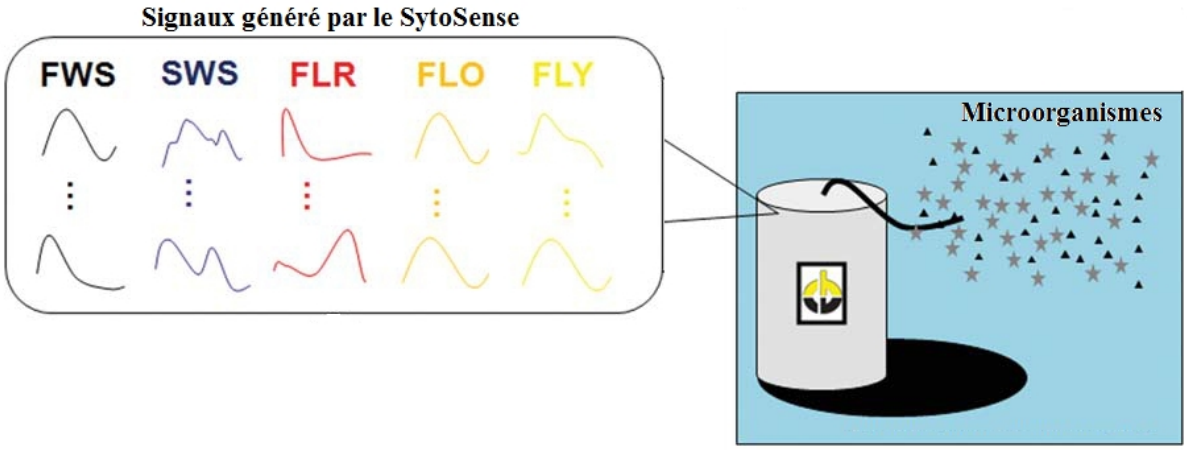

Figure 3.6. Les signaux générés par CytoSense.

Le logiciel Cytoclus (fourni avec l'appareil) réalise une série de calculs mathématiques sur les profils optiques des particules et permet de représenter l'ensemble des particules sur des graphiques appelés cytogrammes. Ils s'agissent de graphiques à nuage de points dont les axes représentent deux paramètres élémentaires calculés sur les profils optiques. Les formes des profils contiennent en elles-mêmes beaucoup d'informations et il devient alors difficile d'analyser une grande quantité de particules. Pour cette raison, un nombre de paramètres était défini pour capturer les caractéristiques les plus importantes des formes de pulses de particule.

Le Cytoclus permet ainsi de sélectionner rapidement une série de particules avec des caractéristiques similaires. Les attributs calculés par le CytoClus sont les suivants:

- **Longueur (Length) :** La longueur est déterminée à partir du temps de passage de la particule à travers du laser, corrigé d'un seuil de 13% (déterminé à partir de considérations physiques du laser).
- **Total (Total) :** Le total est l'intégration du pulse, représentant l'aire du signal.
- **Maximum (Max) :** Le maximum correspond à la valeur maximale détectée le long de la particule.
- **Moyenne (Avg) :** La moyenne est égale à la somme des données divisée par le nombre de données. La moyenne élimine l'effet de taille et de la fluorescence totale.
- **Inertie (Inertia) :** L'inertie est définie comme le second moment de la forme du pulse.
- **Centre de gravité (CG) :** Le centre de gravité est calculé en divisant le premier moment de la forme du pulse par le total.
- **Facteur de remplissage (Fill) :** Le facteur de remplissage donne une indication sur la solidité de la forme du pulse.
- **Asymétrie (Asymm):** L'asymétrie donne une indication sur la distribution du signal le long de la particule.
- **Nombre de cellules (#Cells) :** Le nombre de cellules est calculé à partir d'un algorithme qui identifie le nombre de pics le long de la particule.
- **TOF :** (Time Of Flight) correspond au temps de passage de la particule devant le laser.

Après calculs, les données sont stockées dans un fichier d'extension fcm. Un exemple est représenté au tableau 3.1. Ces attributs sont calculés pour chaque signal, à l'exception du TOF qui est identique pour tous les signaux.

Pour que Cytoclus fasse la discrimination optimale des groupes, il faut représenter une multitude de combinaisons graphiques, ce qui rend difficile leur interprétation.

Tableau 3.1. Un exemple de fichier d'extension fcm.

TOF FWS	Length FWS	Length SWS	Length FL Re	Length FL Or	Length SWS	Length FL Re	Length FL Or	Total FWS	Total SWS HS	Total FL Red	Total FL Orar	Total SWS LS	Total FL Red	Total FL Orar	Max FWS	Max SWS HS	Max FL Red H
135	91,20644	91,27193	85,49666	86,32919	89,26673	83,32904	86,39359	648587,8	734404,9	217308	2771,844	29913,79	3593,555	1499,985	4136,784	4173,864	2063,967
181	106,23	118,5159	16,12242	58,79391	89,73131	17,75036	56,24363	807612,8	842064	288669,3	3555,559	46697,59	5225,074	2047,64	4099,857	4136,784	3301,035
108	99,47431	96,76285	62,65062	80,12054	96,84839	61,00801	72,37006	502105,5	345826,6	72253,64	1342,357	10680,26	1430,708	919,0649	2780,542	2657,857	673,6961
127,5	92,10058	116,2573	26,58497	77,11072	77,43169	76,14236	56,74368	637531,4	696886,6	183398,9	2355,114	23554,8	3093,051	1493,118	4136,784	4173,864	2856,99
194	178,3452	185,173	8,411015	148,6642	88,04398	7,689813	149,0851	1155347	930746,4	250432,3	3174,535	27053,44	4531,236	1922,056	4173,864	4173,864	3098,697
173	164,365	161,9397	161,533	162,7061	162,9609	151,9977	155,0727	857017,4	475368,7	111949,1	2175,817	14786,98	2168,798	1463,973	2805,414	1754,204	399,0295
124	115,1856	114,491	112,4349	111,6093	115,0671	111,6968	113,6969	630641,3	349547	134113,4	1798,456	10966,37	2242,958	976,5923	3015,534	1990,722	984,222
216	200,3779	200,9835	177,5167	186,8409	199,5005	159,6286	195,766	1208568	1075770	300806,2	3981,445	37496,57	5066,226	2244,789	4173,864	4173,864	1802,423
194	179,0048	80,35187	63,17048	64,58894	64,4826	59,68296	67,06131	992865,3	803781,9	126766,4	2348,399	27816,86	2260,811	1501,816	4136,784	4173,864	1179,083
114	84,51261	86,00964	82,2415	78,48177	27,31533	82,35699	83,08493	526712,8	549219,5	105446	1890,164	19747,46	1870,937	959,1968	4173,864	4173,864	1222,419
231,5	220,6191	212,5253	210,5973	213,6024	219,2852	156,2311	211,5983	1123845	1055174	158760,5	2827,54	32876,83	2654,347	1475,417	2908,567	4099,857	785,3939
118	108,2329	100,2459	46,75099	96,23664	103,0958	54,91706	55,80608	609756,2	708399,2	89284,8	1690,42	26544,24	1383,099	835,902	4099,857	4173,864	1244,85
308,5	298,7538	279,3576	212,8766	299,3238	84,11684	212,1136	238,3735	1644977	1650250	287558,4	5274,057	67153,7	5164,953	2969,909	4173,864	4173,864	1399,884
128,5	117,1219	118,4298	90,98875	119,2533	118,6706	102,0094	101,8297	622870,9	355279,4	125641,8	1751,915	9846,188	2399,06	1181,829	2908,567	2178,717	984,222
475,5	233,8417	205,3388	168,3502	118,0917	120,7224	111,8352	125,9806	1814931	1756119	1471049	20470,14	271400,5	29717,25	12889,95	4173,864	4173,864	4327,372
152	140,4747	105,2903	66,07124	129,517	101,4994	131,3901	97,71069	723272,4	764345,6	109892	2173,071	21350,14	2161,168	1176,183	4025,849	4099,857	1057,924
135	125,1485	122,3242	91,14793	2,692144	115,8016	67,16581	3,369521	600516,6	703207,8	77005,07	3256,173	20583,97	1485,641	1238,136	3015,534	4099,857	655,5376
119	110,426	105,814	87,74129	85,94237	107,0794	85,92918	106,7823	503682,3	497333,2	129467,5	1977,599	14562,82	2220,069	1132,542	2856,99	4025,849	1211,585

En général, quelques attributs permettent d'interpréter facilement et de représenter l'essentiel de l'information contenue dans le nuage de points pour l'extraction de quelques groupes. Il s'agit de la longueur, du maximum, du total du signal ainsi que du nombre de cellules.

3.2.3. FLOWCAM

Le FLOWCAM utilise une combinaison d'un microscope et d'un laser permettant de donner plus d'informations et ainsi fournir la possibilité d'une meilleure discrimination pour la distinction des cellules dans un groupe donné. Le principe de fonctionnement est schématisé dans la figure 3.7.

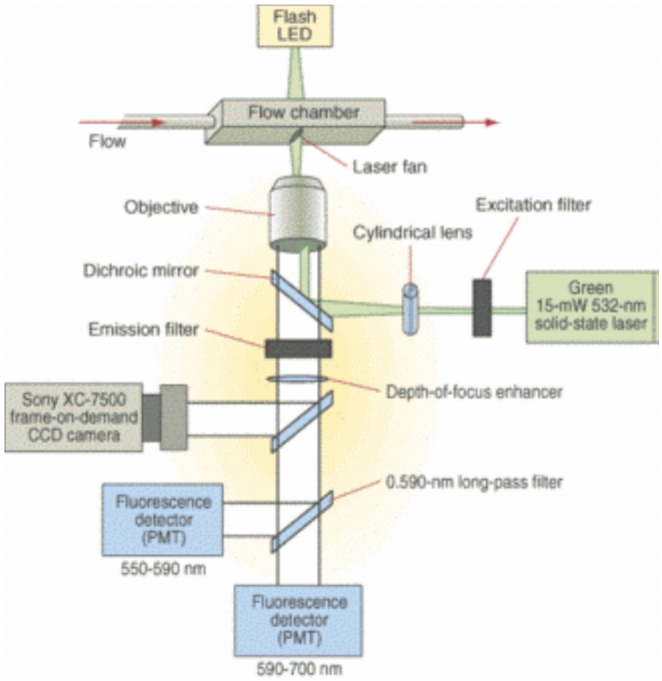

Figure 3.7. La combinaison de microscope et de laser du FLOWCAM [110].

Comme tous les cytomètres en flux, les données cytométriques sont visualisées sous forme de nuages de points qui regroupent des organismes qui sont caractérisés par leur taille et leur fluorescence. De par la conception de l'appareil, des espèces distinctes se confondent dans ces groupements. Sur la figure 3.8, un exemple de nuages de points contenant trois couleurs, chaque couleur étant associée à un ensemble de caractéristiques distinctes.

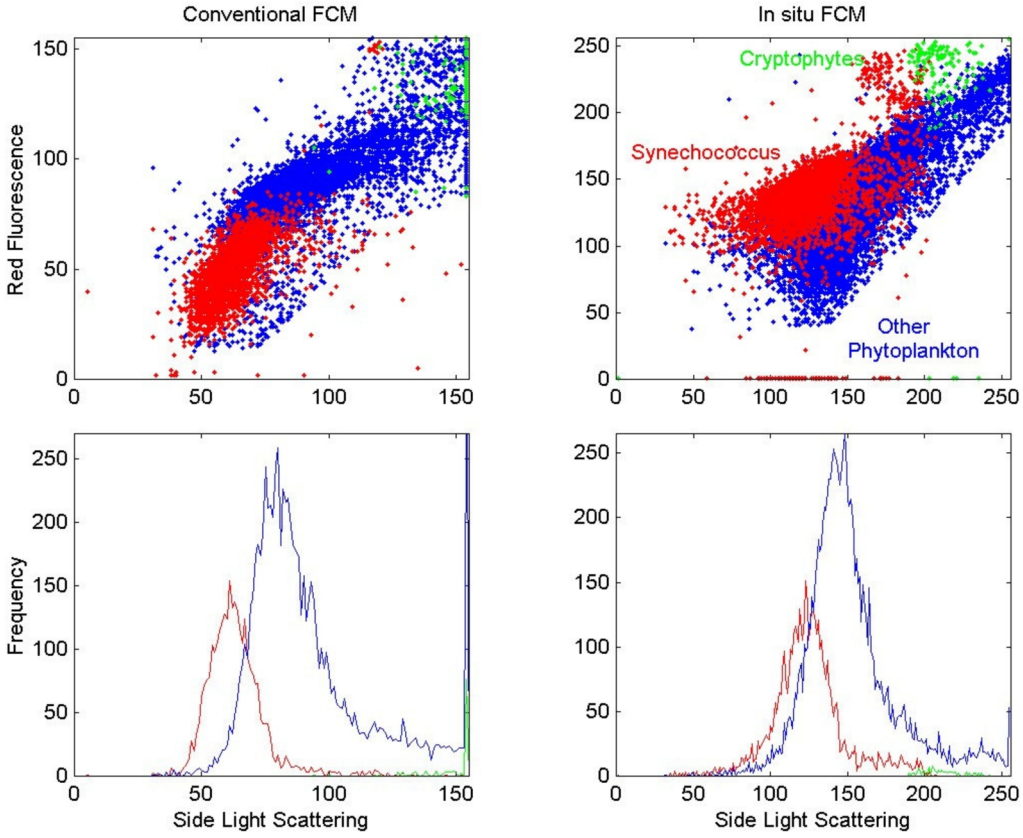

Figure 3.8. Représentation par nuages de points d'organismes multiples [111].

La fusion de l'image et des paramètres cytométriques fournis par le FLOWCAM est une technique viable qui permet la détection rapide d'algue à partir de ces données.

Dans la figure 3.9, des images fournies par le FLOWCAM sont présentées. Le microscope ayant servi à produire ces images avait un grossissement compris entre 20x et 40x. Le cytomètre en flux associé est équipé d'un laser vert avec un faisceau de largeur de 5 µm (laser à état solide, 532 nm, 15 mW) et deux photomultiplicateurs PMT1 (550-590 nm) et PMT2 (590-700 nm).

Le FLOWCAM produit des paramètres caractéristiques pour chaque cellule passant devant ses capteurs. En plus de l'image, on pourrait disposer d'une combinaison de signaux provenant du cytomètre. Pour l'instant, l'appareil ne donne que la valeur maximale et le temps de passage de la cellule (largeur de l'impulsion). La forme complète des signaux, si elle était disponible, serait une information très pertinente pour la reconnaissance.

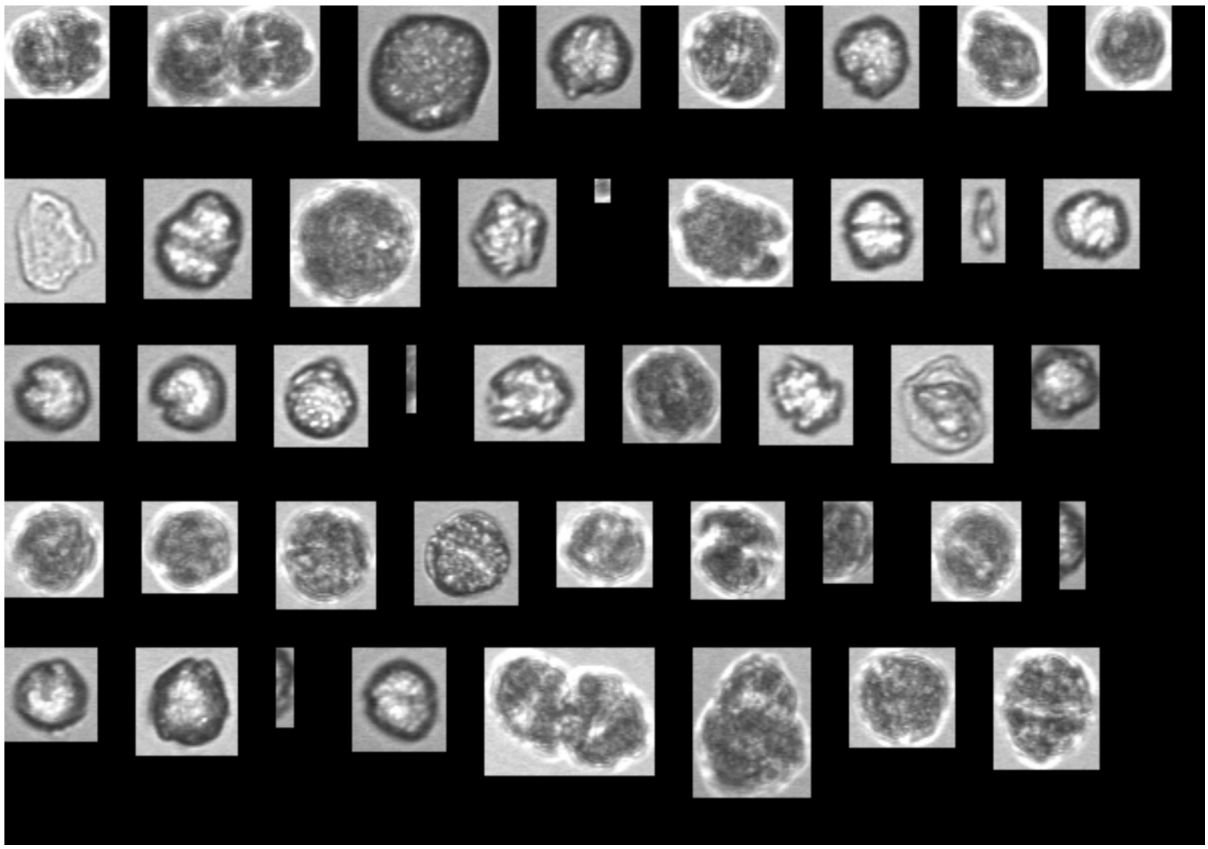

Figure 3.9. Exemples d'images de FLOWCAM (Alexandrium Tamarense).

Sur la figure 3.10, on peut voir un exemple d'une telle forme d'onde. Celle-ci est composée de trois signaux provenant de trois capteurs détectant les intensités de fluorescence de trois longueurs d'onde spécifiques. Ces signaux serviront, à mieux caractériser les cellules à reconnaître, par le choix et l'extraction de paramètres variés en fonction de la forme des signaux. Dans la figure 3.10, les signaux sont associés à l'espèce Alexandrium tamarense.

Pour chaque échantillon analysé, le logiciel intégré du FLOWCAM génère deux fichiers avec des images d'extension tif et des fichiers contenant des paramètres avec l'extension fcm. Les informations spécifiques pour chacune des cellules sont composées des paramètres de l'image et du cytomètre. Ces paramètres indiquent : le nombre de cellules apparaissant dans l'image, la longueur et la largeur de la cellule, le centre de gravité, la fluorescence. Cette fluorescence est engendrée par l'excitation du pigment de la chlorophylle et de la phycoérythrine par un ou plusieurs lasers. Ces paramètres pouvant caractériser une cellule par rapport à d'autres.

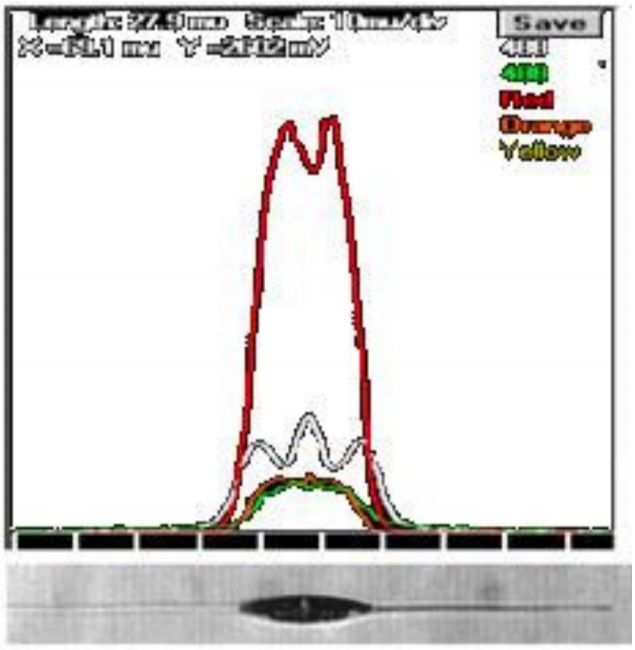

Figure 3.10. Signaux cytométrique provenant du FLOWCAM (Alexandrium tamarense).

La qualité des images produites par le FLOWCAM joue énormément sur le calcul des paramètres. Si l'image est très bruitée alors les valeurs des paramètres seront faussées. Les principaux paramètres enregistrés dans le fichier généré par le FLOWCAM apparaissent dans le tableau 3.2. Les paramètres «Cell Area», «FFT Area», «Feret_max_diameter», «Feret_min_diameter», «Cellx», «Celly» et «Cell ESD» sont liés à la forme de la cellule. Ces paramètres de formes sont le résultat d'un traitement d'image effectué par le FLOWCAM.

Le paramètre «Chlorophyl peak» correspond à la valeur maximale de fluorescence, de la présence de pigment de chlorophylle, enregistrée lors du passage de la cellule devant les capteurs. Le paramètre «Chlorophyl TOF» désigne le temps de passage en milliseconde du signal de fluorescence de la chlorophylle. Le paramètre «Phyco peak» est la valeur maximale de fluorescence obtenue lors de l'excitation par un laser de longueur d'onde spécifique du pigment de phycoérythrine contenu dans une cellule tandis que le paramètre « Phyco TOF » est le temps de passage en milliseconde du signal de fluorescence de la phycoérythrine. Ces paramètres proviennent de la partie cytométrique du FLOWCAM.

Le FLOWCAM est contrôlé par le logiciel FlowCAMphyco fournis par le producteur pour l'acquisition des données et l'analyse des images et des données cytométriques. Pour plus de documentation, le site de producteur est : http://www.fluidimaging.com.

La figure 3.11, présente les diagrammes de dispersion bidimensionnelle (sous forme de nuage de points) du logiciel associé au Flowcam.

La sélection d'un ensemble de données cytométriques permet de voir les images des cellules associées à deux ou trois paramètres au maximum. Cela permettra aux scientifiques de visualiser les phytoplanctons qui ont les mêmes caractéristiques présentées sur la courbe.

Le tableau 3.3 présente l'ensemble des données générées par le FLOWCAM. Elles sont collectées et enregistrées dans un fichier de type fcm.

Tableau 3.2. Principaux paramètres du fichier fcm du FLOWCAM.

Cell area
Chlorophyll peak
Chlorophyll TOF
Phyco peak
Phyco TOF
Particles per chain
FFT Area
Cellx
Celly
Cell ESD
Feret_min_diameter
Feret_max_diameter

Dans l'exemple de la figure 3.11, 1330 cellules sont représentées sous forme de nuage de points dans deux graphes qui sont: l'intensité logarithmique de la chlorophylle et de la phycoérythrine en fonction du diamètre de la cellule.

Le problème dans cette méthode de discrimination est la dimension énorme des données ignorées dans la sélection et la possibilité presque sûre d'avoir des cellules d'espèces différentes avec les mêmes paramètres.

Tableau 3.3. Partie des données générées par le FLOWCAM.

Time of Event	File Number	Cell Area	Chlorophyll	Chlorophyll TOF	Particles per ch	FFT Area	Phyco Peak	Phyco TOF	feret_max_(	feret_min_c	Cell X	Cell Y	Cell ESD	Pat. rec. sco	FlowCam Vo	Manual Vote	Stuck Particl	histogram b	Bulk Fluor (uVolts)		
170057817	170135235	25	2289	6430	1	2	2264	7970	8	6	103	234	5	0	0	-1	0	0	0	0	2289
170057817	170135235	26	2362	6586	1	11	2238	7488	8	5	342	373	5	0	0	-1	0	0	0	0	2362
170057817	170135235	24	2385	5457	1	25	2376	7319	6	4	179	16	5	0	0	-1	0	0	0	0	2385
170057817	170135235	47	2457	6979	1	4	2373	7280	9	8	492	89	7	0	0	-1	0	0	0	0	2457
170057817	170135235	73	2246	6923	1	2	2240	7113	13	9	462	405	9	0	0	-1	0	0	0	0	2246
170057817	170135235	26	2264	7256	1	3	2244	7440	9	5	442	254	5	0	0	-1	0	0	0	0	2264
170057817	170135235	55	2355	7662	1	12	2342	7707	13	10	364	396	8	0	0	-1	0	0	0	0	2355
170057817	170135235	38	2356	2659	1	3	2353	2658	10	7	187	256	6	0	0	-1	0	0	0	0	2356
170057817	170135235	19	2367	1986	1	18	2386	2035	7	4	473	191	4	0	0	-1	0	0	0	0	2367
170057817	170135235	26	2350	7847	1	3	2316	7866	9	5	591	338	5	0	0	-1	0	7	0	0	2350
170057817	170135235	78	2285	6900	1	4	2272	7311	18	12	288	32	9	0	0	-1	0	42	0	0	2285
170057817	170135235	49	2363	6817	1	7	2254	7422	10	7	617	186	7	0	0	-1	0	47	0	0	2363
170057817	170135235	74	2495	12966	3	8	2389	7263	27	4	265	17	9	0	0	-1	0	42	0	0	2495
170057817	170135235	33	2352	6802	1	4	2381	1168	8	5	261	221	6	0	0	-1	0	24	0	0	2352
170057817	170135235	44	2352	6802	1	4	2381	1168	10	7	209	283	7	0	0	-1	0	40	0	0	2352
170057817	170135235	39	2535	852	1	6	2341	1230	9	7	282	105	7	0	0	-1	0	17	0	0	2535
170057817	170135235	27	2535	852	1	6	2341	1230	7	6	143	356	5	0	0	-1	0	4	0	0	2535
170057817	170135235	46	2584	6686	2	27	2265	7846	19	7	7	42	7	0	0	-1	0	3	0	0	2584
170057817	170135235	33	2439	7444	1	12	2385	8166	8	5	136	300	6	0	0	-1	0	4	0	0	2439
170057817	170135235	43	2433	1292	1	3	2295	7562	12	6	322	262	7	0	0	-1	0	7	0	0	2433
170057817	170135235	42	2433	1292	1	3	2295	7562	9	6	373	331	7	0	0	-1	0	1	0	0	2433
170057817	170135235	27	2424	7286	1	5	2273	7746	8	5	485	52	5	0	0	-1	0	0	0	0	2424
170057817	170135235	30	2491	4984	1	5	2357	7914	9	5	402	10	6	0	0	-1	0	0	0	0	2491
170057817	170135235	38	2420	7166	1	8	2386	8193	11	7	519	247	6	0	0	-1	0	0	0	0	2420
170057817	170135235	79	2537	6554	2	8	2294	431	21	6	399	386	10	0	0	-1	0	0	0	0	2537
170057817	170135235	38	2366	7520	1	11	2358	7766	9	5	218	312	6	0	0	-1	0	0	0	0	2366
170057817	170135235	26	2356	8093	1	2	2396	2162	13	5	189	43	5	0	0	-1	0	0	0	0	2356
170057817	170135235	42	2353	7040	2	5	2292	1164	23	5	400	371	7	0	0	-1	0	0	0	0	2353

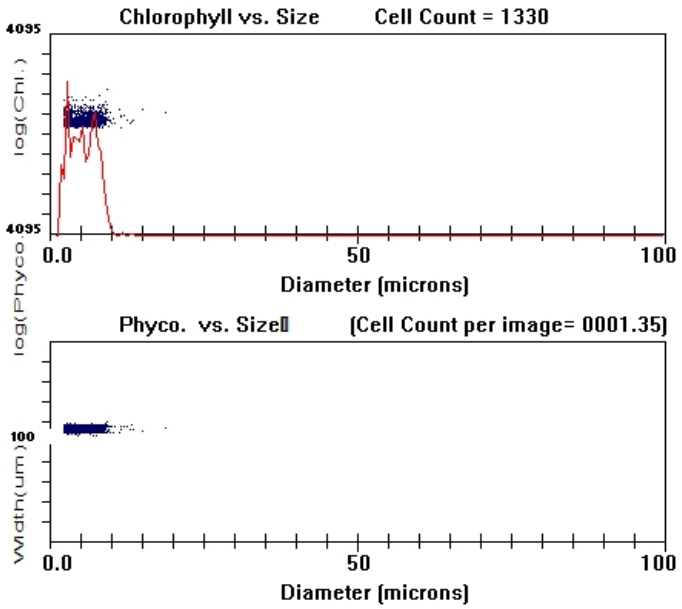

Figure 3.11. Présentation sous forme de nuage de points du logiciel associé au FLOWCAM.

3.3. Traitements automatiques des données cytométrique

Le traitement des données par Cytoclus et FLOWCAM ne suffissent pas pour identifier toutes les espèces ; donc il faut s'orienter vers d'autres traitements plus puissants et intelligents.

3.3.1. État d'art

Les logiciels associés aux instruments de CF ont beaucoup de difficultés pour distinguer les cellules de phytoplancton. Ils sont basés sur la gestion de certaines variables mesurées et des diagrammes de dispersion bidimensionnels. Cette technique est limitée et exige l'expertise d'un opérateur pour l'identification.

Il est montré que les classes formées par des nuages de points en deux dimensions pourraient être les mêmes pour différentes espèces [112]. Ainsi, le potentiel de l'analyse de données CF ne peut être atteint que par des méthodes de classifications multidimensionnelles qui fonctionnent avec la même vitesse que l'instrument.

De nos jours, en utilisant les données CF, de nombreuses études ont été menées pour faire la distinction entre les différents groupes de phytoplancton. On peut citer: l'analyse en composante principale [113], les SVM [114], les RNA [110, 115-117], la classification floue

[118], le regroupement séquentiel super paramagnétique (SSC) [119] et le modèle des mélanges gaussiens [120-123].

Dans cette étude trois méthodes sont utilisées afin de déterminer un bon classificateur. Nous avons appliqué deux méthodes: le modèle des mélanges gaussiens et la méthode d'arbre de décision pour la classification des données de CytoSense. Nous avons appliqué aussi deux méthodes: le réseau de neurones MLP et la méthode d'arbre de décision pour la classification des données de FLOWCAM. Dans ce qui suit, nous allons détailler les deux méthodes : le modèle des mélanges gaussiens et la méthode d'arbre de décision. Le réseau de neurone MLP été introduit dans la section 2.3.3 du chapitre 2.

3.3.2. Modèles des mélanges gaussiens

Le mélange de Gaussiennes GMM (Gaussian mixture modele) est un outil très utilisé dans l'ingénierie informatique [124-126]. Il peut en effet servir à modéliser des données numériques [127, 128] ou encore à réaliser le clustering d'un ensemble d'individus [129].

Nous allons présenter succinctement les GMM et le principe de l'un des algorithmes d'estimation le plus simple.

3.3.2.1. Définition

Un modèle de mélange suppose que l'ensemble de la population est représenté par une distribution de probabilité qui est un mélange de *s* distributions de probabilités associées aux classes. L'objectif final de cette méthode est de définir les *s* distributions en estimant leurs paramètres.

Dans la théorie des probabilités, on a plusieurs types de distributions où chaque distribution suit une loi spécifiée. Nous citons quelques distributions : gaussienne, Poisson, Dirac, Bose-Einstein, Zeta ... etc. Nous nous intéressons dans notre travail à la distribution gaussienne qui suit la loi normale.

Soit un individu *x* représenté sur l'espace vectoriel $\mathbb{R}$. La densité d'un tel individu selon la loi normale (ou gaussienne) de moyenne μ et matrice de covariance Σ est donnée par l'équation (3.1).

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad (3.1)$$

En superposant et pondérant K gaussiennes on définit un mélange gaussien. On note habituellement w_k le poids (sous les contraintes $w_k > 0$ et $\sum_{k=1}^{K} w_k = 1$) et respectivement μ_k et Σ_k, respectivement la moyenne et la matrice de covariance de la k ème composante. Egalement, on notera $\theta_k = \{ w_k, \mu_k, \Sigma_k \}$, ainsi que $\theta = \{ \theta_k \}$. La densité d'un individu *x* selon la distribution de probabilité paramétrée par θ est donnée par l'équation (3.2). Cette équation est un jeu de données qui est une matrice et dont chaque ligne caractérise un individu x_k.

$$P(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{3.2}$$

On note un tel échantillon $X = (x_1, \dots, x_N)^T$ si on suppose que celui-ci est indépendamment et identiquement distribué, la probabilité jointe de cet échantillon est alors donnée par l'équation (3.3). L'utilisation de *log* de $P(x|\theta)$, est appelée vraisemblance.

$$P(x|\theta) = \prod_{n=1}^{N} P(x_n|\theta) \tag{3.3}$$

Intuitivement, si on conçoit un algorithme qui maximise cette quantité pour un échantillon donné, on obtiendra alors un modèle bien adapté à l'échantillon. On a alors résumé l'échantillon par une représentation beaucoup plus légère.

3.3.2.2. L'algorithme EM

Le courbe de la vraisemblance en fonction de θ a une forme très compliquée, avec souvent de très nombreux maxima locaux. En pratique, maximiser cette quantité directement n'est pas possible ; même pour atteindre un maximum local, le recours à un algorithme est nécessaire.

L'algorithme EM (Expectation–maximization algorithm) [130] est une méthode générale d'estimation de données manquantes, qui peut être appliquée à l'ajustement d'un mélange de Gaussiennes. Tel que formulé dans [129], il se base sur une variable binaire implicite $z = \{ z_k \}$ pour chaque individu x_n tel que $p(z_k = 1) = w_k$. On peut interpréter cette probabilité comme celle de choisir une des composantes au hasard. La probabilité conditionnelle $p(x|z_k = 1)$ est alors une loi normale.

L'algorithme EM utilise cette nouvelle variable inconnue ou manquante pour proposer une approche en 2 étapes garantissant la convergence vers un maximum local donné par l'équation (3.3).

- On initialise les paramètres du modèle à θ_0.
- La première étape E (Expectation) consiste à affecter une valeur aux z avec les paramètres de modèle courant fixés θ^{old}.
- La seconde étape M (maximisation) consiste à utiliser les valeurs courantes de z pour former une expression de l'espérance jointe de *x* et *z*. En maximisant cette expression selon θ, on obtient θ^{new}.

On alterne les étapes E et M jusqu'à convergence de la valeur observée de l'équation (3.3), c'est-à-dire la valeur calculée à partir des z et θ courants.

L'algorithme précédent permet d'ajuster un mélange de gaussiennes sur un jeu de données quelconque. La figure 3.12 présente un exemple de plusieurs étapes de l'algorithme EM. A la fin de l'algorithme, les distributions gaussiennes présentent vraiment des classes séparées de données.

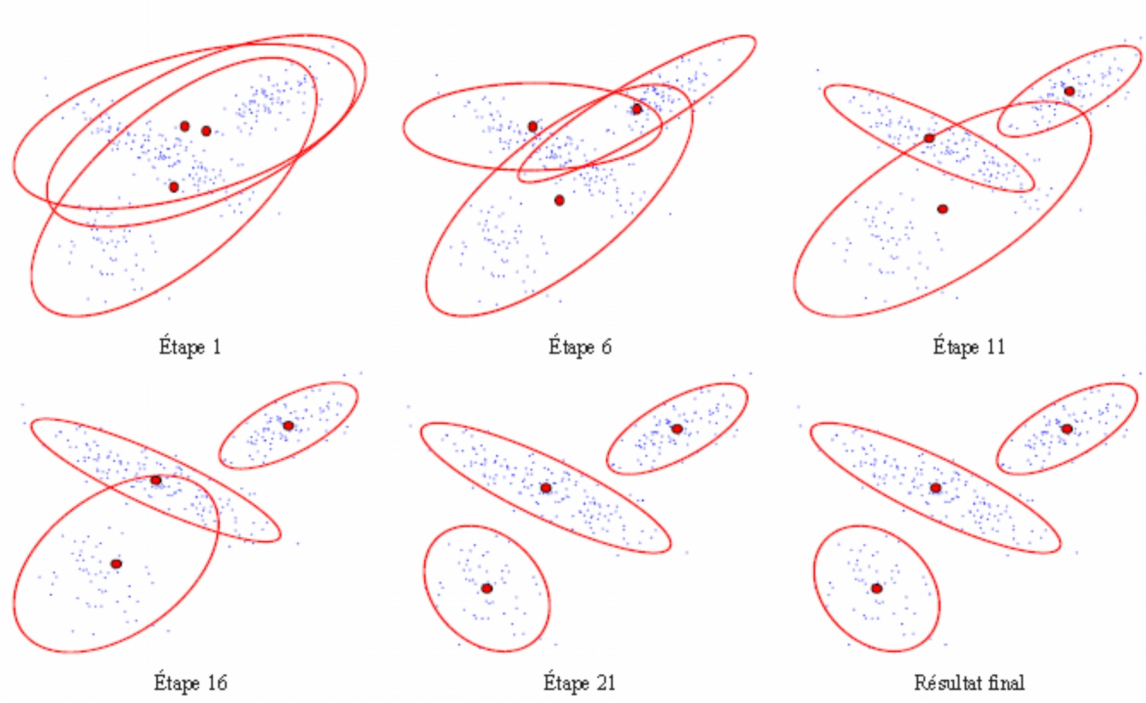

Figure 3.12. Quelques étapes de l'algorithme itératif EM en classification automatique [129].

3.3.2.3. Défauts de l'algorithme EM standard

Dans l'algorithme EM, il est nécessaire de choisir une valeur de K à priori et il appartient à l'expérimentateur de connaître ce nombre. Cela pose un problème car plus K est grand plus la vraisemblance sera élevée.

La solution la plus simple consiste à utiliser des critères pénalisant une complexité trop importante, c'est-à-dire un K trop élevé comme AIC (Akaike Information Criterion) [131] et BIC (Bayesian Information Criterion) [132]. Toutefois cela impose de faire tourner l'algorithme d'ajustement autant de fois que l'on veut avec des valeurs de K différentes.

Généralement, l'algorithme EM converge vers un maximum local, mais nous n'avons aucune garantie que ce maximum soit bon, c'est-à-dire si il est proche en valeur au maximum global. Pour se rapprocher de ce but, on doit encore alourdir le calcul.

Enfin, EM standard présente le risque de trouver des solutions dégénérées présentant au moins une composante ajustée sur un seul individu. La vraisemblance d'une telle gaussienne peut être infinie et se rapproche d'un pic de Dirac, mettant ainsi à mal notre algorithme.

3.3.2.4. EM variationnel

L'algorithme EM variationnel permet de pallier à une partie des problèmes de l'algorithme EM. La référence [133] présente les détails techniques de cette méthode.

Cet algorithme procède à une estimation bayésienne, c'est-à-dire introduisant une distinction entre la solution à priori et la solution a posteriori. A chaque étape E ou M, les paramètres à priori joueront un rôle sur l'initialisation ; c'est une forme de régularisation. Ces paramètres à priori de l'algorithme EM variationnel contribueront à une pénalisation infligée aux valeurs dégénérées.

En réalité cette technique est beaucoup plus puissante que l'algorithme EM. Pour un K initial donné, en conservant que le modèle le plus vraisemblable possible. Ainsi, on se retrouve avec un nombre final de composantes $K' < K$.

Toutefois cette technique nécessite un K initial assez grand pour couvrir la surface des vraisemblances possibles. Le choix de K va provoquer un certain surcoût de calcul, mais qui reste cependant nettement inférieur aux techniques comparables.

3.3.2.5. Classification

L'attribution d'une classe à un individu se fait toujours en estimant la probabilité d'appartenance de l'individu à chaque classe. Ce calcul peut être effectué de plusieurs façons différentes. Il faut tout d'abord estimer la probabilité due à chaque gaussienne puis combiner les probabilités des mêmes classes. Différentes stratégies sont alors possibles comme conserver la probabilité la plus élevée ou calculer la probabilité moyenne… etc.

3.3.2.6. Regroupement de GMM

Dans certaines situations on peut avoir besoin de regrouper : plusieurs mélanges de gaussiennes, des groupes d'experts, les indexations des contenus représentés par des mélanges [134]...etc. On peut donc envisager plusieurs possibilités :

a) **Addition puis moyenne de tous les mélanges :** Une méthode applicable, mais on se retrouve avec un nombre de composantes qui explose rapidement. Cela augmente le coût de transmission et de calcul de similarités (notamment la divergence de Kullback-Leibler).
b) **Ré-échantillonnage :** On gagne en précision, mais c'est la solution la plus coûteuse en termes de calcul. On doit réajuster un modèle sur un nombre d'éléments qui peut rapidement devenir très grand.

Pour solutionner ce problème, l'article [134] propose une approche avec un faible coût de calcul. Cette approche est une variante de l'algorithme EM qui est combinée avec la technique d'échantillonnage virtuelle présentée en [135].

3.3.3. Arbres de décision

Dans cette section, nous examinons brièvement les arbres de décision (AD). Ils sont des systèmes de décision à plusieurs étages dans lequel les classes sont successivement rejetées jusqu'à ce que nous atteignons une classe finalement acceptée.

La construction des arbres de décision à partir de données est une discipline ancienne [136]. La popularité de la méthode repose en grande partie sur sa simplicité. Un exemple d'arbre de décision est présenté à la figure 3.13.

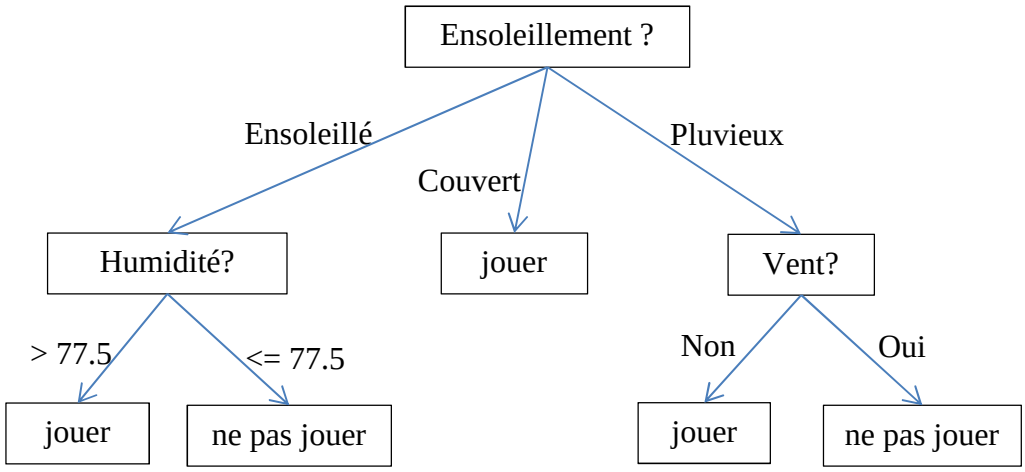

Figure 3.13. Arbre de décision pour décider de jouer ou non.

Cette méthode est basée sur la partition d'ensemble de données à des groupes plus homogènes possible du point de vue de la variable à prédire.

On prend en entrée un ensemble de données classées et on fournit en sortie un arbre où:

- ✓ Chaque nœud final (feuille) représente une décision (une classe)
- ✓ Chaque nœud non final (interne) représente un test.
- ✓ Les branches représentent les résultats des tests

L'exemple précédent de la figure 3.13 été généré à partir du tableau 3.4 [136]. Il présente un fichier composé de 14 observations qui explique le comportement des individus par rapport à un jeu {jouer, ne pas jouer} à partir des prévisions météorologiques.

Tableau 3.4. Données "weather" [136]

N°=	Ensoleillement	Température	Humidité	Vent	Jouer
1	Soleil	75	70	Oui	Oui
2	Soleil	80	90	Oui	Non
3	Soleil	85	85	Non	Non
4	Soleil	72	95	Non	Non
5	Soleil	69	70	Non	Oui
6	Couvert	72	90	Oui	Oui
7	Couvert	83	78	Non	Oui
8	Couvert	64	65	Oui	Oui
9	Couvert	81	75	Non	Oui
10	Pluie	71	80	Oui	Non
11	Pluie	65	70	Oui	Non
12	Pluie	75	80	Non	Oui

En effet, toutes les données ayant l'attribut Ensoleillement="Soleil" et l'attribut Humidité > 77.5 appartiennent à la classe 1 ("oui"). Toute nouvelle donnée peut être classée en testant ses valeurs d'attributs l'un après l'autre en commençant de la racine jusqu'à atteindre une feuille ou une décision.

3.3.3.1. Construction

Pour construire un tel arbre, plusieurs algorithmes existent : ID3, CART, C4.5,...etc. On commence généralement par le choix d'un attribut puis le choix d'un nombre de critères pour son nœud. On crée pour chaque critère un nœud concernant les données vérifiant ce critère. L'algorithme continu d'une façon récursive jusqu'à obtenir des nœuds concernant les données de chaque même classe et l'arbre est construit récursivement de haut en bas.

En réalité ce n'est pas aussi simple, plusieurs problèmes doivent être résolus :

- ? Comment choisir l'attribut qui sépare le mieux l'ensemble de données? On parle souvent de la variable de segmentation.
- ? Comment choisir les critères de séparation d'un ensemble selon l'attribut, et comment ces critères varient pour l'attribut numérique ou symbolique?
- ? Quel est le nombre optimal du nombre de critères qui minimise la taille de l'arbre et maximise la précision?
- ? Quels sont les critères d'arrêt de ce partitionnement, sachant que souvent l'arbre et d'une taille gigantesque?

3.3.3.2. Choix d'attribut

Il s'agit de choisir parmi les attributs des données, celui qui les sépare le mieux du point de vue de leurs classes déjà connues. Pour choisir le meilleur attribut, on calcule pour chacun des attributs une valeur appelée "Gain" (basé sur le gain d'entropie de Shannon) qui dépend des différentes valeurs prises par cet attribut.

On distingue :

- ✓ Gain d'information (ID3/C4.5) où tous les attributs sont catégoriels et peuvent être modifiés pour les attributs numériques.
- ✓ L'indice de Gini (IBM Intelligent Miner) où tous les attributs sont continus, et suppose qu'il y a plusieurs divisions possibles pour chaque attribut.

3.3.3.3. Gain d'information

Le gain informationnel est une mesure de segmentation qui utilise l'entropie de Shannon. ID3 et C4.5 [137-141] utilisent le gain pour choisir l'attribut pour représenter le nœud. Il conserve seulement les informations absolument nécessaires pour classer un objet. À chaque fois, qu'on doit choisir un attribut pour partitionner l'ensemble d'exemples, il faut choisir celui dont l'entropie de classification est la plus petite. Le gain privilégie est généralement celui des attributs ayant un grand nombre de valeurs [142]. Pour avoir un arbre de décision concis et suffisant, il ne faut pas seulement traiter les attributs séquentiellement. La richesse de cette mesure consiste à choisir judicieusement les attributs nécessaires comme des nœuds intermédiaires, pour arriver au chemin le plus court qui correspond de plus au plus grand nombre d'exemples dans la même classe.

Principe (1) :

Il sert à sélectionner l'attribut du gain le plus élevé. Si l'on suppose qu'il y a deux classes P et N et soit l'ensemble d'exemples S contenant p exemples de la classe P et n exemples de la classe N. La quantité d'information nécessaire pour décider qu'un exemple dans S appartienne à P ou N est définie par la formule (3. 4).

$$H(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right) \tag{3.4}$$

Supposons qu'en utilisant l'attribut A, un ensemble S sera divisé en {S1, S2,..., Sv} et si Si contient pi exemples de P et ni exemples de N ; l'entropie ou l'information attendus nécessaire pour classifier les objets dans le sous arbre Si est calculée par la formule (3.5).

$$H(S) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} H(S_i) \tag{3.5}$$

Le codage d'information qui peut être gagné en se branchant à A est donné par la formule (3.6).

$$Gain(A) = H(S) - H(A) \tag{3.6}$$

Principe (2) :

Soit un ensemble X d'exemples dont une proportion p_+ sont positive et une proportion p_- sont négative, avec $p_+ + p_- = 1$, donc l'entropie de X est calculé par la formule (3.7)

$$H(X) = -p_+ \log_2(p_+) - p_- \log_2(p_-) \quad (3.7)$$

Avec $0 \leq H(X) \leq 1$.

Si $p_+ = 0$ ou $p_- = 0$, alors $H(X) = 0$, donc si tous les exemples sont positifs, ou négatifs, l'entropie de la population est nulle. Ainsi, s'il y a autant d'exemples positifs que négatifs, l'entropie est maximale, c'est-à-dire si $p_+ = p_- = 0.5$.

$$Gain(X, a_j) = H(X) - \sum_{v \in \{aj\}} \frac{\left|X_{a_j=v}\right|}{|X|} H\left(X_{a_j=v}\right) \quad (3.8)$$

$X_{a_j=v}$ est l'ensemble des exemples, dont l'attribut considéré a_j prend la valeur v. La notation |X| indique le cardinal de l'ensemble X.

3.3.3.4. Indice de Gini (IBM Intelligent Miner)

Le critère Gini est la mesure de segmentation de l'algorithme CART (Classification and Regression Tree). Cet algorithme construit des arbres binaires, c'est-à-dire que les nœuds non terminaux ont seulement deux branches [143, 144]. Lorsqu'un attribut a plusieurs valeurs possibles, on doit faire des regroupements pour être en mesure de le partitionner en deux. Un bon critère d'éclatement doit prendre soin que l'éclatement soit fait à un nœud qui réduit le coût des erreurs de classification de l'arbre. L'indice de Gini est utile lorsque le problème comporte plusieurs classes.

Si une base T contient des exemples de n classes, Gini(T) est défini par la formule (3.9).

$$Gini(T) = 1 - \sum_{i=1}^{n} p_j^2 \quad (3.9)$$

Où p_j est la fréquence de la classe j dans T.

Si la base T est partitionnée en deux bases T1 et T2 de tailles N1 et N2 respectivement, le Gini(T) du partitionnement est défini par l'équation (3.10).

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2) \quad (3.10)$$

L'attribut de $Gini_{split}(T)$ minimum est choisi pour diviser le nœud.

3.3.3.5. Choix de la bonne taille de l'arbre

L'arbre de décision construit peut être d'une taille très importante épuisant alors les ressources de calcul et de stockage. Pour éviter le sur apprentissage, c'est-à-dire qu'on a deux nœuds qui contiennent un seul élément, on enlève les sous-arbres les moins significatifs de l'arbre afin de les remplacer par des feuilles avec un seuil d'erreur acceptable [137, 144]. De cette façon, on généralise les règles extraites de l'arbre [145].

La phase d'élagage consiste à enlever les feuilles les moins significatives de l'arbre. Cette phase est exécutée après la construction de l'arbre. On considère que la racine de l'arbre comprend au moins deux feuilles. Tant qu'il existe un sous-arbre que l'on peut remplacer par une feuille, sans faire croître l'estimation de l'erreur réelle, alors on élague ce sous-arbre. On doit savoir si les nœuds fils sont des feuilles ou des sous-arbres. Si tous les nœuds sont des feuilles, on remplace la racine du sous-arbre par une feuille, si l'erreur de la nouvelle feuille est plus petite que celle de l'ancien sous-arbre.

Pour créer la feuille de remplacement, on crée la liste des possibilités d'éléments des feuilles du sous-arbre et on prend la valeur la plus fréquente pour le nom de l'étiquette et le(s) autre(s) dans le tableau étiquette des autres possibilités. On remonte ensuite jusqu'à la racine jusqu'à ce qu'on ne puisse plus remplacer un sous-arbre par une feuille [142].

Pré élagage :

Le pré élagage se produit pendant la construction de l'arbre, il agit comme un critère d'arrêt dans l'expansion de l'arbre. Il consiste à fixer une condition d'arrêt pour arrêter la construction [146,147]. Cette condition limite l'expansion d'une branche, c'est-à-dire que les attributs restants ne permettent plus de diviser les exemples. Si on n'effectuait pas cet arrêt exceptionnel, la branche croîtrait en prenant les attributs restants.

Les exemples restants au niveau du nœud que le seuil atteint seraient les mêmes que celle au niveau du nœud terminal. Dans les algorithmes C4.5 et CART, lorsque la mesure de segmentation est égale à zéro ou à l'infini, on doit créer un nœud terminal. Il vaut mieux arrêter la phase d'expansion de la branche que de continuer sans avoir d'apport significatif dans la classification.

Post élagage :

Le post élagage est la méthode la plus utilisée dans la plupart des algorithmes, elle s'effectue après la construction de l'arbre en coupant des sous-arbres entiers et en les remplaçant par des feuilles représentant la classe la plus fréquente dans l'ensemble des données de cet arbre.

On commence par la racine et on descend, et pour chaque nœud interne (non-feuille), on mesure la complexité avant et après sa coupure (son remplacement par une feuille). Si la différence est peu importante, on coupe le sous-arbre et on le remplace par une feuille.

C4.5 utilise une estimation de l'erreur réduite de l'arbre, cette estimation est produite à partir de l'erreur apparente de l'arbre.

3.3.3.6. Algorithmes

Algorithme ID3 :

ID3 construit l'arbre de décision récursivement. À chaque étape de la récursion, il calcule parmi les attributs restant pour la branche en cours, celui qui maximisera le gain d'information, c'est-à-dire l'attribut qui permettra le plus facilement de classer les exemples à ce niveau de cette branche de l'arbre. Le calcul se fait à base de l'entropie de Shannon.

Algorithme C4.5 :

C'est une amélioration de l'algorithme ID3, il prend en compte les attributs numériques ainsi que les valeurs manquantes. L'algorithme utilise la fonction du gain d'entropie combiné avec une fonction Splitinfo pour évaluer les attributs à chaque itération.

Algorithme CART :

L'algorithme CART dont l'acronyme signifie "Classification And Régression Trees » construit un arbre de décision d'une manière analogue à l'algorithme ID3. Contrairement à ce dernier, l'arbre de décision généré par CART est binaire et le critère de segmentation est l'indice de Gini.

3.3.3.7. Règles de classification

Un arbre de décision est un bon outil pour représenter des connaissances. Il représente une connaissance experte en utilisant les nœuds pour les attributs, les branches pour les

valeurs des attributs et les feuilles pour les classes. Une règle représente une série de conditions à respecter les caractéristiques d'une classe en particulier. Chacune des feuilles de l'arbre représente une règle de classification.

Les règles de décision sont l'interprétation directe d'un arbre de décision, une règle de décision représente la lecture d'une branche de l'arbre de la racine à un nœud terminal de l'arbre. Un arbre contient plusieurs règles. Une règle de décision est constituée d'une ou plusieurs conditions et d'une classe avec laquelle elle est associée. Une condition est composée d'un attribut à tester, d'un signe et de valeurs de l'attribut qui répond à la condition. Les valeurs d'une condition sont séparées par une clause OU et les conditions d'une règle sont séparées par une clause AND. Voici un exemple d'une règle de décision:

IF Condition 1 = Val 1 **OU** Val 2 **AND** ... **AND** Condition N = Val X **THEN** Classe X

Pour évaluer les règles, on se sert du jeu de tests. Pour qu'un exemple corresponde à une règle particulière, il doit respecter toutes les conditions de cette règle, s'il respecte toutes les conditions et que la classe de l'exemple correspond à la classe de la règle, on incrémente le nombre d'exemples qui correspond à cette règle. Par contre, si un exemple respecte toutes les conditions d'une règle et que la classe de l'exemple est différente du développement de la règle, on incrémente l'erreur de cette règle.

3.4. Conclusion

Ce chapitre nous a apporté beaucoup de connaissances sur les questions liées à la surveillance des systèmes aquatiques en temps réel. Comme nous savons, le nombre d'espèces phytoplanctonique dans l'eau naturelle est inconnu, c'est-à-dire indénombrable. Donc, il est pratiquement impossible de surveiller toutes les espèces existantes.

On doit développer des classificateurs en mesure de trouver avec précision certaines cellules d'intérêts dans un échantillon. Cette tâche, qu'est la reconnaissance de cellules d'intérêt, est très importante dans un grand nombre de domaines : biotechnologie, médicale, environnementale ... etc.

De ce fait, nous avons proposé quatre modèles de traitement automatique pour la reconnaissance d'espèces phytoplanctoniques. Il s'agit d'un modèle à base de GMM et un autre à base d'arbre de décision pour l'identification et la reconnaissance de plusieurs espèces phytoplanctonique, par traitement des données cytométriques issues du CytoSense.

Les deux autres modèles classifient les données, issues du FLOWCAM, des cellules d'une espèce toxique par arbre de décision et le réseau de neurones MLP. Dans le chapitre 4, nous présentons tous les résultats et les discussions des méthodes proposées.

Chapitre 4

Résultats et discussion

4.1. Introduction

Dans ce chapitre, nous allons vous présenter tous les résultats de simulation des modèles proposés dans notre travail de recherche suivi des discussions.

La première partie est dédiée aux détails des différentes phases de traitement de modèle de prédiction présenté dans le chapitre 2 et ses résultats. La deuxième partie réservée aux modèles de reconnaissance automatique présentés dans le chapitre 3. Toutes les simulations sont faites avec des scripts de MATLAB R2014a.

4.2. Prédiction de la concentration cellulaire de D. Acuminata

Après la présentation théorique des phases de traitement dans le chapitre 2 section 2.3, nous allons former notre modèle de prédiction.

4.2.1. Base des données

La base de données présentée dans la section 2.2.4 de chapitre 2 est composée de 37 échantillons prélevés de l'eau de mer du littoral français (Havre France) durant la période juillet, août et septembre de l'année 1985. Ces données présentent l'évolution de la concentration cellulaire du D. ACUMINATA en fonction de paramètres physico-chimiques.

La base de données est divisée en trois séries temporelles, les 20 premiers échantillons prélevés en juillet sont utilisés pour l'apprentissage de réseau MLP, les 7 échantillons de mois d'août sont réservés pour le test et les 10 échantillons de mois de septembre servent pour l'évaluation de prédiction de réseau à court terme.

4.2.2. Optimisation des attributs d'entrées avec PCA

Nous avons appliqué la méthode de PCA sur les 11 variables qui affectent la croissance cellulaire de D. Acuminata. Parmi les 11 variables présentées dans le tableau 2.1 de chapitre 2, l'algorithme PCA va déterminer et sélectionner les variables les moins corrélées. Ces variables sélectionnées correspondent à la plus grande inertie de la base de données.

Les calculs sont réalisés par le scripte PCA de Matlab qui génère automatiquement:

- ✓ La matrice des coefficients de corrélation des 11 variables ; présentée dans le tableau 4.1.

- ✓ La matrice des valeurs des composants principales de la matrice de variance ; présentés dans le tableau 4.2.
- ✓ Le vecteur de pourcentages de variance de chaque variable ; présentée dans le tableau 4.3.

Tableau 4.1. La matrice des coefficients de corrélation des 11 variables obtenus par le scripte PCA de Matlab.

	T	O2	NO2	NO3	NH4	PO4	COT	COD	Pigm	CHLOR	SI
T	0,0489	-0,0081	-0,0508	0,0060	0,1035	0,0040	-0,0958	0,9819	-0,0573	-0,0806	0,0327
O2	0,0136	0,0097	-0,0080	0,0052	0,0255	0,0086	0,0334	0,0878	0,0393	0,8213	-0,5604
NO2	0,2214	0,0821	-0,2280	0,0082	0,6110	0,6782	0,1324	-0,0687	0,1881	-0,0364	0,0053
NO3	-0,5203	0,7875	0,1646	-0,1533	0,2212	-0,0421	-0,0824	0,0088	-0,0308	-0,0017	-0,0016
NH4	-0,2461	-0,3226	0,6276	0,4919	0,4391	-0,0722	0,0330	-0,0051	-0,0217	-0,0013	0,0005
PO4	-0,0008	0,0447	-0,0341	-0,0127	0,0161	-0,0279	0,6542	0,0203	-0,5123	-0,3147	-0,4536
COT	-0,0072	0,0378	-0,0306	-0,0020	0,0168	-0,0263	0,4875	0,0443	-0,2622	0,4588	0,6912
COD	-0,0773	0,0533	0,0642	0,0107	-0,1345	-0,1570	0,5443	0,1154	0,7908	-0,0895	-0,0355
Pigm	0,4805	0,0513	0,6353	-0,5982	0,0666	-0,0043	0,0205	0,0078	-0,0040	0,0034	0,0080
CHLOR	0,5636	0,5095	0,1262	0,6100	-0,1792	-0,0401	-0,0264	-0,0063	-0,0177	-0,0062	0,0047
SI	-0,2501	-0,0205	0,3140	0,0656	-0,5643	0,7108	0,0349	0,0865	-0,0398	0,0096	0,0017

Tableau 4.2. La matrice des composants principaux de la matrice de covariance obtenue par le scripte PCA de Matlab.

	T	O2	NO2	NO3	NH4	PO4	COT	COD	Pigm	CHLOR	SI
T	151,322809	0	0	0	0	0	0	0	0	0	0
O2	0	74,26056878	0	0	0	0	0	0	0	0	0
NO2	0	0	56,60076617	0	0	0	0	0	0	0	0
NO3	0	0	0	30,46525243	0	0	0	0	0	0	0
NH4	0	0	0	0	16,8465086	0	0	0	0	0	0
PO4	0	0	0	0	0	6,335441964	0	0	0	0	0
COT	0	0	0	0	0	0	1,551771664	0	0	0	0
COD	0	0	0	0	0	0	0	0,761306619	0	0	0
Pigm	0	0	0	0	0	0	0	0	0,610448451	0	0
CHLOR	0	0	0	0	0	0	0	0	0	0,447542321	0
SI	0	0	0	0	0	0	0	0	0	0	0,05070573

Tableau 4.3. Le pourcentage de variance pour chacune des variables obtenues par le scripte PCA de Matlab.

	variance %
T	44,605
O2	21,889
NO2	16,684
NO3	8,980
NH4	4,966
PO4	1,867
COT	0,457
COD	0,224
Pigm	0,180
CHLOR	0,132
SI	0,015
	100,000

Les cinq variables avec 97% de variance totale de la base des données, c.-à-d. T, O2, NH4, NO2, NO3, sont sélectionnées comme des entrées principales de notre système. Les cinq variables sont présentées dans la figure 4.1. Ces cinq paramètres représentent les données les plus significatives dans la base de données à cause de leur faible corrélation.

On a rajouté deux variables aux cinq variables obtenues par PCA pour prendre le maximum d'inertie possible. Plusieurs combinaisons ont été testées et finalement trois regroupements G1, G2 et G3 correspondent aux meilleurs résultats obtenus étaient créés ; ils sont représentés dans le Tableau 4.4.

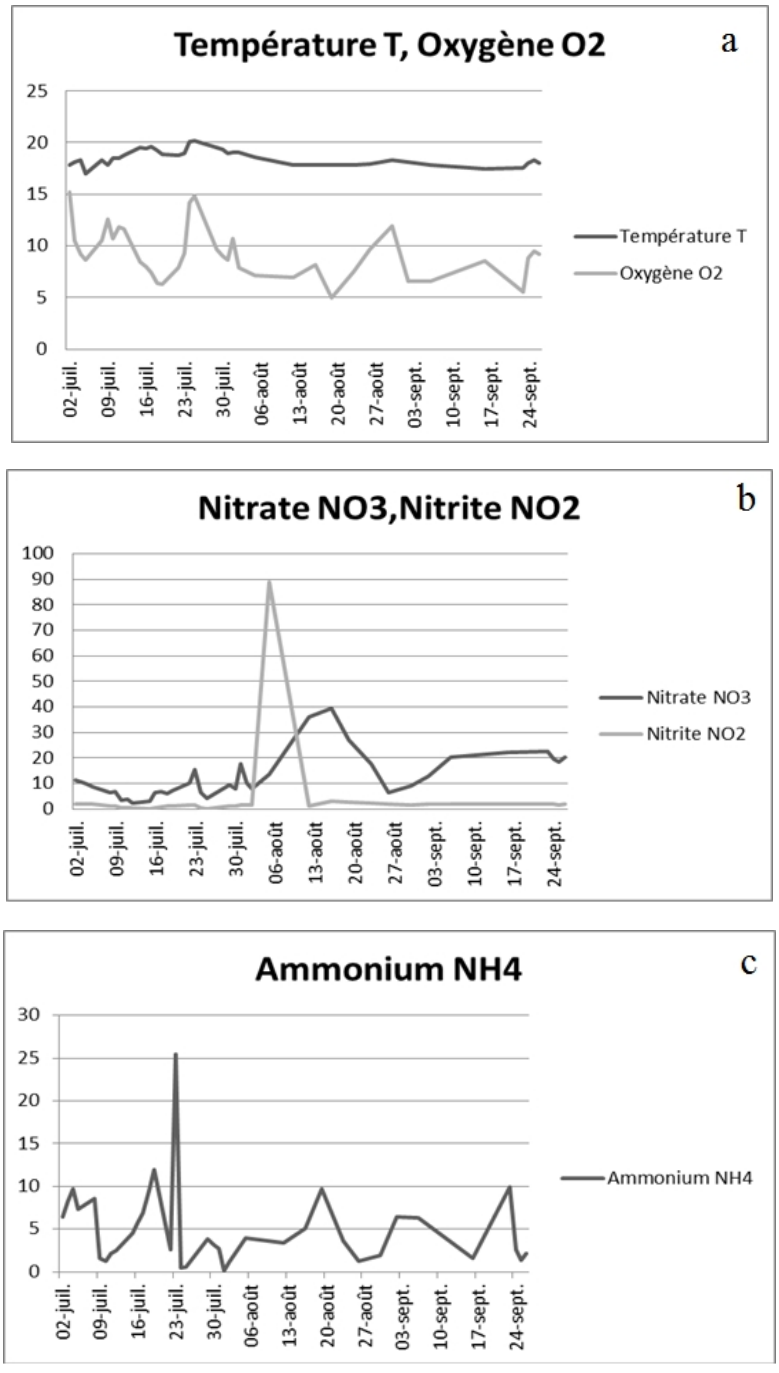

Figure 4.1 Les courbes des variations des attributs prépondérants déterminés par le scripte PCA de Matlab : a. (T, O2), b. (NO3, NO2), et c. (NH4).

Tableau 4.4. Les trois regroupements réalisés.

G1	G2	G3
T	T	T
O2	O2	O2
NO3	NO3	NO3
NO2	NO2	NO2
NH4	NH4	NH4
COT	PIgm	PO4
COD	CHLOR	SI

4.2.3. Simulations et résultats

Nous avons procédé à des simulations avec un modèle de réseau de neurones BP, présenté dans la section 2.3.3.2 du chapitre 2. Le réseau proposé contient une couche cachée. Nous avons testé ce modèle avec plusieurs variantes de nœuds dans la couche cachée (10, 9, 8 7, 6).

Toutes les données ont été normalisées selon la formule (4.1).

$$P_{normalisé} = (P - P_{min})/(P_{max} - P_{min}) \qquad (4.1)$$

Ou P représente une donnée.

Toutes les couches contiennent des entrées biaisées $b = 1$. Une fonction de transfert linéaire est utilisée dans les couches de sortie. Les couches cachées contiennent des fonctions de transfert sigmoïdes. Le réseau est entrainé avec l'algorithme de Levenberg-Marquardt.

L'organigramme d'apprentissage et de validation de réseau est représenté au Figure 4.2. Après plusieurs essais, on a fixé les MSE d'apprentissage à 0.01 et de test à 0.02 pour avoir les meilleurs résultats possibles. Les contraintes de la phase de test conduisent à de bons résultats de prédiction de D. Acumunata.

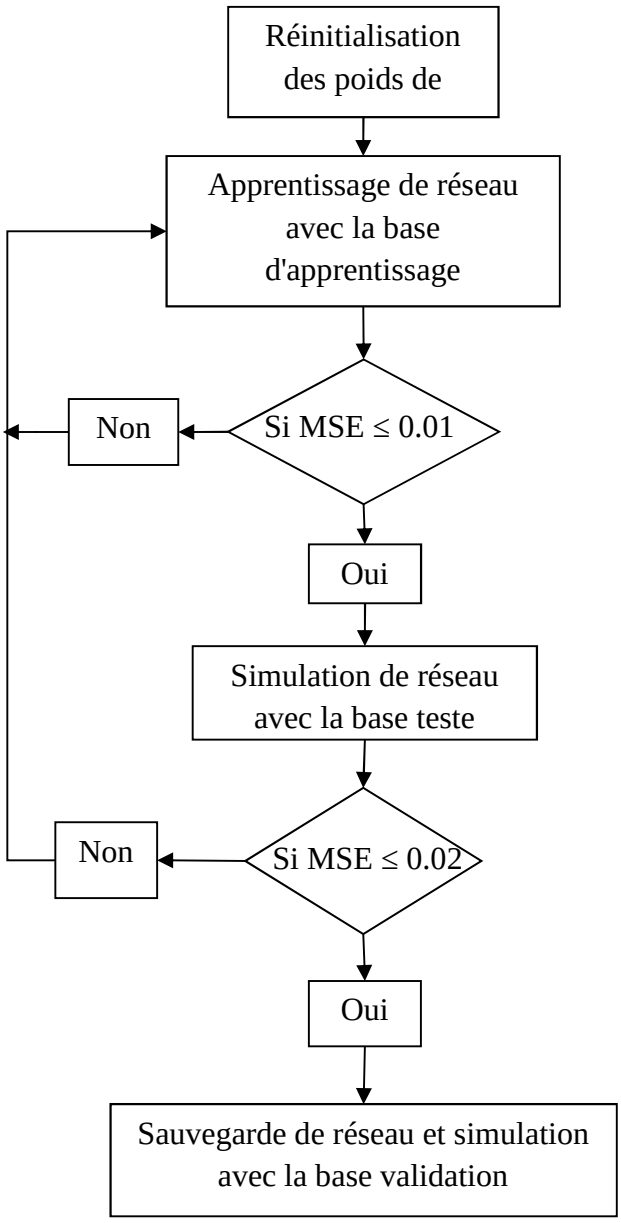

Figure 4.2 Organigramme de la méthode utilisée pour l'apprentissage et la validation de réseau MLP.

Les architectures des réseaux utilisées ont été testées et simulées sous Matlab. Les MSE obtenues entre la cible et la sortie du réseau avec les différents regroupements d'attributs à savoir G1, G2 et G3 sont représentées dans les Tableaux 4.5, 4.6 et 4.7 et sous forme de graphes dans les figures 4.3, 4.4 et 4.5.

Tableau 4.5. Les valeurs de MSE pour le regroupement d'attributs G1 avec les différentes architectures.

Nombre de neurones dans la couche cachée	MSE de G1: T, O2, NO3, NO2, NH4, COT, COD		
	Apprentissage	Test	Prédiction
6n	0.0016	0.0181	0.1053
7n	0.0078	0.0166	0.0870
8n	0.0084	0.0123	0.1305
9n	0.0064	0.0145	0.0831
10n	0.0045	0.0124	0.0917
10-6n	0.0032	0.0169	0.0458

Tableau 4.6. Les valeurs de MSE pour le regroupement d'attributs G2 avec les différentes architectures.

Nombre de neurones dans la couche cachée	MSE de G2: T, O2, NO3, NO2, NH4, PIgm, CHLOR		
	Apprentissage	Test	Prédiction
6n	0.0082	0.0160	0.1312
7n	0.0187	0.0147	0.0786
8n	0.0244	0.0222	0.1084
9n	0.0278	0.0141	0.0762
10n	0.0053	0.0190	0.1208
10-6n	0.0098	0.0184	0.1313

Tableau 4.7. Les valeurs de MSE pour le regroupement d'attributs G3 avec les différentes architectures.

Nombre de neurones dans la couche cachée	MSE de G3: T, O2, NO3, NO2, NH4, PO4, SI		
	Apprentissage	Test	Prédiction
6n	0.0270	0.0147	0.1267
7n	0.0308	0.0275	0.0869
8n	0.0374	0.0186	0.1046
9n	0.0307	0.0218	0.0829
10n	0.0086	0.0203	0.1210
10-6n	0.0183	0.0133	0.1253

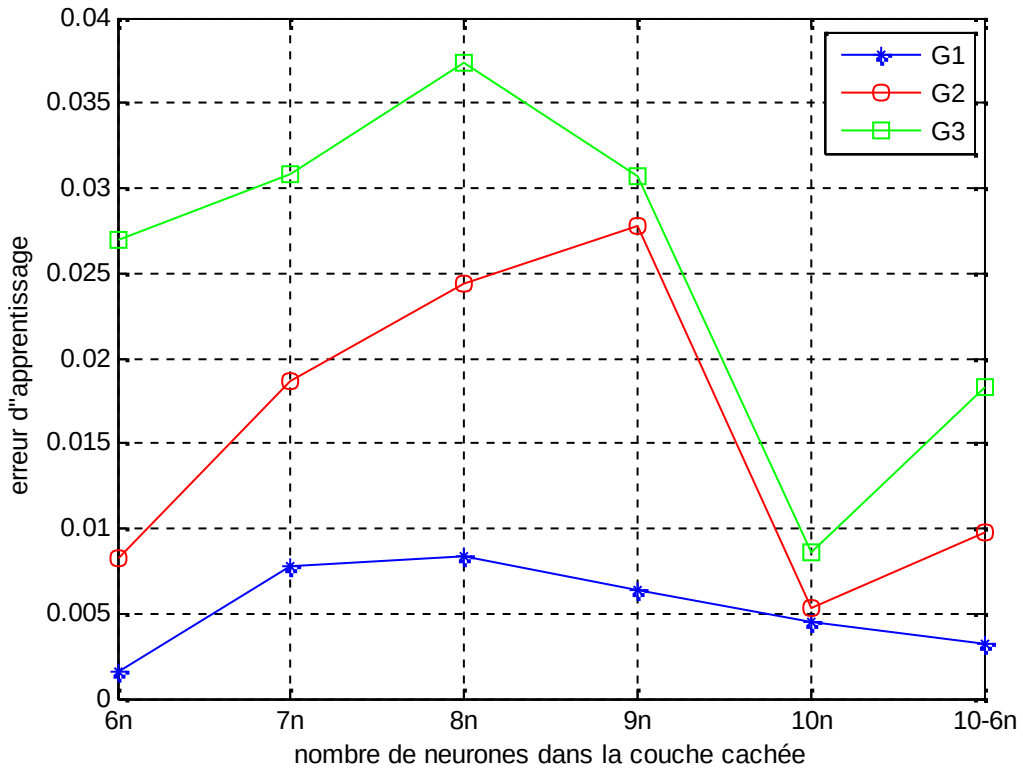

Figure 4.3 L'erreur d'apprentissage pour les différents regroupements G1, G2, G3.

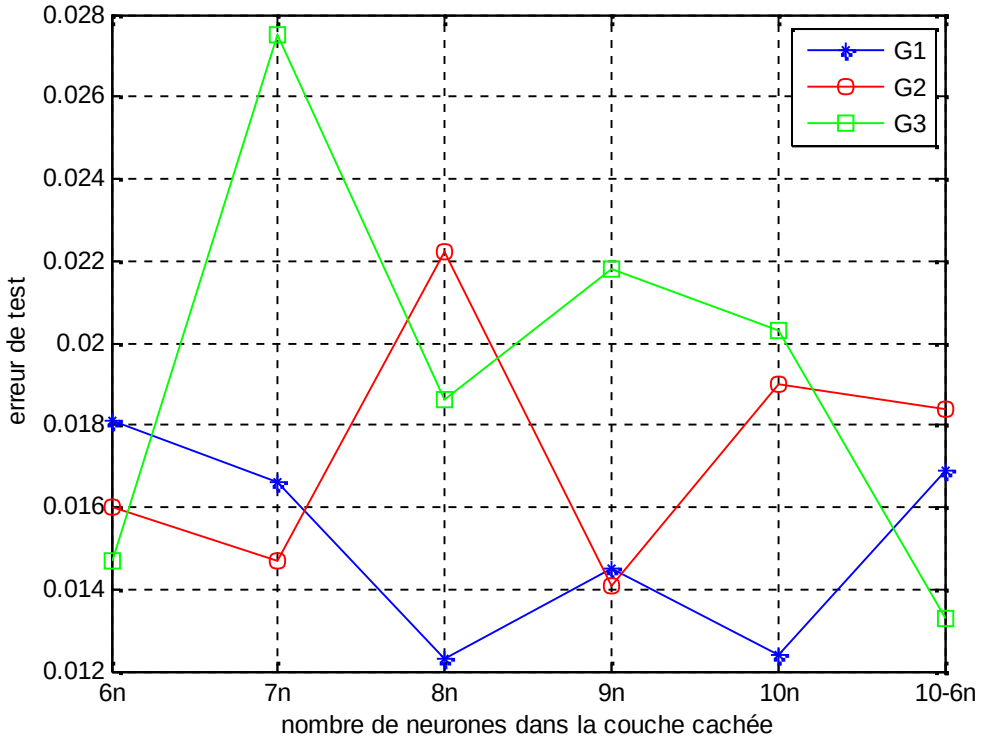

Figure 4.4 L'erreur de test pour les différents regroupements G1, G2, G3.

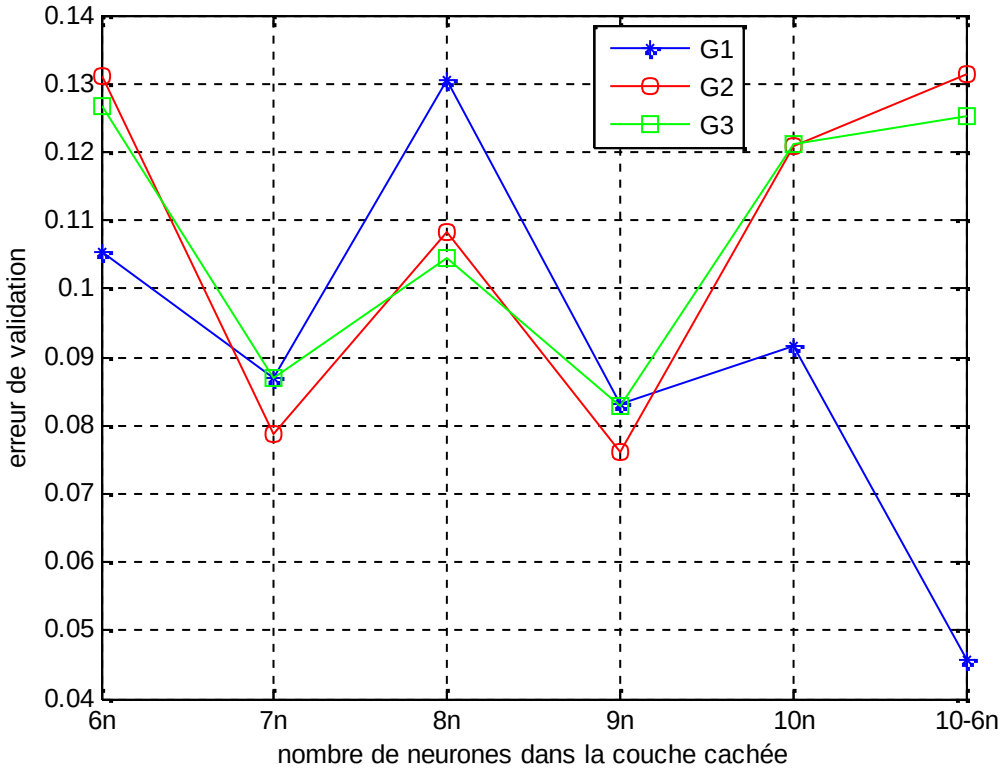

Figure 4.5 L'erreur de validation pour les différents regroupements G1, G2, G3.

Le regroupement G1 (T, O2, NO3, NO2, NH4, COT, COD) est celui qui a donné les meilleurs résultats de simulations. Présentés au tableau 4.5, les MSE de ce groupe varient pour l'apprentissage entre 0.0016 et 0.0084 et pour le test entre 0.0123 et 0.0181.

Le nombre des neurones dans la couche cachée influence aussi la prédiction de réseau. Le Tableau 4.5 présent les MSE d'apprentissage, ils varient entre 0.0016 (pour 6 neurones) et 0.0084 (pour 8 neurones), pour le test entre 0.0123 (pour 8 neurones) et 0.081 (pour 6 neurones) et pour la validation entre 0.0831 (pour 9 neurones) et 0.1305 (pour 8 neurones). Les meilleurs résultats sont présentés dans les Figures 4.6 et 4.7 et ils correspondent respectivement aux architectures à 9 neurones et 10 neurones.

On observe que les bons résultats sont ceux des architectures les plus peuplées en neurones dans la couche cachée, c'est-à-dire avec 9 et 10 neurones. Les autres architectures de 6, 7 et 8 neurones souffrent de manque des données d'apprentissage. Ils peuvent être utilisés en cas de suivi de processus à l'aide d'une base d'apprentissage plus importante. On constate que l'utilisation d'un plus grand nombre de neurones nous conduit à de bons résultats en phase de prédiction.

On a essayé quelques architectures en deux couches cachées pour mieux améliorer nos résultats. Après plusieurs tentatives et avec une multitude d'architectures, on a abouti à l'architecture 10 neurones dans la première couche et 6 neurones dans la deuxième couche.

Cette architecture nous a permis d'atteindre un grand niveau de précision dans toutes les phases. Le Tableau 4.5 présent le MSE d'apprentissage 0.0032, de test 0.0169 et de validation 0.0458. La figure 4.8 représente le nombre de cellules réelles et le nombre de cellules estimées.

L'utilisation d'un nombre très limité de ressources et de données présente un atout majeur de notre travail [148]. La prédiction à court terme, c.-à-d. 10 jours, d'un bloom avec un réseau de neurones a deux couches, entrainé avec des données de deux mois, est très avantageuse surtout pour la surveillance des vastes zones, où les données sont très laborieuses à collecter.

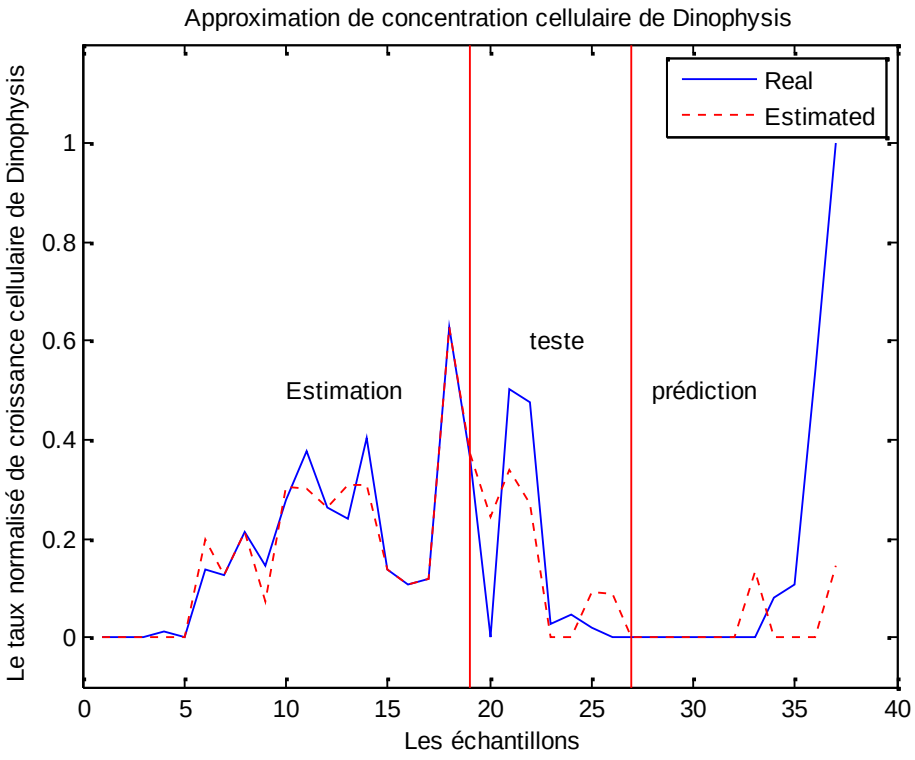

Figure 4.6 L'architecture avec 9 nœuds dans la couche cachée.

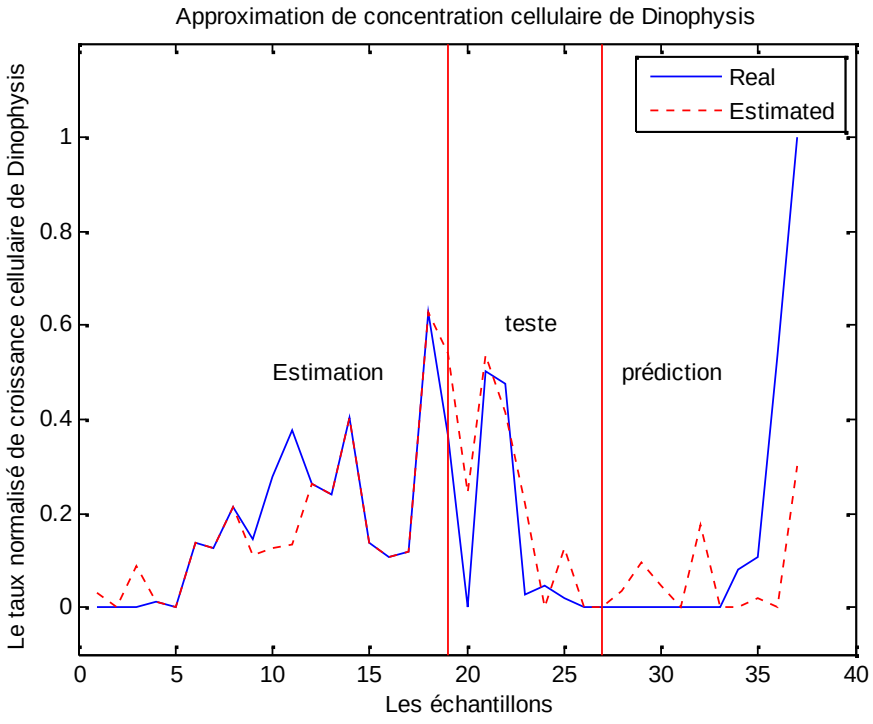

Figure 4.7 L'architecture avec 10 nœuds dans la couche cachée.

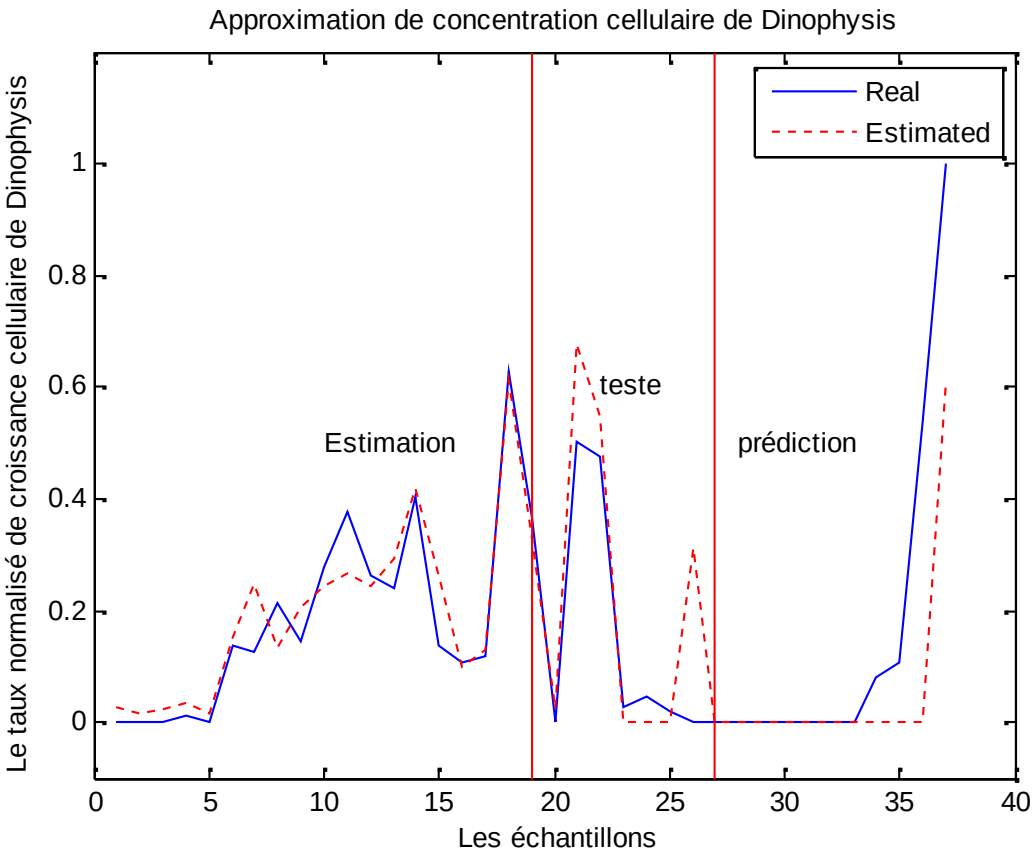

Figure 4.8 L'architecture avec 10 et 6 nœuds dans les couches cachées.

4.3. Reconnaissance et identification des espèces phytoplanctoniques

Dans cette section, nous présentons notre travail pour la réalisation d'un système automatique de reconnaissance et d'identification des espèces phytoplanctoniques. L'instrument de mesure, le cytomètre en flux, ainsi que les techniques de GMM et arbre de décision ont été déjà décrits précédemment dans le chapitre 3.

4.3.1. Base de données

La base de données est la même utilisé dans le travail de A. Malkassian [118]. C'est une collection de 20 espèces phytoplanctoniques analysées avec le cytomètre en flux . Les phytoplanctons appartiennent à plusieurs groupes taxonomiques et à différentes origines, c'est-à-dire eau douce et salé.

Cette base de données est une partie de la Culture Collection Yerseke (CCY) du Centre pour l'écologie estuarienne et marine (Estuarine and Marine Ecology (NIOO)) de Yerseke, Pays-Bas.

Les 20 espèces phytoplanctoniques sont: Anabaena cylindrical, Ankistrodesmus acicularis, Aphanizomenon sp, Chaetoceros muelleri, Chlorella sp, Ditylum brightwellii, Emiliania huxlyei, Gloeothece sp, Hemiselmis sp, Isochrysis sp, Melosira sp, Monoraphidium sp, Nodularia sp, Pavlova sp, Porphyridium sp, Pseudanabaena sp, Pediastrum sp, Rhodomonas sp, Skeletonema costatum, Thalassiosira pseudonana.

Le cytomètre en flux CytoSub génère un fichier de données de 64 paramètres. La figure 4.9 présente 4 exemples d'empreintes numériques générées par le cytomètre en flux pour les espèces : Anabaena cylindrical, Ditylum brightwellii, Melosira sp et Rhodomonas sp.

L'observation des signaux de la Figure 4.9 montre clairement que l'empreinte digitale de chaque espèce est spécifique, donc il y a absolument certaines règles qui séparent chaque espèce des autres.

Nous avons utilisé dans ce travail environ 13330 empreintes de 20 différentes espèces de phytoplanctons. Tous les 64 paramètres cytométrique de chaque cellule ont été utilisés directement sans optimisation.

4.3.2. Modèle GMM

Tous les travaux mentionnés à la section 3.3.1 du chapitre 3 consistent à trouver des caractéristiques appropriées qui permettent une séparation efficace des classes d'une manière supervisée ou nonsupervisé.

Le GMM est un classificateur populaire de données CF [120-123]. Son rôle est de représenter les caractéristiques extraites par une somme pondérée de densités de M gaussiennes, et à classer les données selon leur degré d'appartenance à ces distributions.

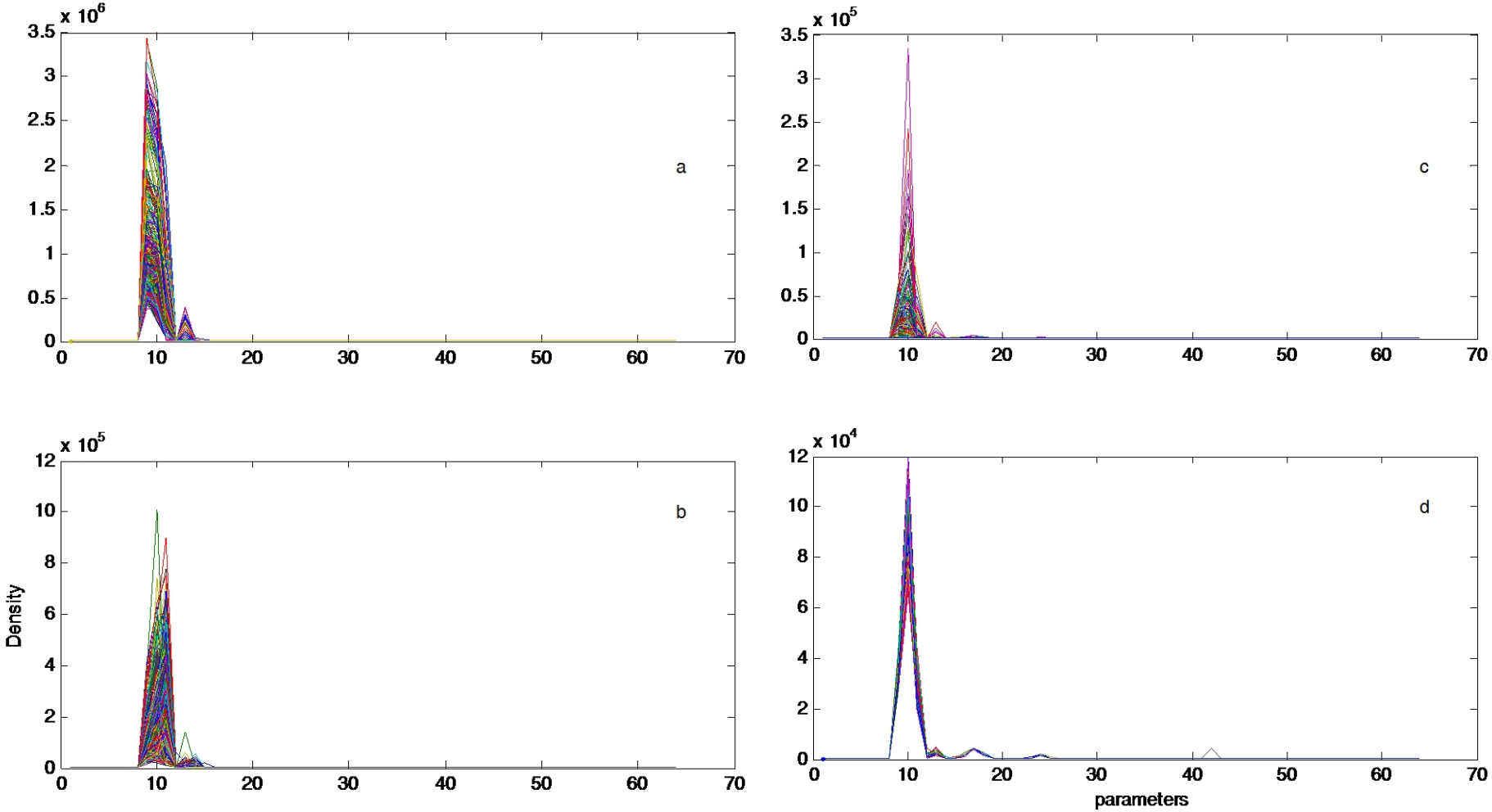

Figure 4.9 Les empreintes numériques générées par le cytomètre en flux pour : (a). Anabaena, cylindrique ; (b). Brightwellii Ditylum ; (c). Melosira sp et (d). Rhodomonas sp.

L'originalité de notre travail provient de:

1) L'application d'une méthode principalement utilisée pour la reconnaissance vocale. Cette méthode classe les données avec les modèles de mélange gaussien (GMM). L'apprentissage consiste à créer un modèle GMM de chaque classe et la classification utilise les distances entre les GMM pour décider l'appartenance des données aux classes déjà formées. Cette approche a démontré son efficacité dans de nombreux travaux de recherche [149-151].

2) La proposition est d'utiliser une nouvelle représentation des données FC qui peut supporter cette méthode. Le vecteur FC de la cellule était changé en une forme matricielle. Nous pouvons ainsi créer un modèle GMM pour chaque cellule à classer. La distance entre le GMM de la cellule et les GMM des classes est le critère de classification.

Nous présentons nos premiers résultats en utilisant la nouvelle approche pour la classification supervisée d'une partie de notre base de données avec une comparaison de nos résultats avec une classification GMM classique. La partie utilisée de la base de données contient 10 espèces qui sont : Ankistrodesmus acicularis, Chlorella sp, Gloeothece sp, Hemiselmis sp, Melosira sp, Nodularia sp, Pavlova sp, Pseudanabaena sp, Rhodomonas sp, Skeletonema costatum.

4.3.2.1. Représentation des données

Un fichier FC de CytoSub (un exemple est présenté au tableau 3.1, chapitre 3) est stocké comme une matrice N× (P*S), où N est le nombre d'échantillons (cellules), P le nombre de paramètres mesurés de chaque signal FC et S est le nombre de signaux. Le CytoSub enregistre S = 7 signaux et calcule P = 9 paramètres de ses signaux plus le paramètre TOF du signale FWS. Le paramètre TOF n'est pas utilisé comme entrée.

Le Tableau 4.8 contient tous les signaux et les paramètres d'un vecteur FC. Cette forme de tableau 4.8 est celle utilisée ultérieurement comme forme matricielle de chaque cellule. La section 3.2.3 du chapitre 3 détaille chaque paramètre et signal.

Tableau 4.8. Signaux et paramètres d'un fichier FC de CytoSub.

signaux	**paramètres**								
	Length	**Total**	**Max**	**Avg**	**Inertia**	**CG**	**Fill**	**Asymm**	**#Cells**
FWS	Length FWS	Total FWS	Max FWS	Avg FWS	Inertia FWS	CG FWS	Fill FWS	Asymm FWS	#Cells FWS
SWS HS	Length SWS HS	Total SWS HS	Max SWS HS	Avg SWS HS	Inertia SWS HS	CG SWS HS	Fill SWS HS	Asymm SWS HS	#Cells SWS HS
FL Red HS	Length FL Red HS	Total FL Red HS	Max FL Red HS	Avg FL Red HS	Inertia FL Red HS	CG FL Red HS	Fill FL Red HS	Asymm FL Red HS	#Cells FL Red HS
FL Orange HS	Length FL Orange HS	Total FL Orange HS	Max FL Orange HS	Avg FL Orange HS	Inertia FL Orange HS	CG FL Orange HS	Fill FL Orange HS	Asymm FL Orange HS	#Cells FL Orange HS
SWS LS	Length SWS LS	Total SWS LS	Max SWS LS	Avg SWS LS	Inertia SWS LS	CG SWS LS	Fill SWS LS	Asymm SWS LS	#Cells SWS LS
FL Red LS	Length FL Red LS	Total FL Red LS	Max FL Red LS	Avg FL Red LS	Inertia FL Red LS	CG FL Red LS	Fill FL Red LS	Asymm FL Red LS	#Cells FL Red LS
FL Orange LS	Length FL Orange LS	Total FL Orange LS	Max FL Orange LS	Avg FL Orange LS	Inertia FL Orange LS	CG FL Orange LS	Fill FL Orange LS	Asymm FL Orange LS	#Cells FL Orange LS

La base de données de chaque fichier FC (espèce) était divisée en deux groupes: 75% des échantillons utilisés pour l'apprentissage et 25% sont utilisés pour le test.

Pour chaque espèce, les données ont été normalisées autour de leur écart type. Ensuite deux procédures d'apprentissage sont exécutées en fonction de la représentation des cellules. La figure 4.10 présente la création de deux procédures :

- ✓ **Premier groupe :** L'apprentissage avec les données FC brutes ; c'est-à-dire toutes les cellules sont présentées comme des vecteurs.
- ✓ **Deuxième groupe :** L'apprentissage avec les données FC modifiées sous forme matricielle. Les cellules sont représentées comme étant des petites matrices (7×9).

L'idée du deuxième groupe est venue des systèmes de reconnaissance de la parole [149] où le signal acoustique est divisé en trames (petits signaux) et chaque trame est caractérisée par l'extraction de ses paramètres. Le signal acoustique est alors représenté par une matrice de taille (nombre de trames x nombres de paramètres).

À l'aide des données de CytoSub, chaque cellule de données FC est représentée par une matrice (7×9) qui est représentée dans le tableau 4.8. Toutes les matrices cellules sont recueillies pour former une seule matrice qui présente l'espèce.

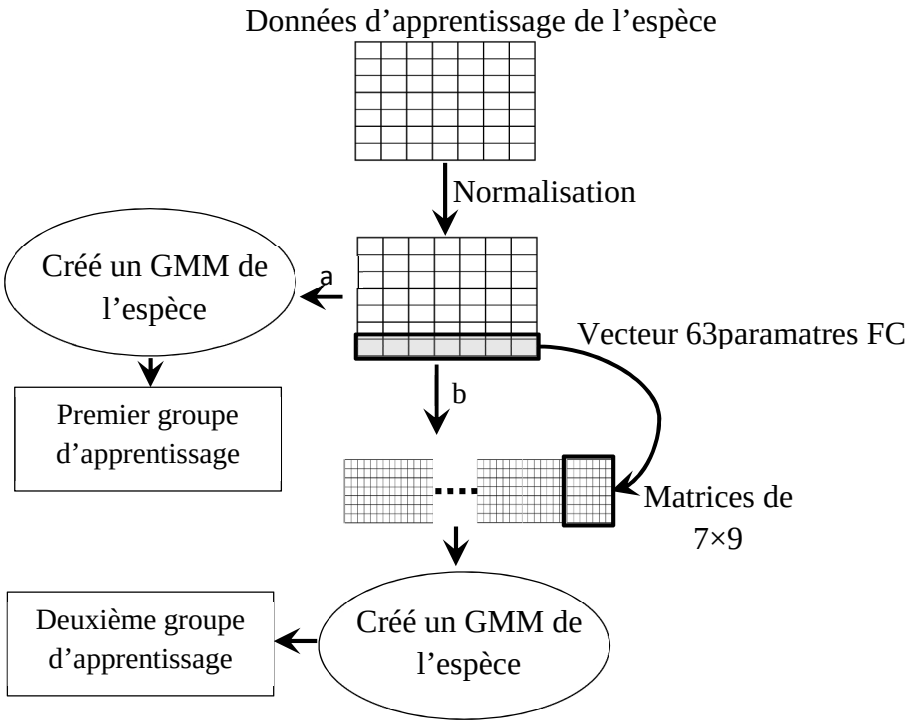

Figure 4.10 Les d'eux groupes d'apprentissage créés avec : a) représentation vectorielle ; b) représentation matricielle

4.3.2.2. Distance de Fréchet

L'approche proposée dans ce travail utilise une distance pour la classification au lieu d'utiliser la probabilité d'appartenance. L'idée principale de la distance entre les GMM est de bénéficier de toutes les informations du modèle GMM qui sont incluses dans les poids, les moyennes et la matrice de covariance. On notera que la distance euclidienne ne parvient pas à mesurer la distance entre les distributions.

Pour résoudre ce problème, des méthodes qui ont la capacité de mesurer la distance entre les distributions sont utilisées. Dans de nombreuses applications de reconnaissance de formes, la distance Fréchet, la divergence de Kullback-Leibler et la distance de Bhattacharyya sont largement utilisées pour mesurer la distance entre les distributions [152-154]. Nous avons opté dans notre travail d'utiliser la distance de Fréchet pour uniquement sa simplicité.

La distance de Fréchet [155] entre deux distributions normales multivariées X et Y avec respectivement μ_x, μ_y leurs moyennes et Σ_x, Σ_y leurs covariances est donnée par la formule (3.11).

$$d^2(X,Y) = \left|\mu_x - \mu_y\right|^2 + tr\left(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{\frac{1}{2}}\right) \quad (3.11)$$

L'équation (3.11) est composée de deux termes: une distance euclidienne entre les moyennes et une autre sur l'espace des matrices de covariance.

4.3.2.3. Scénario de simulation

Le nombre de mélanges dans un GMM influe généralement sur la performance du modèle. De nombreuses expériences sont réalisées pour tester le facteur du nombre de mélanges. Toutes les GMM qui modélisent les espèces ont un nombre différent de gaussiennes qui mènent aux meilleurs résultats de classification des données d'apprentissage.

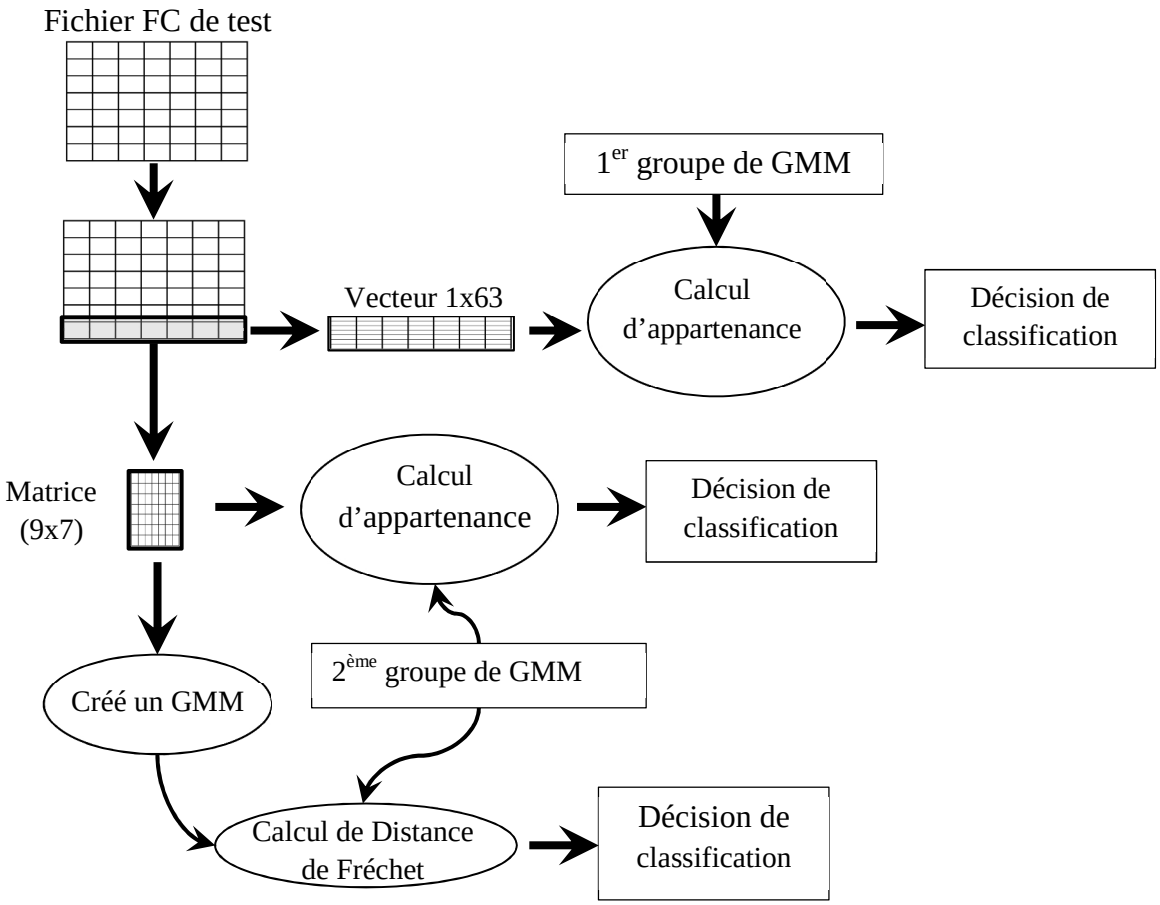

Figure 4.11 L'organigramme des trois expériences

Les GMM sont formés en utilisant l'algorithme EM qui est représenté dans la section 3.3.2.2, afin d'obtenir le maximum de vraisemblance (maximum likelihood ML). L'initialisation d'algorithme EM est assurée avec l'algorithme des K-means. La création des GMM est assurée avec la boîte à outils DCPR de MATLAB. Trois tests ont été effectués et sont représentés dans la figure 4.11.

4.3.2.4. Évaluation de la méthode

Les performances de chaque modèle d'espèces pourraient être présentées par la matrice de confusion présentée dans le tableau 4.9 [156].

Tableau 4.9. La matrice de confusion calculée pour chaque modèle d'espèces : sens: la sensibilité ; spec: spécificité ; ACC: exactitude ; VAN: valeur prédictive négative ; et prec: précision [156].

Modèle de décision	Cible		
	Positive	Négative	
Positive	a	b	$Prec = \frac{a}{a+b} * 100$
Négative	c	d	$Prec = \frac{a}{a+b} * 1$ $NPV = \frac{d}{c+d} * 100$
	$Sens = \frac{a}{a+c} * 100$	$Spec = \frac{d}{d+b} * 100$	$NPV = \frac{d}{c+d} * 100$ $Acc = \frac{a+d}{a+d+c+b} * 100$

Dans le tableau 4.9, *a* présente les cellules d'espèces correctement identifiées, *b* les cellules d'espèce classées incorrectement, *c* les cellules d'autres espèces classées comme l'espèce à identifier et *d sont des* cellules d'autres espèces qui sont classées correctement.

Avec la matrice de confusion, nous pouvons calculer les paramètres de performance tels que la sensibilité, la spécificité, la précision et l'exactitude. Dans ce travail, nous utilisons seulement la sensibilité et la précision. La sensibilité est définie comme étant le rapport entre les cellules correctement classées et le nombre total de toutes les cellules de l'espèce. La précision est la proportion du nombre total de cellules qui ont été classées correctement.

4.3.2.5. Résultats et discussion

Les simulations sont réalisées avec MATLAB 8.3 R2014a en utilisant les boîtes à outils statistiques et DCPR. Les résultats de classification par la méthode proposée ainsi ceux de la méthode classique GMM sont présentés dans le tableau 4.10.

Tableau 4.10. Les résultats de la classification avec la méthode proposée par rapport à la classification classique GMM.

algues	Nombre de cellules tests	Premier groupe				Deuxième groupe							
		sans bruit		avec bruit		Distance entre GMM				Degré d'appartenance avec cellule sous forme matricielle			
						sans bruit		avec bruit		sans bruit		avec bruit	
		Sens %	Acc %	Sens %	Acc %	Sens %	Acc %	Sens %	Acc %	Sens %	Acc %	Sens %	Acc %
Ankistrodesmus acicularis	294	100	100	99,6	99,9	100	100	100	100	99,6	99,9	96,5	99,6
Chlorella sp	277	100	100	99,6	99,9	100	100	100	100	100	100	99,6	99,9
Gloeothece sp	147	100	100	31,2	95,8	91,1	99,4	88,4	99	82,9	98,9	2	94
Hemiselmis sp	177	100	100	83	98,7	100	100	100	100	100	100	20,9	94,3
Melosira sp	56	92,8	99,8	5,3	97,7	76,7	99,4	67,8	99,2	92,8	99,8	3,5	97,7
Nodularia sp	84	98,8	99,9	0	96,5	67,8	98,8	38	97,8	80,9	99,3	26	97,4
Pavlova sp	300	99,3	99,9	4	88,5	100	100	100	100	98,6	99,8	12	89,5
Pseudanabaena sp	335	100	100	17	87	100	100	100	100	99,7	99,9	46,8	92,9
Rhodomonas sp	176	99,4	99,9	99,4	99,9	100	100	100	100	99,4	99,9	93,7	99,5
Skeletonema costatum	291	99,3	99,9	28,5	91,7	77,6	97,4	68	96,3	79	97,5	5,4	89

Les espèces simulées sont: Ankistrodesmus acicularis, Chlorella sp, Gloeothece sp, Hemiselmis sp, Melosira sp, Nodularia sp, Pavlova sp, Pseudanabaena sp, Rhodomonas sp, Skeletonema costatum.

Les résultats montrent qu'aucun progrès n'a été donné par l'approche proposée. Dans la plupart des cas, les performances de classification de la méthode proposée sont identiques à la classification classique. Pour certaines espèces l'approche classique pourrait avoir de meilleurs résultats de sensibilité ; comme dans le cas des espèces : Melosira sp, Nodularia sp et Skeletonema costatum.

En ajoutant un bruit gaussien dans le fichier de test, l'approche proposée a donné de bons résultats. Sa résistance est très forte au bruit par rapport à la classification classique GMM (voir tableau 4.10).

Le tableau 4.10 montre que l'utilisation de la représentation de la matrice de cellules n'apporte aucun avantage par rapport à la classification classique. Cela signifie que la méthode que nous avons adoptée pour classer les cellules, c'est-à-dire basés sur la recherche de la classe la plus apparente, a provoqué une grave erreur de classification.

En rajoutant du bruit aux données de test, la classification classique GMM a beaucoup diminué en termes de sensibilité, de 98,8% sans bruit à 0% avec le bruit dans le cas de l'espèce Nodularia sp. Cela signifie que le modèle n'a identifié aucune cellule de l'espèce. En ce qui concerne la précision, une très faible baisse a été enregistrée. Cela signifie que les modèles ne classent pas les autres cellules comme étant l'espèce recherchée .En conclusion, la méthode GMM reste toujours un puissant classificateur.

Maintenant, avec l'utilisation de la méthode proposée avec des données de test bruitées, nous atteignons une précision d'au moins 97% avec une diminution légère de la sensibilité. Dans ce cas, la sensibilité est comprise entre 67 à 100%, donc meilleure que l'approche classique. La faible sensibilité enregistrée par la méthode proposée est pour le cas de l'espèce Nodularia sp, mais reste aussi meilleure que la classification classique GMM. D'après ces résultats, nous confirmons que l'utilisation de la distance (modèle proposé) est plus appropriée que la probabilité d'appartenance pour la décision de classification.

Nous notons que dans notre étude, nous avons utilisé la discrimination des classes par la plus petite distance entre modèles GMM. Des approches récentes associent un modèle de règles qui peut séparer correctement les classes. Un exemple est l'utilisation de la machine à

vecteurs de support (SVM) pour donner une décision basée sur la distance entre les modèles GMM [157].

L'inconvénient majeur de l'approche proposée est le temps nécessaire pour l'apprentissage et la classification. L'utilisation d'un PC i3 de 4G de RAM pour la création de la base des modèles GMM prend presque une heure pour chaque espèce, avec une consommation importante de ressource mémoire et CPU. La phase de classification consomme moins de temps et elle dépend de la base d'apprentissage. Ces contraintes doivent être prises en considération pour l'étude des systèmes de discrimination en temps réel.

4.3.3. Modèle avec arbre de décision

Les résultats que nous avons obtenus, par le modèle d'arbre de décision qui est utilisé comme classificateur (méthode proposée), sont comparés avec les travaux de recherche de [118]. Notre choix s'est porté sur le classificateur utilisant l'arbre de décision, car les travaux de l'application de ce classificateur pour le traitement de données FC sont actuellement un créneau vierge à exploiter.

Nous présentons et nous discutons nos résultats pour la classification de la base de données FC de 20 espèces citée dans la section 4.3.1. Puis, nous allons développer un modèle de classification qui permet de classer ces 20 espèces.

Enfin, nous avons conçu 20 autres modèles pour discriminer chaque espèce d'algues parmi les 20 espèces. Il est démontré, dans plusieurs travaux de recherche, que la création d'un modèle adapté à l'identification d'une espèce spécifique à la foi serait plus efficace [158].

4.3.3.1. Création d'arbre de décision

Nous avons utilisé comme modèle de décision un arbre de classification binaire ordinaire (Ordinary Binary Cassification Trees OBCT) [143]. Le but de ce modèle OBCT est de classer toutes les espèces ou d'identifier une espèce spécifique dans un ensemble de données FC.

Le facteur d'évaluation de ces classificateurs est l'écart type (SD) entre la classification d'OBCT et la classification réelle. Ce choix a été fait pour qu'on puisse comparer nos résultats avec les travaux de A. Malkassian [118]. L'écart type sert à mesurer la

dispersion d'un ensemble de données. Plus il est faible, plus le vecteur de classes du modèle proposé et celui des classes réelles se ressemblent.

Pour le modèle de classification, nous avons utilisé la moitié des échantillons choisis de chaque espèce aléatoirement comme entrées d'apprentissage. Pour le test nous avons utilisé toute la base de données où les classes sont numérotées de 1 à 20. Les étapes à suivre sont les suivantes:

1. Choisir au hasard 50% des échantillons de chaque fichier FC et les mélanger pour créer un fichier de données d'apprentissage.
2. Créer un arbre OBCT en MATLAB avec les données d'apprentissage.
3. Tester l'arbre OBCT avec toutes les données.
4. Calculez l'écart standard (SD) entre le classement d'OBCT et le vrai classement.

Pour le modèle de discrimination, c'est-à-dire un modèle pour chaque espèce de phytoplancton, nous utilisons 50% des échantillons à partir d'un fichier FC de l'espèce spécifique et 50% des échantillons de toutes les espèces restantes ; tous choisis aléatoirement comme des entrées d'apprentissage. Pour le test, nous utilisons la base de données entière. Nous allons créer 20 modèles et dont chaque modèle est obtenu en suivant les étapes suivantes:

1. Prendre 50% des échantillons du fichier FC de l'espèce à identifier déjà sélectionnés dans l'étape 1 de modèle OBCT précédent.
2. Mélanger le reste des espèces dans un seul fichier et sélectionner 50% au hasard.
3. Créer une base d'apprentissage avec les échantillons des étapes 1 et 2.
4. Créer un arbre OBCT en MATLAB, avec les données d'apprentissage.
5. Tester l'arbre OBCT avec toutes les données.
6. Calculez l'écart standard (SD) entre le classement d'OBCT et le vrai classement.

4.3.3.2. Résultats et discussion

Les simulations sont réalisées avec MATLAB 8.3 R2014a en utilisant la boîte à outils des statistiques. Les résultats de l'utilisation du modèle OBCT et ceux de A. Malkassian [118] sont représentés au tableau 4.11.

Tableau 4.11. Les résultats de classification de 13332 échantillons de 20 espèces de phytoplanctons utilisant le modèle OBCT et ceux donnés par les travaux de [118].

Les espèces	Ecart type			
	Modèle OBCT		Modèle flou [118]	
	SD		SD	
	Modèle de classification	Modèle de discrimination	Paramatres FC	Paramatres FC et forme de signale
Anabaena cylindrical	0.2226	0.0286	0.396	0.383
Ankistrodesmus acicularis	0.0085	0.0076	0.017	0.240
Aphanizomenon sp	0.0349	0.0029	0.079	0.194
Chaetoceros muelleri	0.9751	0.1055	0.042	0.160
Chlorella sp	0.0096	0.0087	0.000	0.000
Ditylum brightwellii	0.6770	0.0498	0.105	0.069
Emiliania huxlyei	0.0733	0.0599	0.111	0.278
Gloeothece sp	0.5788	0.0111	0.080	0.248
Hemiselmis sp	0.0212	0.0145	0.000	0.325
Isochrysis sp	0.0473	0.0438	0.311	0.223
Melosira sp	0.4055	0.0118	0.429	0.452
Monoraphidium sp	0.0062	0.0354	0.000	0.000
Nodularia sp	0.7392	0.0664	0.131	0.138
Pavlova sp	0.3864	0.0194	0.000	0.000
Porphyridium sp	0.3023	0.0341	0.223	0.481
Pseudanabaena sp	0.0309	0.0081	0.000	0.238
Pediastrum sp	0.1256	0.0111	0.000	0.000
Rhodomonas sp	0.0746	0.0151	0.000	0.000
Skeletonema costatum	0.4115	0.0811	0.281	0.406
Thalassiosira pseudonana	0.2099	0.0065	0.281	0.000

En utilisant un modèle d'arbre de décision comme classificateur, nous pouvons atteindre un SD de l'ordre de 0.0085 en phase de test. Les résultats pour la plupart des espèces sont meilleurs que les travaux de A. Malkassian [118]. Du tableau 4.11, les meilleurs résultats du SD de la méthode d'arbre de décision proposé sont compris entre 0.0085 et 0.4055. Ces résultats montrent la puissance du modèle d'arbre de décision par rapport à la classification floue.

Cependant, la méthode de classification floue utilisée dans l'article [118] est plus stable : les valeurs de SD varient entre 0 et 0.429 pour le modèle qui utilise uniquement les paramètres FC et entre 0 et 0.452 pour celui utilisant les paramétrés FC et la forme de signal. Par contre, dans le modèle d'arbre de décision proposé, les valeurs de SD sont comprises entre 0.0085 et 0.9751: notre modèle favorise la classification des espèces bien discriminées. Il faut aussi noter que le modèle d'arbre de décision proposé nous a donné de mauvaises classifications pour les quatre espèces : Chaetoceros muelleri, Ditylum brightwellii, Gloeothece sp et Nodularia sp.

Nous notons ici que certaines décisions ne sont pas définitives, c'est-à-dire qu'il y a des probabilités dans la décision. Une feuille finale d'un arbre peut donner des différentes probabilités d'appartenance à des classes différentes. La décision d'appartenance est la classe de forte probabilité. Cela nous permet de connaître précisément le niveau de confusion dans l'arbre de décision.

L'utilisation des modèles de la discrimination de chaque espèce nous a conduits à de bons résultats pour toutes les espèces étudiées, tableau 4.11. La valeur du SD a diminué considérablement par rapport au modèle de classification. Les quatre espèces, dont les plus grandes valeurs de SD sont comprises entre 0.5788 et 0.9751 sont minimisées, en utilisant ces modèles de discrimination, respectivement à 0.0111 et 0.1055. Cette contribution nous a amené à des résultats très performants.

Avec ces modèles de discrimination d'arbre de décision, nous obtenons une très grande stabilité : les valeurs de SD des modèles de discrimination proposée varient maintenant entre 0.0029 et 0.1055. Ces résultats sont remarquablement meilleurs que ceux des travaux de A. Malkassian [118] dans le cas d'utilisation des paramètres FC et de la forme du signal.

Ces modèles de discrimination sont très utiles pour la recherche des cellules d'intérêt dans l'ensemble des échantillons analysés par le cytomètre en flux. L'inconvénient majeur de cette méthode de discrimination est qu'elle est dépendante de la base de données.

L'utilisation de toutes les variables (paramètres CF) est un grand avantage. L'utilisation d'autres variables liées aux espèces comme celle de la forme du signal qui est présentée dans l'article [118] peut aussi améliorer nos résultats et en particulier dans le cas des espèces de la même famille des phytoplanctons.

Le processus d'apprentissage prend entre 5 et 8 secondes sur un PC i3 de 4G RAM. Cependant, le modèle généré ne prend que quelques secondes pour classer toutes les données (plus que treize mille échantillons). On notera que pour un système hardware, le nombre de variables que nous avons utilisé (63 paramètres CF) utilisera beaucoup de ressources mémoire. Nous présentons ce modèle comme étant un classificateur en temps réel de données cytométrique [159].

4.4. Reconnaissance d'espèces phytoplanctoniques toxique

Nous allons présenter nos résultats pour la création d'un modèle de classification d'un ensemble de données cytométrique spécifique d'espèce Alexandrium Tamarens. Donc, notre objectif est la classification d'un ensemble de données issues du FLOWCAM de l'espèce Alexandrium par deux modèles supervisés : un réseau neuronal (RNA) et un arbre de décision (AD). Le meilleur modèle trouvé nous servira comme un système hardware de reconnaissance en temps réel.

4.4.1. Description des données

La base de données est une partie de la base utilisée par [110]. Les données proviennent du prototype FLOWCAM du département de biologie de l'Université Laval Canada. Ces données sont collectées durant l'année 2004 avec un FLOWCAM. C'est un instrument de la compagne Fluid Imgaging qui est composé d'un cytomètre en flux couplé à une caméra.

Les données qui nous intéressent sont celles issues de la partie cytométrique. Les paramètres de la partie cytométrique de FLOWCAM sont présentés dans la section 3.2.3 (voir tableau 3.2). Dans notre étude, les images générées par la caméra ne sont pas prises en considération.

Notre but est de réaliser un classificateur capable d'identifier les cellules toxiques de l'algue Alexandrium tamarense. Cette espèce a une forme très simple (cylindrique déformée). Il faut noter qu'il existe des espèces de la même famille qui soient très apparentées en forme et qui ne présentent pas la toxicité de l'espèce Alexandrium tamarense. À partir des paramètres cytométriques, nous verrons s'il y a une possibilité de discriminer les espèces apparentées de l'espèce toxique.

Alexandrium tamarense est une espèce marine d'algues unicellulaires. Certaines cellules peuvent produire des neurotoxines paralysantes. Les moules ou les myes peuvent accumuler ces toxines en se nourrissant de ces cellules. Des humains qui consommeraient des bivalves ainsi devenus toxiques risquent d'être victimes de troubles neurologiques plus ou moins graves (allant jusqu'à une paralysie pouvant entraîner la mort dans les cas les plus graves) en fonction de la dose de toxines qu'ils auront absorbée [160]

En plus des cellules viables, les cultures contiennent également une proportion de bruit comme : les particules inorganiques, des débris cellulaires... etc. Dans notre étude, plus que 3000 échantillons de cellules ont été utilisés.

La base de données est composée de 2000 cellules d'Alexandrium tamarense toxiques et de cellules de la famille Alexandrium qui ne sont pas toxiques. Ces données sont repartitionnées en deux classes «ALX» pour les cellules d'Alexandrium Tamarens et «NALX» pour les autres cellules. Les 80% des données, prises aléatoirement, sont destinés pour l'apprentissage et les 20% des données restantes pour les tests. Pour chaque répartition de données, les données ont été normalisées autour de leur écart type.

Pour la validation de nos modèles (AD et MLP), on a utilisé une autre base de données constituée de 1000 échantillons (Alexandrium Tamarens et famille Alexandrium).

4.4.2. Résultats et discussion

Le réseau MLP utilisé dans ce travail est celui proposé par les travaux de thèse de [110]. Le nombre optimum de neurones dans les différentes couches et le nombre de couches cachées sont obtenus avec l'approche d'essai-erreur [161,162]. Plusieurs architectures ont été testées afin de déterminer les réseaux qui ont un bon compromis entre la taille et le taux d'erreur.

L'utilisation de plusieurs architectures nous a conduits à un modèle de réseau MLP de deux couches cachées qui contiennent respectivement 20 et 8 neurones. Une couche d'entrée de 11 neurones et une couche de sortie d'un seul neurone (classification ALX ou NALX). Cette architecture (12-20-8-1) nous a permis d'atteindre des erreurs de classification minimes lors de l'apprentissage et le test (les travaux de thèse [110]).

Tableau 4.12. Pourcentage de reconnaissance pour Arbre OBCT et Réseau MLP pour chaque phase.

Méthode	**Pourcentage de reconnaissance**	
	Phase apprentissage et test	***Phase validation***
Réseau MLP	93.40 %	57 %
Arbre OBCT	96.36 %	30 %

L'utilisation du modèle neuronal MLP, dont les résultats sont représentés au tableau 4.12, donne de bons taux de reconnaissance et de test. Le modèle MLP généré arrive à un taux de l'ordre de 93% de reconnaissance, tableau 4.12. Les cas de rejet sont tous dus aux cellules des espèces d'Alexandrium non toxique (Cellules identiques en forme et en paramètres de fluorescence).

L'utilisation du modèle d'arbre de décision, représenté sur la figure 4.12, nous a permis d'atteindre un taux de 96.3% de reconnaissance dans les phases d'apprentissage et de test. L'avantage de l'utilisation des arbres de décision est la réduction du nombre de variables utilisées. L'arbre n'utilise que quelques variables pour arriver rapidement à séparer et classifier les cellules toxiques est non toxique. On notera aussi que sur la figure 4.12, pour certains cas, les classes ne sont pas définitives (feuille terminale avec des probabilités par exemple 0.6 et 0.5) c'est-à-dire avec une probabilité d'appartenance à une classe autour de 0.5 (Erreur pour chaque décision prise). Pour d'autres cas, une décision finale d'appartenance est prise (feuille terminale avec probabilité 1 ou 0).

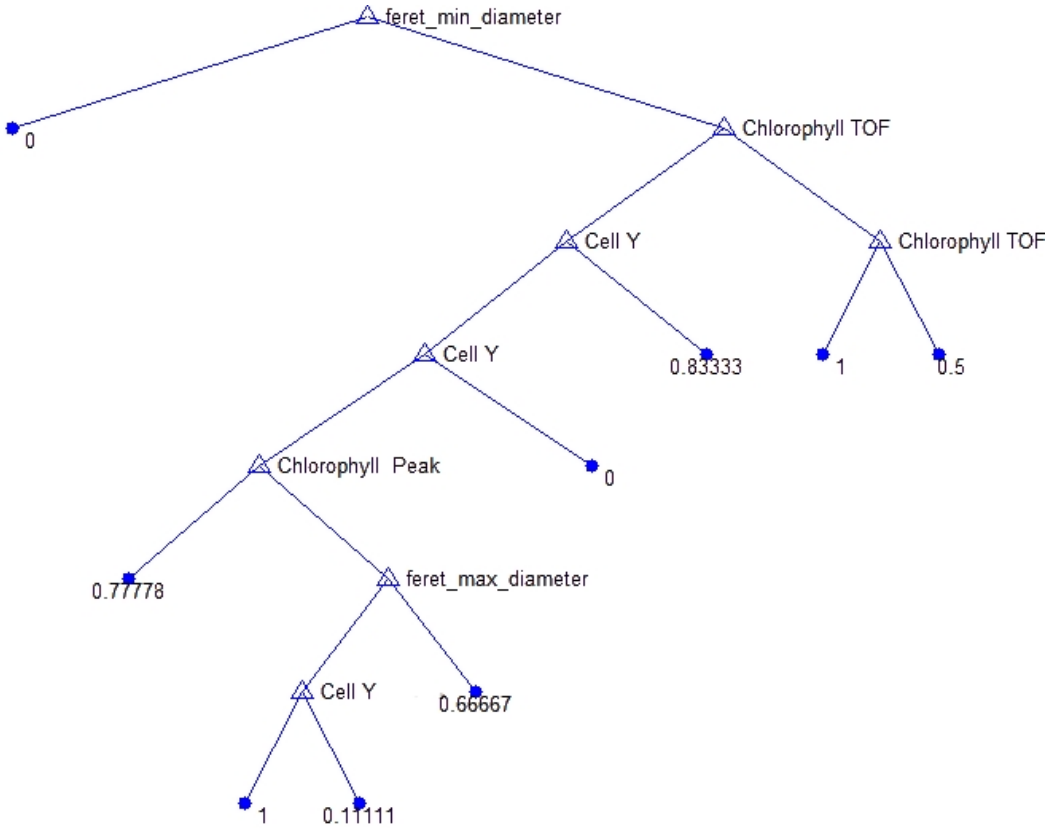

Figure 4.12 L'arbre de décision proposé pour la reconnaissance d'Alexandrium Tamarens.

L'avantage de réseau de neurones MLP est sa prédiction. Dans la phase de validation de modèle MLP, on arrive à un taux de reconnaissance de 57%, tandis que l'arbre de décision ne dépasse pas les 30% de reconnaissance tableau 4.12. Cela s'explique par le fait que nous avons utilisé des données qui ne figurent pas dans les bases d'apprentissage et de test.

Les cellules qui ne figurent pas dans la base d'apprentissage ne peuvent pas affecter le modèle MLP parce qu'il a utilisé un nombre suffisant de variables pour la discrimination (la totalité des attributs). Ces nouvelles cellules sont directement classées comme Alexandrium Tamarens dans le cas des arbres de décision ; qu'elles soient toxiques ou non.

Donc pour que le modèle d'arbre de décision proposé puisse être utilisé comme un système d'alerte authentique sur site ; il faut que la base de données d'apprentissage soit suffisamment grande avec une mise à jour périodique.

4.5. Conclusion

Ce chapitre présente nos résultats de recherche à propos de deux thèmes liés à la télésurveillance de la pollution biologique aquatique.

Le premier thème est celui de la prédiction des blooms à partir des données physiques et chimiques de milieu aquatique. Notre but était la création d'un système de prédiction à moindre coût (capteurs) et ressource (données).

Présentée au chapitre 2, une modélisation de la prédiction des blooms reste très compliquée et dépend essentiellement du choix des variables, des méthodes à suivre (les modèles) et de la durée de prédiction. Le choix des variables reste très difficile et cela est dû au grand espace de variables (mesures) lié au phénomène et aux coûts élevés des capteurs.

Finalement, on a abouti à des mesures qui sont liées au processus et qui sont obtenues facilement avec des sondes spécifiques à faible coût. Avec ces variables on a pu construire plusieurs architectures de réseaux de neurones MLP qui donnent de très bons résultats en suivie et en prédiction à court terme.

Pour les architectures étudiées à une seule couche cachée, les phases d'apprentissages sont très bonnes, mais pour les phases de test et de prédiction les résultats ne sont pas très performants. Cependant, les résultats de l'architecture à deux couches cachées que nous avons proposée sont meilleurs et sûrs lors des 3 phases. Les contraintes de cette architecture résident dans l'implémentation du modèle sur un support hardware. Les causes sont la taille du réseau ainsi que le temps de calcul qui est un peu élevé par rapport aux architectures à une seule couche cachée.

Le deuxième thème est la reconnaissance automatique d'algues. Ce travail apporté est complémentaire aux travaux de modélisation traitée dans le premier thème.

Nous avons représenté nos premiers résultats de l'utilisation d'une nouvelle approche d'analyse des données FC. Puis, nous avons proposé une présentation matricielle des données FC pour chaque cellule, au lieu de sa forme originale qui était un vecteur. Nous avons testé une approche déjà appliquée dans la classification de l'information multimédia qui utilise une approche GMM modifiée. Cette méthode classe les données en fonction de la distance entre les GMM des références (classes) et des échantillons modélisés par GMM. Les premiers résultats étaient prometteurs ; la résistance d'approche proposée au bruit été fort. Cette

propriété rend la méthode proposée plus appropriée pour l'utilisation au suivi *in situ*. Nous devrons augmenter la discrimination entre les classes en testant d'autres distances et d'autres outils de discrimination comme la machine à vecteurs de support (SVM) [157].

Nous proposons l'arbre de décision de type OBCT pour classer les 20 espèces de phytoplanctons. Les résultats obtenus sont comparés à d'autres résultats obtenus avec un travail déjà réalisé de A. Malkassian [118]. Les résultats sont favorables pour le modèle OBCT proposé [159].

Idéalement, le modèle OBCT devrait être formé et testé sur des données obtenues à partir du site ou mesurées en ligne. Pendant le processus d'analyse, les particules d'intérêt (espèces toxiques) peuvent être détectées assez rapidement en utilisant des modèles de discrimination automatisés. L'amélioration de notre travail est d'ajouter d'autres paramètres CF qui seront calculés à partir des paramètres CF initiaux.

Pratiquement, nous devons tout d'abord nous concentrer sur les espèces les plus toxiques ou les plus représentatives de leur environnement. Par conséquent, il serait préférable de viser directement l'espèce d'intérêt. Nous avons proposé un modèle de classification OBCT. Les résultats obtenus sont comparés avec ceux du modèle RNA de type MLP déjà réalisé [110]. Les résultats favorisent le modèle OBCT dans la phase d'apprentissage, mais il est loin d'être utile sur n'importe quel autre site. Le modèle MLP peut s'adapter aux nouvelles particules rencontrées sur site.

Les résultats trouvés sont encourageants et nous incitent à compléter notre recherche ; afin de créer un classificateur qui est en mesure de trouver des cellules et des espèces d'intérêt. Enfin, essayer de réaliser un système de discrimination et classification en temps réel avec la mise en œuvre (logiciels et implémentation matérielle) de ces méthodes proposées.

Conclusion Générale et Perspectives

Conclusion Générale et Perspectives

A l'échelle mondiale, les zones aquatiques comptent parmi les milieux où la pression humaine a le plus fortement augmenté depuis trois siècles. Les plus grandes villes du monde se trouvent en bord de mer et la moitié de la population des pays industrialisés vit à moins d'un kilomètre des rivages marins. Il s'avère donc indispensable de surveiller ces milieux pour préserver ces environnements uniques de par leurs richesses et leurs spécificités.

Cette thèse fait partie du suivi de ces milieux par la contribution à la surveillance de la pollution biologique aquatique. La pollution biologique ciblée dans cette thèse est la contamination de milieux aquatiques par les phytoplanctons.

Les phytoplanctons sont d'une grande diversité, dont la dynamique reste encore mal connue à cause des méthodes d'observations utilisées actuellement. De plus, le phytoplancton avec son potentiel d'adaptation présente un bon bio-indicateur de l'état du milieu aquatique. Cependant, cette propriété peut causer des phénomènes nocifs de pollution biologique comme les blooms et les phycotoxines. Les études sur les outils de la surveillance automatique de ces microorganismes sont donc d'une importance capitale.

La première partie de cette thèse traite le problème des blooms d'algues toxiques. Ces phénomènes émergent dans le monde entier et touchent principalement les ressources d'eaux potables et les zones côtières. Leurs effets sur la vie humaine, surtout dans le cas des phytoplanctons toxiques, rendent le développement d'un système de surveillance un créneau porteur de recherche.

La modélisation de bloom, à partir de données mesurables, aide à prédire et à comprendre la dynamique de développement de ces microorganismes. Ces données peuvent être liées au phytoplancton lui-même, à l'environnement ou au deux en même temps.

Les analyses ponctuelles par microscopie offrent le plus de précision pour l'identification de la composition spécifique et la reconnaissance des espèces responsables des blooms. Mais dans ce genre de traitement, le temps d'analyse est trop long et les manipulations exigent un personnel qualifié.

Dans le cas de notre étude, la modélisation utilise les données environnementales. Ce choix a été fait pour la simplicité des mesures physico-chimiques, le coût réduit des instruments utilisés et la disponibilité d'une base de données à court terme. A partir des données mesurées, un modèle de réseaux de neurones MLP a été conçu afin de prédire

l'évolution de la concentration des cellules phytoplanctoniques de l'espèce toxique D. Acuminata.

La technique de modélisation utilisant les réseaux de neurones demeure une voie de recherche en plein développement. Actuellement, il n'existe aucune théorie permettant d'évaluer la configuration optimale d'un réseau indépendamment de l'application envisagée.

Les résultats obtenus de la prédiction à court terme de 10 jours, à partir d'apprentissage de modèle MLP avec une base de données de deux mois, sont satisfaisants. Un modèle de prédiction à long terme est envisagé en utilisant une base de données plus importante avec plus de paramètres mesurés. Ce modèle sera implémenté sur circuit programmable de type FPGA. Une proposition théorique de la réalisation d'un tel système a été déjà publiée [148].

Malgré les avantages de ce modèle, le problème qui se pose encore est celui du taux de reconnaissance qui dépend de l'espèce d'algue. Il existe plusieurs espèces qui prolifèrent dans les mêmes conditions. Donc l'identification de l'espèce à étudier est une question importante. La confirmation de la décision d'alerte générée par notre système par les analyses manuelles (la présence d'algue surveillée) va nous prendre beaucoup de temps, c'est-à-dire la décision est très lente et dépend du personnel expertisé.

Dans cette thèse nous avons proposé un autre système de reconnaissance automatique complémentaire au premier. Ce système va permettre d'identifier les espèces phytoplanctoniques ciblées dans les échantillons d'eau.

L'analyse des données générées par le CytoSense ou FLOWCAM permet d'obtenir les principales informations de la morphologie des cellules et de traiter une quantité suffisante de données pour analyser ces phytoplanctons à une fréquence plus élevée. Ces informations permettent ainsi de discriminer les groupes (espèces) phytoplanctoniques sur la base de la fluorescence des pigments, de la diffusion de la lumière et des images.

Le temps d'analyse avec ces appareils dépend de la densité des cellules présentes dans l'échantillon, mais ils n'excèdent pas en général 15 min. Ces types de techniques peuvent opérer *in situ* dans le cadre de suivis à haute fréquence. Le temps d'analyse est d'une heure pour le traitement des données, ce qui reste 3 fois inférieur au temps nécessaire à l'analyse par microscopie.

Les compagnies CytoBuoy et fluidimaging ont développé des instruments capables de fournir le profil des signaux optiques de fluorescence et des images des phytoplanctons. L'exploitation de tels profils par le biais de méthodes de machine d'apprentissage a été engagée au cours de ce travail. Nous avons fourni nos résultats prometteurs dans la

reconnaissance et la discrimination d'espèces. Ces méthodes reposent sur l'apprentissage d'un modèle de reconnaissance par des exemples des cellules possédant les mêmes propriétés optiques, pour l'identification ou la classification.

Nous avons présenté tous les résultats dans le chapitre 4 dans l'ordre chronologique inverse de la réalisation. Un modèle de classification GMM a été réalisé, ensuite une étude comparative entre les travaux de recherche de A. Malkassian [118] et une proposition d'un modèle d'arbre de décision est donnée. Enfin, les résultats de la classification d'espèces toxiques Alexandrium tamarense sont comparés avec ceux des travaux de recherche de [110].

Dans le cas du modèle GMM proposé, une représentation matricielle des données cytométriques a été exposée. Il est difficile de comparer de façon appropriée les résultats obtenus dans cette recherche avec d'autres travaux qui utilisent des représentations simples des données CF. Les résultats obtenus montrent la grande résistance du modèle proposé au bruit. Cette propriété présente un avantage lors des analyses *in situ*, où le bruit est important à cause des cellules déformées et des déchets microscopiques. Ces débris sont généralement filtrés au laboratoire avant de procéder à des analyses avec le cytomètre en flux.

Le problème rencontré avec ce type de modèle est la quantité de calculs nécessaires pour créer les GMM des cellules et les classer par rapport aux GMM des espèces. A cet effet, cette technique ne peut pas s'appliquer dans sa forme actuelle pour la reconnaissance en temps réel.

Le deuxième modèle proposé est basé sur les arbres de décision. Les avantages d'utilisation de ces classificateurs sont la simplicité de la méthode, la possibilité d'apprendre à reconnaître la forme des signaux et l'extraction des différentes règles qui séparent les classes. Nous avons obtenu de bons taux de reconnaissance de plusieurs espèces à la fois. L'utilisation des modèles de discrimination, qui consiste à cibler une espèce à chaque fois, a amélioré nos résultats précédents. La rapidité des tâches d'apprentissage et de classification est le grand avantage de cette méthode. Les limites de cette méthode proposée (arbre de décision) sont la non-flexibilité pour les nouveaux cas. Le modèle dépend de la base d'apprentissage et nécessite alors un grand nombre de données d'apprentissage.

Le troisième travail abordé est le problème de la discrimination des espèces toxiques et non toxiques de la même famille d'algues. Premièrement, il y a plusieurs difficultés à étendre la méthodologie de classification automatique aux cultures dans le laboratoire à des populations hétérogènes dans la mer. Le nombre des classes dans tous les échantillons d'eau de mer naturelle est inconnu. Deuxièmement, si on connait d'avance le nombre de classes

(classification supervisée) dans un prélèvement, plusieurs cas de phytoplanctons de la même famille sont proches en forme et en structure interne. Le modèle doit donc viser la discrimination et non la classification de l'espèce toxique parmi plusieurs cellules et surtout ceux de la même famille.

Le modèle proposé d'arbre de décision donne de bons résultats pour la discrimination de l'espèce toxique Alexandrium tamarense pour des données d'apprentissage et de test connues. Pour la validation avec d'autres données, hors la base d'apprentissage et de test ; le modèle d'arbre de décision se présente comme un mauvais classificateur par rapport au modèle MLP. Alors, il serait préférable d'élargir la base de données et de créer une bibliothèque de données cytométriques pour toutes les sous-espèces d'Alexandrium et d'y associer des modèles de classification pour pouvoir détecter et différencier toutes les sous-espèces.

En perspective, nous proposons :

- La réalisation d'un système de reconnaissance sur circuit électronique semble prometteuse surtout pour l'analyse *in situ*, où le système de surveillance prend en charge automatiquement la prédiction et l'alerte.
- Des études plus approfondies seront également nécessaires sur les instruments (CytoSub, FLOWCAM), l'extraction des paramètres du traitement (signaux CF et images), l'optimisation des données et les capteurs afin de réduire le coût des calculs.
- Le développement du modèle GMM proposé semble une idée très prometteuse surtout avec l'association d'autres modèles de discriminations comme les machines à support de vecteur ou les arbres de décision. Cela nous poussera à développer d'autres logiciels plus performants de classification aux laboratoires.
- Dans notre recherche biotechnologique, nous visons aussi à l'application de ces modèles que nous avons développés (logiciels) pour l'amélioration de la reconnaissance automatique de cellules vivantes. En perspective, nous sommes à la recherche d'un cytomètre en flux en Algérie pour pouvoir appliquer nos travaux de recherche dans le domaine de la médecine pour la détection des bactéries et de cellules cancéreuses.

Références

Références

[1] Ménesguen A, Aminot A, Belin C. L'eutrophisation des eaux marines et saumâtres en Europe, en particulier en France. Rapport IFREMER DEL/EC/01.02 pour la Commission Européenne – DG.ENV.B1. Ifremer, Brest, 2001.

[2] Smith VH, Tilman GD, Nekola JC. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environmental Pollution. 100, 179-196, 1999.

[3] Estrada M. Persistence and variability of phytoplankton assemblages. Implications for monitoring. Bio-ecological Observations in Operational Oceanography. EuroGOOS Publication, Southampton, pp 14-15, 2000.

[4] Smayda TJ. Novel and nuisance phytoplankton blooms in the sea: Evidence for a global epidemic. Toxic marine phytoplankton: Proceedings of the 4th International Conference on Toxic Marine Phytoplankton. Elsevier, pp. 29-40, 1990.

[5] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities EN, vol. 2000/60/EC, 2000.

[6] http://www.cdph.ca.gov/healthinfo/environhealth/water/Pages/Shellfish.aspx

[7] http://www.appl.dz/spip.php?article311

[8] Aminot A, Kérouel R. Hydrologie des écosystèmes marins. Paramètres et analyses. Ifremer, pp 336, 2004.

[9] Mantoura RFC, Llewellyn CA. The rapid determination of algal chlorophyll and carotenoid pigments and their breakdown products in natural waters by reverse-phase high-performance liquid chromatography. Analytica Chimica Acta. 151, pp 297-314, 1983.

[10] Jeffrey SW, Mantoura RFC, Wright SW. Phytoplankton pigments in oceanography. Unesco, Paris, pp 661, 1997.

[11] Utermöhl H. Zur Vervollkommnung der quantitativen Phytoplankton Methodik. Mitteilungen Internationale Vereinigung für Theoretische und Angewandte Limnologie. 9, pp 1-39, 1958.

[12] Shapiro HM. Practical flow cytometry. 4th ed. Wiley-Liss, Hoboken, New Jersey, pp736, 2003.

[13] Green R. E, Sosik H. M, Olson, R. J., DuRand, M. D. Flow cytometric determination of size and complex refractive index for marine particles: Comparison with independent and bulk estimates, Applied Optics, 42:3, pp 526–541, 2003.

[14] Katsugari T, Tani Y. Screening for microorganisms with specific characteristics by flow cytometry and single-cell sorting, Journal of Bioscience and Bioengineering, 89:3, pp217–222, 2000.

[15] http://portail.umons.ac.be/FR/universite/facultes/fs/services/institut_bio/ecologie_numerique_milieux_aquatiques/Pages/default.aspx

[16] Cloern JE. Our evolving conceptual model of the coastal autrophication problem. Marine Ecology Progress Series. 210, pp223-253, 2001.

[17] Furnas MJ. Net in situ growth rates of phytoplankton in an oligotrophic, tropical shelf ecosystem. Limnology and Oceanography. 36, pp13-29, 1991.

[18] Smith VH, Tilman GD, Nekola JC. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environmental Pollution. 100:1, pp79-196, 1999.

[19] KALFF J. Limnology. Prentice-Hall. pp 592, 2002.

[20] PITOIS S, JACKSON M et WOOD B. Sources of the eutrophication problems associated with toxic algae: An overview. Journal of Environmental Health, 64, pp 25-32, 2001.

[21] WETZEL R.G. et LIKENS G. E. Limnological Analyses 3rd edition Springer-Verlag. pp 429, 2000.

[22] SKULBERG O.M., CODD G. A. et CARMICHAEL W.W.. Toxic blue-green algal blooms in Europe: A growing problem. Ambio, 13, pp 244-247, 1984.

[23] SMITH V. Eutrophication of freshwater and coastal marine ecosystems - A global problem. Environmental Science and Pollution Research, 10, 126-139, 2003.

[24] Wyatt T, Horwood J. Model which generates red tides Nature, 244, 238–240, 1973.

[25] Truscott JE, Brindley J. Ocean plankton populations as excitable media. Bull Math Biol, 56, 981–998, 1994.

[26] Ebenhoh W, Kohlmier C, Radford PJ. The benthic biological submodel in the European regional Seas Ecosystem Model. Neth J Sea Res, 33, 423–452, 1995.

[27] Grover JP, Crane KW, Baker JW, Brooks BW, Roelke DL. Spatial variation of harmful algae and their toxins in flowing water habitats: a theoretical exploration. J Plank Res, 33, 211–227, 2011.

[28] Banerjee M, Venturino E. A phytoplankton-toxic phytoplankton-zooplankton model. Ecol Complex, 8, pp 239–248, 2011.

[29] Subhendu Ch, Ulrike Feudel Harmful algal blooms: combining excitability and competition, Theor Ecol 7:pp 221–237, 2014.

[30] Kim J. K, Youn H. J, "HABs Prediction method in Yeosu bay using Remote Sensing", In proceeding of spring Conference of Korean Society GIS, pp.47-57, 2001.

[31] Li Y and Smayda T, "Heterosigma akashiwo (Raphidophyceae): On prediction of the week of bloom initiation and maximum during the initial pulse of its bimodal bloom cycle in Narragansett Bay", Plankto Biol. ecol., 47 : 2, pp. 80-84, 2000.

[32] Song B. H., Jung M. A., Lee S. R., "A Design and Implementation Red Tide Prediction Monitoring System using Case Based Reasoning", Journal of Korea Information and Communications Society, 35:12, pp.1819-1826, 2010.

[33] Fdez-Riverola F., Corchado J. M., "FSfRT: Forcasting System for Red Tides", Applied Intlligence, 21, pp.251-264, 2004.

[34] Rong Z., Hong Y., Liping D., "Research on Prediction of Red Tide Based on Fuzzy Neural Network", Marine Science Bulletin, 8:1, 2006.

[35] Zaiwen L, Xiaoyi W, Lifeng C, Xiaofeng L, Jiping X. Intelligent technology for predicting water bloom engendering, Industrial Electronics. IECON 2008. 34th Annual Conference of IEEE, pp1896-1900, 2008.

[36] Teles L. O, Pereira E, Saker M, Vasconcelos V. Time Series Forecasting of Cyanobacteria Blooms in the Crestuma Reservoir (Douro River, Portugal) Using Artificial Neural Networks. Environmental Management, 38:2, pp 227–237, 2006.

[37] Xiaoyi W, Zaiwen L, Shiping Z, Jun D, Chenling Z, Minghua Y. Research on One Intelligent Prediction Method for Water Bloom. Dependable, Autonomic and Secure Computing, 2009. DASC '09. Eighth IEEE International Conference on, pp 682-685, 2009.

[38] Xiaoyi W, Jun D, Zaiwen L, Xiaoping Z, Dong S, Zhiyao Z, Zhang M. The lake water bloom intelligent prediction method and water quality remote monitoring system, Natural Computation (ICNC), 2010 Sixth International Conference on, 7, pp 3443-3446, 2010.

[39] Shiping Z, Zaiwen L, Xiaoyi W, Jun D. Water-Bloom Medium-Term Prediction Based on Gray-BP Neural Network Method, Dependable, Autonomic and Secure Computing. DASC '09. Eighth IEEE International Conference on, 2009, pp 673-676, 2009.

[40] Sohyun C, Byungjin L, Jaewoon J, Sangdon K, Hyunmi C, Jonghwan P, Seoksoon P, Jae K.P. Factors affecting algal blooms in a man-made lake and prediction using an artificial neural network, Measurement, 53, pp 224–233, 2014.

[41] Zaiwen L, Lifeng C, Xiaoyi W, Siying L. The Method of Soft Sensing for Water Bloom in River and Lakes Based on RBF Neural Network, Control Conference, 2007 China., pp108-111, 2007.

[42] Siying L, Zaiwen L, Xiaoyi W, Lifeng C. Short-term predicting model for water bloom based on Elman neural network. Control Conference 2008, pp 218 – 221, 2008.

[43] Huajun L, Defu L, Yingping H. Artificial neural network modeling of algal bloom in Xiangxi Bay of Three Gorges Reservoir, Intelligent Control and Information Processing (ICICIP), 2010 International Conference on, pp 645 – 647, 2010.

[44] Xiu L, Jin Y, Zhuo J, Jingdong S. Harmful algal blooms prediction with machine learning models in Tolo Harbour, Smart Computing (SMARTCOMP) International Conference on, 2014 , pp 245 – 250, 2014.

[45] Kuo J, Hsieh M, Lung W, She N. Using artificial neural network for reservoir eutriphication prediction. Ecol. Model., 200, pp 171–177, 2007.

[46] Yong X, Changchun C, Zhang Y, Dong Z. Identification of algal blooms based on support vector machine classification in Haizhou Bay, East China Sea, Environ Earth Sci, 71, pp 475–482, 2014.

[47] Zaiwen L, Xiaoyi W, Lifeng C, Xiaofeng L, Jiping X. Research on Water Bloom Prediction Based on Least Squares Support Vector Machine. Computer Science and Information Engineering, 2009 WRI World Congress on, 5 , pp764 - 768, 2009.

[48] Peretyatko A, Teissier S, De Backer S, Triest L. Classification trees as a tool for predicting cyanobacterial blooms, Hydrobiologia, 689, pp131–146,2012.

[49] Déath G. and Fabricius K. E.. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology , 81, pp 3178–3192, 2000.

[50] Peretyatko A, Teissier S, De Backer S and Triest L. Assessment of the risk of cyanobacterial bloom occurrence in urban ponds: probabilistic approach. Annales De Limnologie (International Journal of Limnology) ,46, pp121–133, 2010.

[51] Park S, A Jung Mn, Seong R. L, Se Jun P, Jae H.P, Kong S.K. Yinsoo Park. Prediction of red tide blooms using decision tree model. ICT Convergence (ICTC), 2011 International Conference on, pp 710 – 713, 2011.

[52] Laanemets J, Lilover M.J, Raudsepp U, Autio R, Vahtera E, Lips I and Lips U. A fuzzy logic model to describe the cyanobacteria Nodularia spumigena blooms in the Gulf of Finland, Baltic Sea, Hydrobiologia, 554, pp 31–45, 2006.

[53] Park S, Jong G.J, Jangwoo K, Seong R.L. Red tides prediction using fuzzy inference and decision tree. ICT Convergence (ICTC), 2012 International Conference on. pp 493 – 498, 2012.

[54] Jong-Kuk C, Jee-Eun M, Jae H.N, Tai-Hyun H, Suk Y, Young J.P, Jeong-Eon Moon, Jae-Hyun Ahn, Sung Min Ahn, Jae-Hun Park. Harmful algal bloom (HAB) in the East Sea identified by the Geostationary Ocean Color Imager (GOCI). Harmful Algae, 39,pp 295–302,2014.

[55] Bing Z, Qian S, Junsheng L. Monitoring water quality of urban water supply sources using optical remote sensing. Urban Remote Sensing Event, 2009 Joint, pp1–5, 2009.

[56] Wei Q, Jiang N, Lu H, Hu B. A System for Dynamically Monitoring and Warning Algae Blooms in Taihu Lake Based on Remote Sensing. Image and Signal Processing, 2009, 2nd International Congress on, pp1-5, 2009.

[57] Pettersson L.H., Durand D., Johannessen O.M., Svendsen, E., Noji T., Soiland, H, Groom, S., Regner, P. Monitoring and model predictions of harmful algae blooms in Norwegian waters, Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International, 3, pp1146 - 1148, 2001.

[58] Chen Q, Mynett A.E. Predicting phaeocystis globosa bloom in Dutch coastal waters by decision trees and non-linear piecewise regression. Ecological Modeling, 176 pp 277-290, 2004.

[59] Rousseeuw K, Caillault E.P, Lefebvre A, Hamad D. Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling, Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International, pp 3962 – 3965, 2013.

[60] Zijian W, Zhao Z, Dong L, Li C. Data-Driven Soft Sensor Modeling for Algal Blooms Monitoring. Sensors Journal, IEEE, 15:1, pp 579 – 590, 2015.

[61] Muttil N and Chau K. Neural network and genetic programming for modelling coastal algal blooms. International Journal of Environment and Pollution, 28, pp223-238, 2006.

[62] Hu J, Ji P and Zhang C. Prediction Model for Red Tide at Yantai Sishili Bay Based on LMBP Algorithm. Journal of System Simulation, 19, pp 60-68, 2009.

[63] Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model., 90, pp 39–52, 1996.

[64] Niu Z.G, Zhang H. W, Liu H. B. Application of neural network to prediction of coastal water quality. J. Tianjin Polytechnic Univ., 25, pp. 89–92, 2006.

[65] Shu J. Using neural network model to predict water quality. North Environ., 31, pp. 44–46, 2006.

[66] http://en.wikipedia.org/wiki/Dinophysis_acuminata

[67] Díaz P, Reguera B, Ruiz-Villarreal M, Pazos Y, Velo-Suárez L, Berger H, Sourisseau M. Climate variability and oceanographic settings associated with interannual variability in the initiation of Dinophysis acuminata blooms. Marine Drugs , 11:8, pp 2964–2981, 2013.

[68] Ka Jeong L, Mok Jong S, Song K.C, Hongsik Y, Jung J.H, Kim J. H. Geographical and annual variation in lipophilic shellfish toxins from oysters and mussels along the south coast of Korea. Journal of Food Protection, 74:12, pp 2127–2133, 2011.

[69] Naustvoll L.J, Gustad E, Dahl E. Monitoring of Dinophysis species and diarrhetic shellfish toxins in Flødevigen Bay, Norway: inter-annual variability over a 25-year time-series. Food Additives & Contaminants: Part A, 29 :10, pp1605–1615, 2012.

[70] Myung G.P, Sunju K, Hyung S.K, Geumog M, Kang Y.G, Wonho Y. First successful culture of the marine dinoflagellate Dinophysis acuminata, aquatic microbial ecology, 45, 101–106, 2006.

[71] Barre N. Le phytoplancton toxique des lagunes méditerranéennes : de la surveillance à la recherche, Conseil Scientifique et Technique du Pôle-relais lagunes méditerranéennes, pp 5, 2005.

[72] Bougis P. Ecologie du plancton marin. I. Le phytoplankton. Masson et Cie, Paris: pp 196, 1974.

[73] Paulmier G et OOLY J.P. Manifestation de Dinophysis Acuminata sur le littoral normand. 47, pp.149-157, 1985.

[74] Smayda T.J. Eutrophication and phytoplankton. In Drainage Basin Nutrient Inputs and Eutrophication: an Integrated Approach. University of Tromsø, Norway, pp. 89–98, 2004.

[75] LASSUS P, MARTIN A.G, BERTHOME J.P, LANGLADE A, BACHERE E et MAGGI P. Extension du D. Acuminata en Bretagne sud et conséquence pour les cultures marines. Rev. Trav. Inst. Pêche maritime, 47, pp 122-133, 1985.

[76] Vollenweider RA. Coastal marine eutrophication: principles and control. Marine coastal eutrophication. Elsevier, London, pp 1–20, 1992.

[77] Martin Y, Joncour P, Saliba P, Tanguy B et Descatoire J. Fixation biologique du C02 en milieu marin. Contrat Elf n°8398, avenant n°4, 1996.

[78] Geider R.J, MacIntyre H.L, Kana T.M. A dynamic regulatory model of phytoplanktonic acclimatation to light, nutrients and temperature. Limnol. Oceanogr., 43 :4, 679-694, 1998.

[79] Shuter B. A model of physiological adaptation in unicellular algae. J. Theor.Biol., 78: 519-552. 1979.

[80] http://www.cetsm.eu/var/storage/images/medias-ifremer/medias-cetsm/illustrations/illustration_lot-e_sonde-ctd/815778-1-fre-FR/Illustration_Lot-E_Sonde-CTD.jpg

[81] Nathalie D. Mesures en océanographie physique. Article Techniques de l'ingénieur, R2340, 1996.

[82] http://www.ioos.noaa.gov/images/gliders_combined620.jpg

[83] http://cobs.pol.ac.uk/cobs/gliders/images/glider_schematic_is.jpg

[84] Koroleff F. Direct determination of ammonia in natural waters as indophenol blue. International Council for the exploration of the sea. C.M. 1969/C:9 Hydr. Comm, 1969.

[85] Bendschneider K, Robinson RJ. A new spectrophotometric determination of nitrate in seawater. Journal of Marine Research. 11, pp 87-96, 1952.

[86] Woods ED, Amstrong FAJ, Richards FA. Determination of nitrate in sea water by cadmiumcopper reduction to nitrite. Journal of Marine Biological Association of United Kingdom. 47, pp 23-31, 1967.

[87] Murphy J, Riley JP. A modified single solution method for the determination of phosphate in natural waters. Analytica Chimica Acta. 12 , pp 31-36, 1962.

[88] http://www.spuvvn.edu/upload/medialibrary/home_science/Autoanalyser.JPG

[89] Debliquy M. Capteurs chimiques. Article Techniques de l'ingénieur, R420, pp 3-9, 2010.

[90] http://www.nico2000.net/datasheets/elgrplrg.html

[91] http://fr.wikiversity.org/wiki/Capteur/D%C3%A9tection_et_mesure_d%27%C3%A9l%C3%A9ments_ou_de_mol%C3%A9cules#L.27amp.C3.A9rom.C3.A9trie

[92] http://www.bravofrance.fr/fr/p/glacier-artisanal-methodologies-de-production

[93] Sugimura Y, Suzuki Y. A high temperature catalytic oxidation method for the determination of non-volatile dissolved organic carbon in sea water by direct injection of a liquid sample. Marine Chemistry. 24, pp105-131, 1988.

[94] Cauwet G. HTCO method for dissolved organic carbon analysis in sea water: influence of catalyst on blank estimation. Marine Chemistry. 47, pp 55-64, 1994.

[95] Maier H.G, and Dandy G.C. Neural networks for the prediction and forecasting of water resources variables, a review of modelling issues and applications. Environmental Modelling and Software, 15, pp101–124, 2000.

[96] Maier, H. G., and G. C. Dandy. "Neural network based modelling of environmental variables", a systematic approach. Mathematical and Computer Modelling, 33, pp 669–682. 2001.

[97] Parizeau M. Réseaux de neurons. Université Laval, 2004.

[98] Rosenblatt F. The Perceptron : probabilistic model for information storage and organization in the brain. Psychological Review, 65, pp 386-408, 1958.

[99] https://en.wikipedia.org/wiki/Flow_cytometry

[100] Green R.E, Sosik H.M, Olson R. J, and DuRand M.D. Flow cytometric determination of size and complex refractive index for marine particles: Comparison with independent and bulk estimates. Applied Optics, 42, 526–541, 2003.

[101] Katsugari T, and Tani Y. Screening for microorganisms with specific characteristics by flow cytometry and single-cell sorting. Journal of Bioscience and Bioengineering, 89:3, 217–222, 2000.

[102] Collier J. L. Flow cytometry and the single cell in phycology. Journal of Phycology, 36, 628–644, 2000.

[103] Minor E. C and Nallathamby P. S. Cellular vs. ‘‘detrital’’ POM: A preliminary study using fluorescent stains, flow cytometry, and mass spectrometry. Marine Chemistry, 92, 9–21, 2004.

[104] Dubelaar G. B. J and Geerders, P. J. F. Innovative technologies to monitor Plankton dynamics scanning flow cytometry: A new dimension in real-time, insitu water quality monitoring. Sea Technology, 15–21, 2004.

[105] Dubelaar G. B. J, Geerders P. J. F and Jonker, R. R. High frequency monitoring reveals phytoplankton dynamics. Journal of Environmental Monitoring, 6, 946–952, 2004.

[106] Dubelaar G.B.J, Gerritzen P.L, Beeker A.E.R, Jonker R.R, Tangen K. Design and first results of the CYTOBUOY: an autonomous flow cytometer with wireless data transfer for in situ analysis of marine and fresh waters. Cytometry. 37, 247-254, 1999.

[107] Dubelaar G.B.J, Gerritzen P.L. CytoBuoy: a step forwards using flow cytometry in operational oceanography. Scientia Marina. 64, 255-265, 2000.

[108] http://commons.wikimedia.org/wiki/File:Cytosense.jpg

[109] http://images.slideplayer.fr/3/1296347/slides/slide_5.jpg

[110] LEPAGE R.. Reconnaissance d'algues toxiques par vision artificielle et réseau de neurones, PhD thesis, Université du Québec à Rimouski, Canada, 2004.

[111] http://www.bigelow.org

[112] Cellamare M., Rolland A., Jacquet S. Flow cytometry sorting of freshwater phytoplankton. J Applied Phycology, 22:1, pp 87, 2010.

[113] Lugli E, Pinti M, Nasi M, Troiano L, Ferraresi R, Mussi C, et al. Subject classification obtained by cluster analysis and principal component analysis to flow cytometric data, Cytometry Part A, 71:5, pp 334–344, 2007.

[114] Rajwa B, Murugesan V, Ragheb K, Banada P. P, Hirleman D, Lary T, et al., Automated classification of bacterial particles in flow by multialgle scatter measurement and support vector machine classifier, Cytometry Part A, 73:4, pp 369–379, 2008.

[115] Pereira G.C, Ebecken N.F.F, Combining in situ flow cytometry and artificial neural networks for aquatic systems monitoring, Expert Systems with Applications, 38: pp 9626–9632, 2011.

[116] Busam M. S, McNabb M, Wackwitz A, Senevirathna W, Beggah S, Meer J. R, et al., Artificial neural network study of whole-cell bacterial bioreporter response determined using fluorescence flow cytometry, Analytical Chemistry, 79, pp 9107–9114, 2007.

[117] Kim C. J, Kim H. G, Kim C. H, Oh H. M, Life cycle of the ichthyotoxic dinoflagellate Cochlodinium polykrikoides in Korean coastal waters, Harmful Algae, 6, pp 104–111, 2007.

[118] Malkassian A, Nerini D, A. van Dijk M, Thyssen M, Mante C, Gregori G. Functional Analysis and Classification of Phytoplankton Based on Data from an Automated Flow Cytometer, Cytometry Part A, 79:3, 263-275, 2011.

[119] Gluge S., Pomati F., Albert C., Kauf P., Ott T. The Challenge of Clustering Flow Cytometry Data from Phytoplankton in Lakes, Communications in Computer and Information Science, Nonlinear Dynamics of Electronic Systems, Springer International Publishing, 438, 379-386, 2014.

[120] Lo K, Brinkman R. R, Gottardo R. Automated gating of fow cytometry data via robust model-based clustering, Cytometry PartA, 73A:4, 321-332, 2008.

[121] Lo K, Hahne F, Brinkman R. R, Gottardo R. fowClust: a Bioconductor package for automated gating of flow cytometry data, BMC Bioinformatics, 10, pp145, 2009.

[122] Fraley C, Raftery A. How many clusters? which clustering methods? answers via model-based cluster analysis, Comput J, 41, pp578–588, 1998.

[123] Pyne S, Hu X, Wang K, Rossin E, Lin T, Maler L. M, et al. Automated high-dimensional flow cytometric data analysis, PNAS, 106:21, pp8519–8524, 2009.

[124] McLachlan G. J. and Peel D. Finite Mixture Models. Wiley, New York, 2000.

[125] Biernacki C., Celeux G., Anwuli A., Govaert G., and Langrognet F. Le logiciel mixmod d'analyse de mélange pour la classification et l'analyse discriminante. La Revue de Modulad, 35 , pp 25–44, 2006.

[126] http ://www-math.univ-fcomte.fr/mixmod/index.php

[127] Hammoud R. and Mohr R. Mixture densities for video objects recognition. In International Confe¬rence on Pattern Recognition (ICPR'2000), Barcelona, Spain, pp 71–75, September 2000.

[128] Reynolds D.A. and Rose R.C. Text independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3 :1, 72–83, 1995.

[129] Bishop C. M. Pattern Recognition and Machine Learning, chapter 9, Springer, pp 424–444. 2006.

[130] Dempster A. P., Laird N. M., and Rubin D. B.. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society - B Series, B :39, 1–38, 1977.

[131] Akaike H. A new look at the statistical model identification. IEEE Trans. on Automatic Control, AC-19 :6, 1974.

[132] Schwarz G.. Estimating the dimension of a model. The Annals of Statistics, 6, pp461–464, 1978.

[133] Bishop C. M.. Pattern Recognition and Machine Learning, chapter 10, pp 474–486. Springer, 2006.

[134] Bruneau P., Gelgon M., and Picarougne F.. Parameter-based reduction of Gaussian mixture models with a variational-Bayes approach. In 19th International Conference on Pattern Recognition, 2008.

[135] Vasconcelos N. and Lippman A.. Learning mixture hierarchies. Advances in Neural Information Processing Systems - MIT PressNeural Information Processing Systems, II, pp 606–612, 1998.

[136] Morgan J., Sonquist J.A. Problems in the Analysis of Survey Data, and a Proposal, Journal of the American Statistical Association, 58, pp415-435, 1963.

[137] Quinlan R. C4.5: Programs for Machine Learning, Morgan Kaufman, 1993.

[138] Quinlan J.R.. Induction of Decision Trees, Kluwer Academic publishers, 1986.

[139] Russell S et Norvig P. Intelligence Artificial: a modern approach, Prentice Hall, 1995

[140] Jianshe B, Fan B and Junyi X. Knowledge Representation and Acquisition Approach Based on Decision Tree, International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003

[141] Ruggieri S.. Efficient C4.5. Knowledge and Data Engineering, IEEE Transactions on, 14 :2 , pp 438 – 444, 2002.

[142] Denis F et Gilleron R. Apprentissage à partir d'exemple, http://www.grappa.univ-lille3.fr/polys/apprentissage/index.html.

[143] Breiman L, Friedman J, Stone C J., Olshen R.A. Classification and Regression Tree, California, Wadsworth International, 1984

[144] Quinlan J.R, Kohavi R. Decision Tree Discovery, Handbook of Data Mining and Knowledge Discovery, Klosgën and Zytkow Editors, pp 267-276, 2002

[145] Lawrence O. H, Nitesh C, Kevin W.B. Decision Tree Learning on Very Large Data Sets. Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on, 3, 2579 – 2584, 1998.

[146] Rakotomala R. Arbre de décision, Revue MODULAD 2005, numéro 33

[147] Ming D, Kothari R. Classifiability Based Pruning of Decision Trees. Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on, 3, 1739 - 1743, 2001.

[148] Diaf Y, Kaddeche M. A neural network implementation on FPGA for ecological monitoring. ACTA TECHNICA NAPOCENSIS, 45:3, 1-6, 2013.

[149] El Emary I. M. M, Fezari M, Amara F. Towards developing a voice pathologies detection system, Journal of Communications Technology and Electronics, 59:11,1280-1288, 2014.

[150] Yeoun L. J, A two-stage approach using Gaussian mixture models and higher-order statistics for a classification of normal and pathological voices, EURASIP Journal on Advances in Signal Processing, 252, 2012.

[151] Juan I., Pedro G., Manuel B. Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters, IEEE Transaction on biomedical engineering, 53:10, 1943-1953, 2006.

[152] Hershey J R., Olsen P A., Rennie S J. Variational Bhattacharyya Divergence for Hidden Markov Models. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on.

[153] Hershey J R., Olsen P A., Rennie S J. Variational Kullback-Leibler Divergence for Hidden Markov Models.Automatic Speech Recognition and Understanding, 2007. ASRU. IEEE Workshop on.

[154] Vergés-Llahi J and Sanfeliu A. Evaluation of Distances Between Color Image Segmentations, Lecture Notes in Computer Science, Pattern Recognition and Image Analysis, Springer Berlin Heidelberg, 3523, pp 263-270, 2005.

[155] Dowson D, Landau B. The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis, 12, pp 450–455, 1982

[156] http://www.saedsayad.com/model_evaluation_c.htm

[157] Xiang W., Jianping Z., Yonghong Y. Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach3 Journal of Voice, 25, pp 38–43, 2011.

[158] Pereira G.C. and Ebecken N.F.F. Combining in situ flow cytometry and artificial neural networks for aquatic systems monitoring. Expert Systems with Applications. 38, pp 9626–9632, 2011.

[159] Diaf Y, Kaddeche M And Elakremi S. Automatic recognition of phytoplankton's by combining flow cytometry and decision tree. Asian Jr. of Microbiol. Biotech. Env. Sc. 17:3, 83-88, 2015.

[160] https://fr.wikipedia.org/wiki/Alexandrium_tamarense

[161] Marand F. B. Modélisation et Identification des Systemes Non-Linéaires par Réseaux de Neurones a Temps Continu. PhD thesis, Université de Poitiers, France, 2007.

[162] Karim M. N. and Rivera S. L.. ANN in bioprocess state estimation. Advances in Biochemical Engineering Biotechnology, 46:1–33, 1992.