UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS
Département d'informatique et d'ingénierie

Estimation du niveau de Chlorophylle-A dans des lacs en utilisant l'observation à distance par satellite

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR

PABLO PEDROCCA

Décembre 2013

UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS
Département d'informatique et d'ingénierie

Ce mémoire intitulé :

Estimation du niveau de Chlorophylle-A dans des lacs en utilisant l'observation à distance par satellite

Présenté par
Pablo Pedrocca

Pour l'obtention du grade de maître ès science (M.Sc.)

a été évalué par un jury composé des personnes suivantes :

Dr. Mohand Saïd Allili……..…....…..Président du jury
Dre. Ana-Maria Cretu……..…..Membre du jury
Dr. Marek Zaremba……..Directeur de recherche

Mémoire accepté le : 23 janvier 2014

À Monica, Agustina et Martin.

Remerciements

J’aimerais remercier mon directeur de recherche, Marek Zaremba, pour l’encadrement, ses conseils et son soutien financier. Je tiens aussi à remercier les membres du jury, Dre Ana-Maria Cretu et Dr Mohand Saïd Allili qui ont accepté lire ce mémoire et me donner des conseils.

Ce projet fut réalisé en collaboration avec la firme Noetix, qui fournit les échantillons des mesures in situ, dans le cadre du Projet de collaboration de recherche Canada-China - “*Blue-green Algae Blooms Warning System*” CCRD09-102.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations, sigles et acronymes

ANN : *Artificial Neural Network* (Voir RAN)

BP : *Back Propagation* (voir PA)

CHL-A : Chlorophylle-A

MCI : *Maximum Chlorophylle Index* (Index Maximale de Chlorophylle)

MERIS : *MEdium Resolution Imaging Spectrometer* (Spectromètre d'images de résolution médiane);

MODIS : *MODerate Resolution Imaging Spectroradiometer* (Spectromètre radiométrique d'images de résolution modérée).

NIR : *Near Infra Red* (presque infrarouge).

PA : Propagation en Arrière (voir BP).

RAN : Réseau Artificiel de Neurones (voir ANN).

SeaWIFS : *Sea-viewing Wide Field-of-view Sensor* (un type de satellite comme MODIS ou MERIS)

SVM : *Support Vector Machines* (machines à vecteurs de support)

Résumé

Les eaux côtières, estuariennes et lacustres servent comme habitat à une variété très importante de flore et faune. La surveillance de la concentration de Chlorophylle-A permet d'évaluer l'abondance de phytoplancton, et à son tour, d'évaluer l'écosystème marin.

Ce mémoire propose un nouveau système pour l'estimation de la concentration de CHL-A en utilisant la télédétection à distance (*satellite remote sensing*). On vise à développer des modèles intelligents qui soient capables de s'entraîner eux-mêmes ou de subir un apprentissage supervisé, dans le but d'améliorer la précision de l'évaluation de la concentration de CHL-A. Pour ce faire, nous utilisons un système axé sur des méthodes d'apprentissage statistique pour l'estimation du CHL-A. De plus, nous faisons une classification des eaux avant l'estimation. De cette façon, on développe plusieurs modèles pour une même surface, tout en tenant compte du profil de réflexion à un moment donné.

Globalement, nous avons obtenu de meilleur résultats que comparés aux résultats sans classification de l'eau a priori et des résultats légèrement supérieurs à ceux mentionnés dans la littérature qui utilisent les différentes méthodes de classification des eaux. L'évidence présentée dans ce mémoire est la preuve que notre approche fonctionne.

Abstract

Coastal, lake, and river waters are the habitat of an important variety of vegetal and animal wildlife. The monitoring of chlorophyll-A concentration allows for the evaluation of phytoplankton abundance, and thus, the evaluation of the water ecosystem's health as a whole.

This thesis proposes a new system for estimating the CHL-A concentration by using satellite remote sensing data. We aim to develop an intelligent model capable of using supervised learning, with the goal of improving the precision of the evaluation of CHL-A concentration. To achieve this, we use an intelligent system based on statistical learning to classify the waters a priori, before estimating the CHL-A concentration with neural network models. Thus we develop several models for the same surface of water, based on the spectral signature of the samples acquired in-situ.

When performing the water classification a priori, we obtained better results compared to estimations without classifying the waters; moreover, we obtained slightly superior results to those reported in the current literature. The experimental work presented in this thesis demonstrates the advantages offered by the proposed approach.

Chapitre 1 - Introduction

Les eaux côtières, estuariennes et lacustres servent comme habitat à une variété très importante de flore et faune. Lesdites eaux sont importantes aussi pour les humains, soit pour leur valeur récréative, soit pour leur capacité de combler des besoins – par exemple, la provision d'eau pour la consommation humaine.

À cause de cela, il est important que les conditions biophysiques de ces eaux soient évaluées en permanence pour s'assurer que l'équilibre écologique est maintenu. Notamment, l'abondance excessive d'algues est un indicateur important d'un manque d'équilibre et il faut la détecter aussi vite que possible; et dans le cas idéal, il faut pouvoir la prédire. Pour y arriver, il nous faut évaluer des éléments associés à la présence d'algues, comme le phytoplancton.

La concentration de Chlorophylle-A (CHL-A) est un descripteur de l'abondance de phytoplancton. En effet, ladite mesure de concentration permet d'évaluer l'écosystème marin. Si la quantité d'éléments nutritifs dans l'eau est trop élevée, le phytoplancton se propage plus rapidement et la concentration de Chlorophylle devient plus élevée. Cependant, une concentration élevée de Chlorophylle peut entraîner des problèmes : quand les eaux sont trop nutritives, le phytoplancton se propage très rapidement, tout en épuisant des grandes quantités d'oxygène dans l'eau. Ceci provoque la suffocation et la mort des autres formes de vie marines. Ce phénomène est appelé « marée rouge » [1].

La concentration de Chlorophylle-A (CHL-A) est aussi utile pour la détection des fleurs d'eau et leur produit- les cyanobactéries (« *algal blooms* »). Les algues (« *bluegreen algae* ») sont communes et apparaissent naturellement dans plusieurs systèmes aquatiques autour du monde. Les cyanobactéries sont perçues négativement, car elles peuvent causer la mort massive de poissons, la contamination des fruits de mer avec des toxines et altérer l'écosystème d'une façon négative. Les fleurs d'eau dans des lacs et les eaux côtières sont devenues des enjeux importants non seulement pour leur impact dans les écosystèmes eux-mêmes, mais aussi dans la communauté humaine.

La plupart des systèmes développés pour surveiller et prédire les fleurs d'eau se concentrent dans les cyanobactéries toxiques en utilisant des données cueillies in situ. L'analyse the cyanotoxines disponible dans des laboratoires commerciaux utilise soit la chromatographie liquide / spectrométrie de masse (« *liquid chromatography-mass spectometry (LC-MS)* »), soit la méthode immuno-enzymatique (« *enzyme-linked immunosobrent assay* (ELISA) »)

Cependant, la prise des échantillons in situ ne permet pas d'obtenir des mesures fréquentes. À présent, la seule façon de prendre des mesures fréquentes du CHL-A est en utilisant la télédétection à distance par satellite (*satellite remote sensing*). Les techniques de télédétection à distance offrent des avantages significatifs par rapport à la surveillance in situ en fonction de la couverture temporelle et spatiale et la réduction de coûts.

Étant donné que la CHL-A est un pigment photosynthétique, elle cause des changements distinctifs dans la couleur de l'eau en absorbent et dispersant la lumière du soleil. La concentration de CHL-A peut être estimée à partir des données spectrales de réflexion captées à distance, et en faisant la relation entre la lumière reflétée dans des longitudes d'onde spécifiques et la concentration de CHL-A.

Cette tâche ne s'avère pas facile. En effet, la réflexion des rayons de soleil dans l'eau change d'un environnement à l'autre. Ces changements sont occasionnés par le type d'eau (lacustre, côtière, etc.), les caractéristiques de sa flore et sa faune, de même que le niveau de pollution de l'eau. À cause de cela, les modèles paramétriques qui ont été développés pour une surface (un lac, par exemple) ne fonctionnent pas bien dans d'autres lacs.

Ce projet de mémoire se penche sur ce sujet – plus précisément, il aborde la création d'un modèle d'évaluation des eaux lacustres axé sur les données. Dans ce cas, on entraînera un réseau de neurones avec des données satellites, plus des observations in situ de la concentration de CHL-A. On va encore plus loin, car on a décidé de classifier les eaux d'une même surface avant d'appliquer un modèle. De cette façon, on développe plusieurs modèles pour une même surface, tout en tenant compte du profil de réflexion à un moment donné.

L'objectif général de ce projet est de démontrer que la classification des eaux a priori permet d'obtenir une meilleure estimation de la concentration du CHL-A. Pour ce faire, nous avons quelques objectifs spécifiques :

- Faire une classification des eaux par la forme de leur signature spectrale, c'est-à-dire en utilisant la courbe caractéristique du profil de réflexion.
- Pour chaque type d'eau obtenue, entrainer un réseau de neurones pour générer un modèle d'entrainement.
- Avec ledit modèle, utiliser les réseaux de neurones pour obtenir des estimés du CHL-A.
- Comparer les résultats avec des estimations obtenues sans la classification, ainsi que des autres résultats dans des recherches similaires.

L’état de l’art est exposé dans la première partie de ce document. Cela permettra aux lecteurs de se familiariser non seulement avec les recherches existantes, mais aussi de comprendre les technologies disponibles et se familiariser avec la terminologie. Ensuite, on dresse la définition du problème et la motivation de cette recherche. De plus, on présente les objectifs, la méthodologie et les résultats de notre recherche ainsi qu’une comparaison avec des autres méthodes existantes. Enfin, on présente la conclusion.

Chapitre 2 - L'état de l'art

2.1 Aperçu de la télédétection par satellite

Il y a plusieurs définitions de télédétection, mais la plus appropriée dans notre contexte semble être la suivante :

> « La télédétection est la technique qui, par l'acquisition d'images, permet d'obtenir de l'information sur la surface de la terre sans contact direct avec celle-ci. La télédétection englobe tout le processus qui consiste à capter et à enregistrer l'énergie d'un rayonnement électromagnétique émis ou réfléchi, à traiter et à analyser l'information, pour ensuite mettre en application cette information. » [2]

Autrement dit, la télédétection est le processus par lequel on obtient des informations sans entrer en contact avec la source de ces informations (« cible »). Dans un schéma de télédétection par satellite, on peut trouver plusieurs composants, tel qu'illustré dans la Figure 1 :

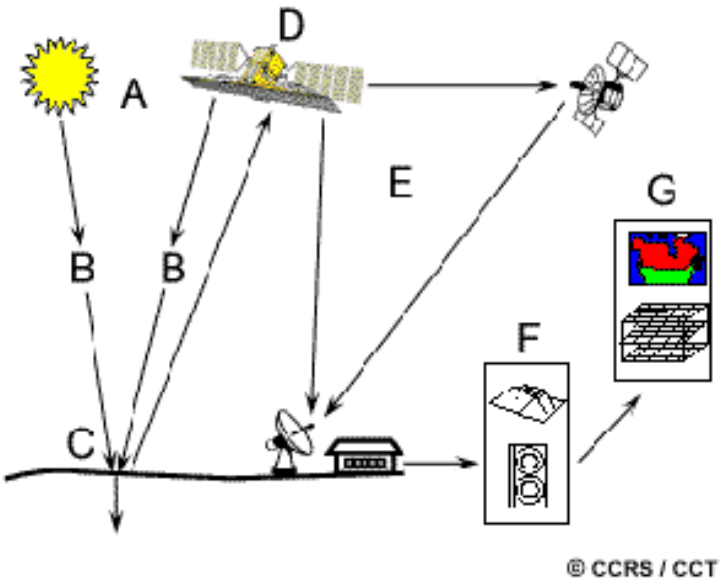

Figure 1: Composants élémentaires d'un système de télédétection [2]

A. **Source d'énergie :** Il nous faut avoir une source d'énergie pour illuminer la cible. Cette énergie peut être les rayons du soleil eux-mêmes ou elle peut être émise par le dispositif de télédétection (par exemple, dans le cas du radar). On appelle cette énergie « le rayonnement ».

B. **Interaction avec l'atmosphère :** Durant son parcours à travers l'atmosphère, l'énergie utilisée pour illuminer la cible interagit avec l'atmosphère, et ce, dans l'aller et le retour. Selon les conditions atmosphériques, il faudra plus tard faire une correction pour compenser les altérations introduites par l'interaction avec l'atmosphère.
C. **Interaction avec la cible :** Une fois l'énergie arrivée à la cible, elle interagit avec celle-là. L'énergie peut être reflétée vers le capteur (satellite) ou pourrait être réfractée (dans le cas des lacs et rivières). Cela dépend toujours de la nature de la cible ainsi que de l'angle d'incidence du rayonnement.
D. **Télédétection :** Le rayonnement de retour émis par la cible revient aux capteurs à distance et est enregistré.
E. **Transmission des données :** Les données enregistrées par les capteurs sont transmises à des stations de réception où l'information est transformée en images.
F. **Interprétation et analyse :** Une interprétation visuelle ou numérique de l'image est faite pour obtenir des informations que l'on désire obtenir de la cible. Dans cette partie du processus, on fait aussi des corrections aux images – par exemple, des corrections atmosphériques ou par rapport à l'angle du satellite.
G. **Application :** Une fois les images obtenues, on utilise cette information pour mieux comprendre la cible ou pour résoudre un problème particulier.

2.1.1 Résolution des images

Il y a plusieurs façons de définir la résolution d'une image satellite :

Résolution spatiale : La résolution spatiale donne la relation entre la taille de la surface et chaque pixel de l'image. Par exemple, si on dit qu'un capteur a une résolution spatiale de 20 m, ceci veut dire que chaque pixel de l'image représente une aire de 20 m x 20 m.

Résolution spectrale : La résolution spectrale définit la capacité d'un capteur de percevoir des bandes différentes dans le spectre électromagnétique. En effet, on peut diviser le spectre électromagnétique en plusieurs régions, aussi appelées bandes, tel qu'illustre la Figure 2; ces bandes vont des micro-ondes, passant par l'infrarouge, vers la partie visible du spectre (rouge, vert, bleu), jusqu'au l'ultraviolet.

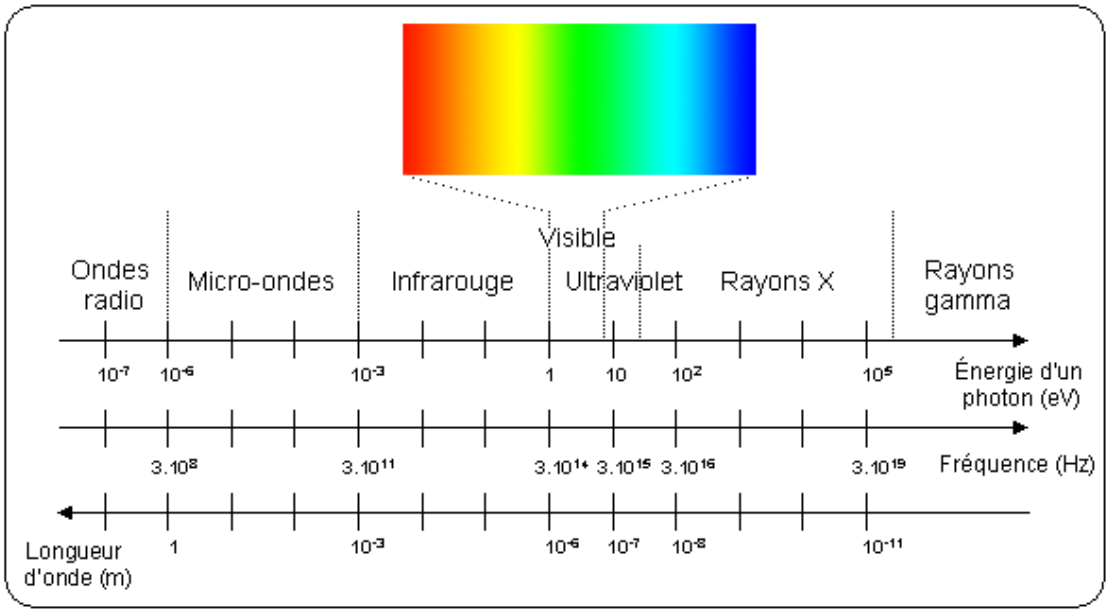

Figure 2 Le spectre électromagnétique [3]

Chaque classe d'objet (l'eau, la végétation, les roches) a une réponse différente en termes de longueur d'onde; pour faire une classification plus détaillée du type d'objet, une résolution spectrale élevée s'impose.

Chaque bande est une région limitée du spectre électromagnétique. Normalement, on énonce une bande comme « la bande de 681nm ». Cela veut dire qu'il s'agit d'une bande qui est centrée dans la longitude d'onde de 681 nanomètres.

La relation entre la longueur d'onde et la fréquence est formulée comme suit :

$$f = \frac{c}{\lambda} \qquad (1)$$

Où f représente la fréquence (en Hertz), c est la vitesse de la lumière en mètres/secondes et λ (lambda) est la longueur d'onde en mètres. D'après l'équation ci-dessous, notre longueur d'onde de 681nm est équivalente à une fréquence de 440.220 GHz.

La bande est alors définie par la longueur d'onde sur laquelle elle est centrée. Un autre concept important est la *longueur de la bande*. Un senseur est alors capable de détecter des fréquences dans le centre de la bande ainsi que dans les zones contigües qui se trouvent dans la largeur de la bande.

Par exemple, la bande centrée à 681nm a une longueur de bande de 10nm. Cela veut dire que le senseur a une réponse optimale aux signaux entre 676nm et 686nm, c'est-à-dire entre 437.020 GHz et 443.480 GHz.

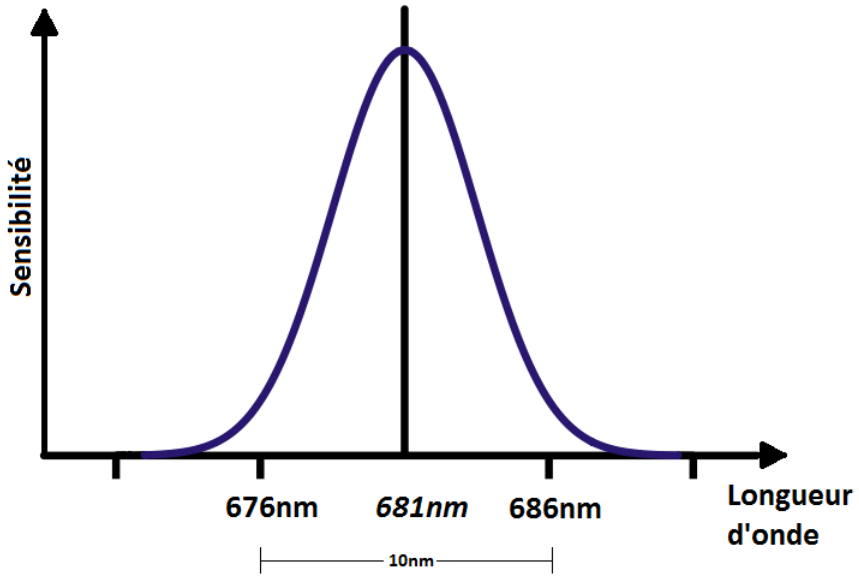

Figure 3 Réponse en fréquence d'un capteur centrée sur 681nm

La courbe montrée ci-dessus est la courbe de sensibilité du capteur. Cette courbe n'est qu'un exemple, car chaque capteur à des caractéristiques de la sensibilité spectrale différentes. Si l'on prend cette courbe comme le résultat d'une fonction de sensibilité ***F***, et les signaux reçues par le capteur sont issues d'une fonction ***S***, la fonction de sensibilité ***F*** est utilisée comme noyau (« *kernel* ») pour le calcul de la convolution avec la fonction de signal ***S***; autrement dit, l'intensité ***R*** du signal à la sortie du capteur sera égale à l'intégrale du produit de la sensibilité ***F*** et le signal ***S***.

$$R = \int F \times S \qquad (2)$$

Résolution radiométrique : La résolution radiométrique est la capacité d'un capteur à détecter des variations d'amplitude (c.-à-d. intensité) du signal capté. Cette résolution est toujours exprimée en puissances de deux (2^n). Si un capteur a une résolution radiométrique de 4 bits, il peut détecter $2^4 = 16$ échelons d'intensité et ainsi de suite.

Pour exprimer l'intensité des signaux, on utilise l'acronyme « R_{rs} », pour « *Remote Sensing Reflectance* ». On doit interpréter R_{rs} comme l'intensité du signal mesuré par les capteurs du satellite. Cet acronyme peut être suivi de la longueur d'onde pour spécifier la bande correspondante :

$R_{rs}(\lambda)$ = Valeur de réflectance d'une bande centrée dans la longueur d'onde lambda

$R_{rs}(681)$ = Valeur de réflectance dans une bande centrée dans la longueur d'onde de 681nm.

$R_{rs}(\lambda_{681})$ = Valeur de réflectance dans une bande centrée dans la longueur d'onde de 681nm.

Les expressions $R_{rs}(681)$ et $R_{rs}(\lambda_{681})$ sont équivalentes.

Résolution temporelle : La résolution temporelle est l'intervalle qui s'écoule entre les prises d'images d'une aire déterminée. En effet, le période de revisite d'un satellite peut comprendre plusieurs jours (c.-à-d. pour qu'un satellite survole la même aire, entre 1 et 5 jours peuvent s'écouler).

2.1.2 Senseurs actifs et passifs

Dépendamment du type de satellite, il peut y avoir des senseurs actifs ou passifs.

Senseur passif: Un senseur passif capte l'énergie provenant des objets dans la surface. Cette énergie peut être reflétée par l'objet ou émise par l'objet lui-même. Dans le cas de l'énergie reflétée, on parle des objets illuminés par la lumière du soleil. Dans le cas de l'énergie irradiée par les objets eux-mêmes, on parle des images infrarouges qui détectent la chaleur ou des images nocturnes qui détectent les lumières dans la surface terrestre.

Senseur actif: Un senseur actif fournit sa propre énergie pour illuminer la cible. Deux exemples sont les capteurs radar (SAR, « *Synthetic Aperture Radar* ») et fluorodétecteur au laser.

2.1.3 Interactions avec l'atmosphère

Pour que l'énergie puisse illuminer la cible, elle doit traverser l'atmosphère. Ceci est aussi le cas pour l'énergie reflétée. Autrement dit, il y a une interaction de l'énergie électromagnétique dans l'aller et le retour. Les mécanismes d'interaction sont définis comme l'absorption et la dispersion ou diffusion (« *scattering* »).

Absorption: Ce phénomène fait en sorte que les molécules de gaz dans l'atmosphère absorbent l'énergie dans des longueurs de bandes différentes. Par conséquence, il y aura une diminution de la quantité d'énergie qui aura un impact avec la cible (ou le senseur, au retour).

Diffusion (« *scattering* ») : La diffusion est le processus opposé à l'absorption. Ce phénomène se présente quand des particules ou des molécules de gaz dans l'atmosphère interagissent avec la radiation et font changer sa direction originale.

2.1.4 Interactions avec la cible

Il y a trois façons dont l'énergie incidente dans la cible peut interagir avec elle : absorption, transmission et réflexion :

Absorption : Quand le phénomène d'absorption se produit, l'énergie incidente est prise par la cible. Par exemple, quelques surfaces se chauffent quand elles sont illuminées par le soleil.

Transmission : La transmission se présente quand l'énergie passe à travers d'un objet et continue son parcours.

Réflexion : La réflexion se produit quand l'énergie rebondit dans la cible et prend une autre direction. Dans la plus part des cas, on s'intéresse aux mesures de cette énergie.

Il faut remarquer que les trois types d'interaction peuvent avoir lieu lors de l'interaction avec la cible. Par exemple, quand la lumière du soleil frappe une surface d'eau, une partie est absorbée (l'eau se chauffe), une autre partie est transmise (et illumine les objets au-dessous de la surface) et une autre partie est reflétée.

2.2 Utilisation de la télédétection par satellite pour l'estimation du CHL-A

Tel que mentionné dans les pages précédentes, les techniques de télédétection à distance offrent des avantages significatifs par rapport à la surveillance in situ en fonction de la couverture temporelle et spatiale et la réduction de coûts.

Les types d'instruments les plus utilisés pour la télédétection sont MODIS, MERIS et SeaWiFS. SeaWiFS est l'acronyme pour instrument à grand champ pour l'observation des mers (« *Sea-viewing Wide Field-of-View Sensor* »); il est en orbite depuis 1997 et fournit des images aquatiques en couleur, utiles à l'évaluation des paramètres de la qualité de l'eau [4]. Les observations journalières de SeaWiFS dans la région des grands lacs fournissent des données qui permettent l'obtention des tendances des paramètres de qualité de l'eau. Cependant, la baisse résolution spatiale (1km x 1km) limite l'application de SeaWiFS pour la surveillance des environnements aquatiques intérieurs (lacs et rivières).

D'autres instruments les plus souvent utilisés sont MERIS - *MEdium Resolution Imaging Spectrometer* (Spectromètre d'images de résolution médiane) et MODIS *MODerate Resolution Imaging Spectroradiometer* (Spectromètre radiométrique d'images de résolution modérée). Ces instruments offrent une résolution spatiale et spectrale supérieure[5]. En utilisant des senseurs hyper spectraux, un seul instrument a la capacité de classifier et quantifier des environnements aquatiques complexes, y compris l'identification du phytoplancton par groupe et composants chimiques spécifiques. Un des systèmes qui s'appuient dans des données satellites et des données in situ fut développé par l'Administration Nationale de l'Atmosphère et les Océans (« *National Oceanic and Atmospheric Administration* (NOAA) ») aux États-Unis, pour évaluer la qualité des eaux dans le Golfe de Mexique. Il y a des essais sur des prototypes qui ont été faits dans l'Europe et l'Asie [6]. Un système de télémétrie fut développé pour faire le suivi des changements de fluorescence de chlorophylle, oxygène dissous et autres paramètres [7].

2.3 MODIS et MERIS

Les types d'instruments les plus utilisés pour l'observation de la Terre à moyenne résolution sont MODIS et MERIS. MERIS est le *MEdium Resolution Imaging Spectrometer* (Spectromètre d'images de résolution médiane); il fait partie du Satellite d'Observation Terrestre ENVISAT. MERIS est consacré premièrement à l'observation des océans, mais sa mission fut changée pour inclure les observations des côtes et des surfaces terrestres. MERIS a une résolution spatiale de 260m x 300m, mais sa caractéristique la plus

importante est la capacité de programmer ses senseurs pour établir le centre et la largeur de chacune de ses bandes. MERIS a été conçu pour la télédétection de 15 bandes, entre 390nm et 1040 nm du spectre électromagnétique [8]. La table suivante montre les 15 bandes spectrales qui ont été conçues pour des applications océanographiques et interdisciplinaires :

Table 1: Bandes spectrales du MERIS

MDS Nr.	**Centre de la bande (NM)**	**Longueur de la bande (NM)**	**Applications potentielles**
1	412.5	10	Substances jaunes et détritus
2	442.5	10	Maximum d'absorption de Chlorophylle
3	490	10	Chlorophylle et autres pigments
4	510	10	Sédiments en suspension, marée rouge
5	560	10	Minimum d'absorption de Chlorophylle
6	620	10	Sédiments en suspension
7	665	10	Référence fluoroscopique d'absorption de Chlorophylle
8	681.25	7.5	Fluorescence maximale de Chlorophylle
9	708.75	10	Corrections atmosphériques
10	753.75	7.5	Végétation, nuages
11	760.625	3.75	Absorption d'oxygène R-branch
12	778.75	15	Corrections atmosphériques
13	865	20	Végétation, vapeur d'eau
14	885	10	Corrections atmosphériques
15	900	10	Vapeur d'eau, surface terrestre

MODIS reçoit son nom de *MODerate Resolution Imaging Spectroradiometer* (Spectromètre radiométrique d'images de résolution modérée). MODIS est un instrument essentiel dans les satellites Terra (EOS AM) et Aqua (EOS PM). Terra fait des orbites du nord au sud et fait la traverse de l'équateur pendant les matins, tandis qu'Aqua orbite de sud au nord et traverse l'équateur pendant l'après-midi. Ces deux satellites sont capables de voir la surface entière de la Terre chaque 1-2 jours [9].

À la différence de MERIS (300m x 300m), MODIS a une résolution de seulement 1km x 1km et présente 36 bandes. Les bandes 20-36 sont utilisées uniquement pour des phénomènes atmosphériques (températures atmosphériques, nuages, températures des nuages et de surface, etc.). Par souci de simplicité, on montre ici seulement les bandes 1-19 qui incluent les 16 bandes utilisées dans notre étude :

Table 2: Bandes spectrales du MODIS

Utilisation principale	Bande	Largeur de la bande	Radiance spectrale	
Terre / Nuages /Limites des Aérosols	1	620 - 670	21.8	
	2	841 - 876	24.7	
Terre / Nuages /Propriétés des Aérosols	3	459 - 479	35.3	
	4	545 - 565	29.0	
	5	1230 - 1250	5.4	N'est pas utilisée
	6	1628 - 1652	7.3	N'est pas utilisée
	7	2105 - 2155	1.0	N'est pas utilisée
Couleur des Océans / Phytoplancton/ Biogéochimique	8	405 - 420	44.9	
	9	438 - 448	41.9	
	10	483 - 493	32.1	
	11	526 - 536	27.9	
	12	546 - 556	21.0	
	13	662 - 672	9.5	
	14	673 - 683	8.7	
	15	743 - 753	10.2	
	16	862 - 877	6.2	
Vapeur d'eau atmosphérique	17	890 - 920	10.0	
	18	931 - 941	3.6	
	19	915 - 965	15.0	

2.4 Les modèles paramétriques pour l'évaluation du CHL-A

Un modèle dit paramétrique fonctionne en utilisant une formule fixe ayant comme paramètres quelques coefficients invariables. La valeur d'estimation du CHL-A est alors une fonction de la valeur de l'intensité des bandes comme variables indépendantes.

2.4.1 Les modèles à bandes

La plupart des études récentes se concentrent sur des modèles paramétriques axés sur un modèle à bandes, soit un modèle à trois bandes ou, dans un cas particulier, un modèle à deux bandes. Ces modèles peuvent être décrits comme :

Modèle à trois bandes :

$$CHLA \cong (Rrs_{\lambda 1}^{-1} - Rrs_{\lambda 2}^{-1})R_{\lambda 3}. \quad (3)$$

Modèle à deux bandes :

$$CHLA \cong (Rrs_{\lambda 1}^{-1})Rrs_{\lambda 3}, \quad (4)$$

où chaque bande $R_{\lambda n}$ correspond à une longueur d'onde différente, soit dans MODIS ou MERIS. Le choix des bandes est essentiel pour le bon fonctionnement du modèle. Morel et al. [10] furent parmi les premiers à identifier la région spectrale entre 400nm et 600nm. Les modèles à bandes mentionnés ci-dessus utilisent des facteurs de réflexion dans les régions spectrales rouge et presqu'infrarouge (« *near-infrared* » (NIR))[11], un modèle similaire utilisé par Carder et al. [12] pour l'estimation du CHL-A avec données du MODIS.

Des modèles similaires ont été mis en pratique par O'Reilly et al. [13] qui introduisent le modèle OC2v4. Il s'agit d'un modèle paramétrique à deux bandes, traduit par :

$$CHLA = 10^{(a_0+a_1R+a_2R^2+a_3R^3)} + a_4$$
$$a[a_0, a_1, a_2, a_3, a_4] = [0.319, -2.336, 0.879, -0.135, -0.071] \quad (5)$$
$$R = log_{10}\left(\frac{R_{rs}(490)}{R_{rs}(555)}\right)$$

$R_{rs}(490)$ et $R_{rs}(555)$ représentent les bandes MODIS à 490nm et 555nm respectivement.

Un autre modèle similaire fut introduit, encore une fois, par[14]. Ce modèle, appelé le modèle OC4v4, est donné par :

$$CHLA = 10^{(a_0+a_1R+a_2R^2+a_3R^3+a_4R^4)}$$

$$a[a_0, a_1, a_2, a_3, a_4] = [0.366, -3.067, 1.930, 0.649, -1.532] \quad (6)$$

$$R = log_{10}\left(\frac{\max(R_{rs}(443), R_{rs}(490), R_{rs}(510)}{R_{rs}(555)}\right)$$

Les deux modèles mentionnés ci-dessous s'appliquent au SeaWiFS. Pour MODIS, les mêmes auteurs ont introduit le modèle OC3M. Ce modèle a été officiellement adopté par la NASA pour faire l'évaluation des images MODIS et il est formulé comme suit :

$$CHLA = 10^{(a_0+a_1R+a_2R^2+a_3R^3+a_4R^4)}$$

$$a[a_0, a_1, a_2, a_3, a_4] = [0.2830, -2.753, 1.457, 0.659, -1.403] \quad (7)$$

$$R = log_{10}\left(\frac{\max(R_{rs}(443), R_{rs}(488))}{R_{rs}(555)}\right)$$

Les analogies entre les modèles sont évidentes; en effet, il s'agit des variations d'un même modèle. Le choix de bandes et les coefficients changent pour s'adapter à un environnement géographique. Il s'agit toujours des modèles paramétriques et régionaux.

2.4.2 Le modèle de la forme spectrale

Dans l'article présenté par Gower et al. [15], les auteurs proposent une méthode axée sur la forme de la caractéristique spectrale (« *spectral shape* »). Un modèle de forme spectrale est exprimé comme :

$$SS(\lambda) = R_{rs}(\lambda) - R_{rs}(\lambda^-) - \{R_{rs}(\lambda^+) - R_{rs}(\lambda^-)\} \times \frac{(\lambda - \lambda^-)}{(\lambda^+ - \lambda^-)} \quad (8)$$

Dans cette équation, R_{rs} est une valeur de réflectance, λ est une bande quelconque, tandis que λ^+ et λ^- sont les bandes voisines (avant et après); par exemple, dans la Table 2: Bandes spectrales du MODIS on peut voir la liste de bandes. Si l'on choisit la bande 9 comme λ, λ^+ et λ^- seront les bandes 10 et 8

respectivement. Dans la formule originale où l'on utilise des bandes MERIS, λ est la bande centrée à 681nm, λ^+est 709 nm et λ^- est 665 nm, équivalentes aux bandes 8, 9 et 7 de MERIS respectivement (voir Table 1). Il faut remarquer qu'un modèle de forme spectrale est réduit à une dérivée de deuxième ordre quand les bandes sont espacées régulièrement, c'est-à-dire $(\lambda^+ - \lambda) = (\lambda - \lambda^-)$ [16, 17]; dans ce cas particulier, le ratio entre la différence de bandes devient (681-665) / (709-681) = 0.5714.

À la différence des modèles qui utilisent une relation entre les bandes (« *ratio algorithms* »), un modèle de forme présente l'avantage de ne pas être sensible à la présence des radiances négatives. Ainsi, une correction atmosphérique mauvaise ou incomplète n'a pas d'influence sur les résultats.

Ce type d'modèle nommé MCI (*Maximum Chlorophylle Index*) fut présenté par Gower et al. [15] et repris par d'autres auteurs [16, 17]. Malheureusement, il s'agit d'un produit spécifique à MERIS et il utilise des bandes qui ne sont pas présentes dans MODIS (par exemple, la bande 709).

L'index maximal de Chlorophylle (« Maximum Chlorophyll Index » ou MCI) est devenu un des standards pour l'évaluation de la concentration de Chlorophylle à partir des données recueillies par MERIS. Le MCI montre l'amplitude d'un sommet dans la bande couvrant 705nm (car la bande MERIS est centrée sur 709 nm). Le MCI, illustré d'une façon graphique dans la Figure 4, est exprimé par :

$$MCI = R_{rs}(709) - R_{rs}(681) - \left[\frac{(709-681)}{(753-681)}(R_{rs}(753) - R_{rs}(681))\right] \qquad (9)$$

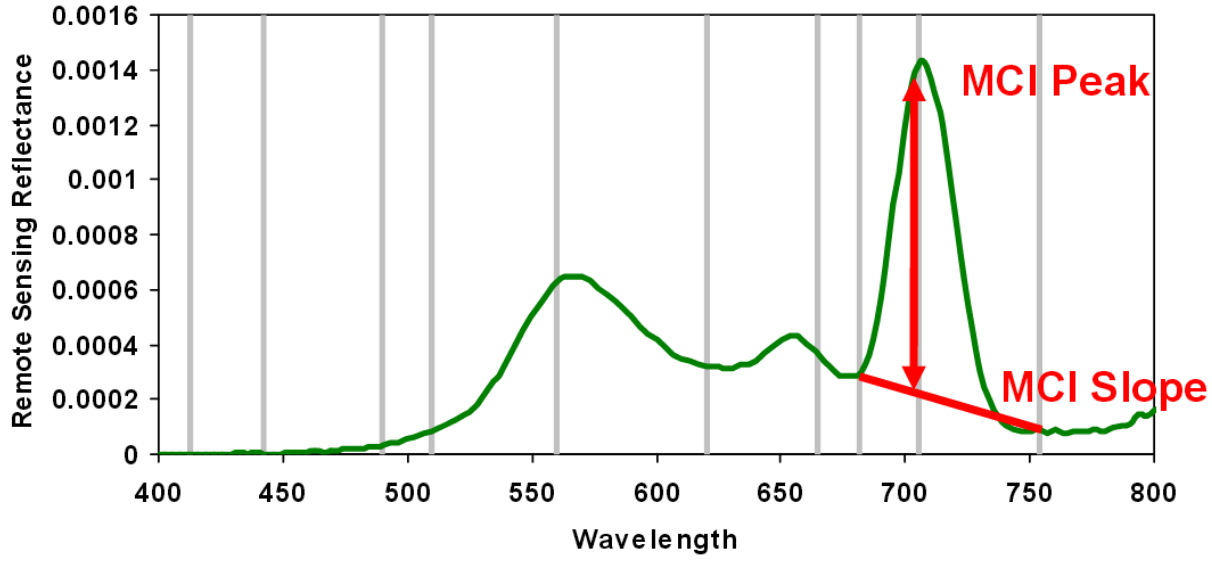

Figure 4: Représentation graphique du MCI [18]

La formule pour calculer le MCI fonctionne de la façon suivante: d'abord, on calcule la pente entre les bandes λ^+ et λ^-; ensuite, on calcule la valeur (élévation) du point de ce pente correspondant à la localisation de la bande centrale λ; finalement, on calcule la différence entre ce point particulier et la valeur de réflectance de la bande λ. Ceci nous donne l'amplitude du sommet de la bande λ par rapport à la pente entre les bandes λ^+ et λ^-.

Parmi les satellites couvrant les eaux côtières et océaniques, cet index est unique à MERIS, car la bande couvrant 705nm (centrée à 709nm dans MERIS) n'est pas disponible dans MODIS ou SeaWiFS [19].

2.5 Les modèles d'apprentissage statistique

Une autre approche pour estimer la concentration de CHL-A est l'utilisation de systèmes d'apprentissage statistique qui peuvent être entraînés (sujets à un apprentissage supervisé). Cela élimine le besoin de calculer des paramètres spécifiques pour chaque surface d'eau différente. Au contraire, il faut « entraîner » le système pour chaque environnement et développer ainsi un modèle d'interprétation qui sera utilisé par le système intelligent.

En général, les méthodes d'apprentissage statistique montrent des meilleurs résultats par rapport aux modèles paramétriques. Liu et al. [20] font une comparaison entre la Régression Multiple Statistique (un modèle paramétrique), les réseaux artificiels de neurones (RAN) et des machines à vecteurs de support (SVM, « *Support Vector Machines* »), tout en obtenant les meilleurs résultats avec les RAN. Il faut remarquer qu'ils n'utilisent pas seulement les données de télédétection, mais aussi d'autres données telles que la température de l'eau, la densité des solides en suspension, la profondeur de l'eau, etc.

Une de ces méthodes d'apprentissage est le réseau artificiel de neurones. Les réseaux de neurones sont utilisés par la prédiction numérique et la classification et peuvent être considérés comme des modèles de régression non-linéaires avec un nombre très élevé de coefficients ajustables avec le but de résoudre un problème de minimisation par rapport à l'erreur du modèle .

L'utilisation des réseaux artificiels de neurones n'est pas une idée nouvelle. Plusieurs auteurs se sont déjà penchés sur le sujet et l'ont abordé de plusieurs points de vue [21-23]. D'abord, il y a des recherches qui essaient d'estimer la concentration de CHL-A, non pas en utilisant la télédétection, mais plutôt la concentration des composants chimiques dans l'eau [24-31]. Ces études analysent des

paramètres tels que le pH, l'oxygène dissous, l'azote ammoniacal et le phosphore ainsi que d'autres paramètres comme la température de l'eau. Aussi, Xiaoshen et al. [1] émettent une comparaison entre modèle basés sur des RAN et des méthodes paramétriques comme OC4V4. Liu et al. [32], pour leur part, utilisent une « *feed forward error-back propagation network (FFNN)* » pour estimer la concentration de CHL-A dans des environnement nuageux.

Dans d'autres études [33], les auteurs ont appliqué les données de chaque bande satellite directement au réseau de neurones. Dans quelques autres cas, la modélisation a été faite sans y utiliser la télédétection, mais plutôt des données atmosphériques recueillies *in situ* [20]. Bien que les méthodes exposées ci-dessus ne s'appliquent pas directement à notre recherche, elles sont cependant utiles pour évaluer l'application des réseaux de neurones à l'estimation du CHL-A.

D'autres analyses plus complètes [34] furent effectuées sur plusieurs bandes du satellite Landsat-TM, en n'utilisant que quatre bandes (450-520nm; 520-600nm; 630-690nm; et 760-900nm). Les auteurs ont utilisé trois méthodes : 1) Les données des bandes appliquées directement sur un réseau de neurones, 2) Les coefficients de données appliqués sur le réseau des neurones et 3) L'interpolation des données de la première méthode avec les données de la deuxième méthode. Les auteurs ont constaté que la première méthode (c.-à-d. d'appliquer les données des bandes directement aux réseaux de neurones) donnait de meilleurs résultats avec un coefficient de corrélation de 0.92.

Finalement, d'autres auteurs [35] suggèrent de faire le *clustering* des eaux a priori, avant d'utiliser les réseaux de neurones. Par conséquence, il y aura autant de modèles que de classes d'eaux trouvées. La séparation des types d'eaux se fait par la méthode de *clustering*, plus précisément par l'algorithme des c-moyens flous (« *fuzzy C-Means* »). Avec cela, les auteurs ont trouvé entre 6 et 7 *clusters*, qui, à leur tour, ont été utilisés pour alimenter les réseaux de neurones correspondants.

Chapitre 3 - Estimation du CHL-A

3.1 Description du problème et motivation

Tel que mentionné dans notre introduction, le processus d'évaluation de la concentration the CHL-A en utilisant la télédétection présente les limitations suivantes :

- Les modèles paramétriques sont très dépendants du type d'eau et souvent développés pour un environnement géographique particulier. Ces mêmes modèles paramétriques ne s'adaptent pas bien à d'autres surfaces [11], [36], [15], [17].
- Même si on a développé un modèle paramétrique correct pour un environnement particulier, ledit modèle est souvent appliqué avec une approche « tout-ou-rien », c'est-à-dire qu'on applique le même modèle dans tous les états du lac, sans prendre en considération les changements globaux au niveau de la surface de l'eau. Par exemple, même si la concentration de CHL-A ne change pas mais la température de l'eau change, le profil de réflexion de l'eau sera différent. Il faut alors adapter notre modèle d'estimation the CHL-A à ce nouveau profil de réflexion.
- La plupart des modèles ont été développés pour les images du MERIS. Cependant, MERIS est en train d'être remplacé par MODIS. On vise ici à développer un système indépendant de la source et adaptable à n'importe quel type d'instrument (MODIS, MERIS, SeaWiFIs, etc.).

Dans le cadre de cette recherche, on vise à développer des modèles intelligents qui soient capables de subir un apprentissage supervisé, dans le but d'améliorer la précision de l'évaluation de la concentration de CHL-A. Une amélioration de la précision est importante, car elle permettra une meilleure évaluation des caractéristiques biophysiques des surfaces d'eau ainsi qu'une meilleure surveillance de l'écologie des lieux. Finalement, une meilleure évaluation de la qualité des eaux aura des impacts positifs dans la population dans les cas des lacs utilisés, soit pour des activités récréatives ou pour fournir de l'eau pour la consommation humaine.

3.2 Objectifs poursuivis

L'objectif de cette recherche est de trouver un modèle d'estimation de la concentration de Chlorophylle-A en utilisant des images satellites multi-spectrales comme source. Malgré le fait que l'objectif est simple, la tâche s'avère complexe. En effet, les modèles d'usage courant ont des taux d'erreur très grands qui, souvent, empêchent d'avoir des bonnes mesures d'évaluation.

Pour cette recherche, on essaie de résoudre les problèmes des modèles existants. Parmi ces problèmes, on peut mentionner :

- **Les modèles paramétriques sont des modèles régionaux ou locaux :** En effet, les modèles paramétriques sont adaptés à une surface d'eau en particulier. Dans la plupart de la littérature, on trouve des modèles appliqués à un lac spécifique. En conséquence, il faut adapter les modèles (ou les paramètres) pour chaque environnement particulier.
- **L'application des modèles se fait d'une façon globale :** On applique le même modèle sur toutes les observations, sans discrimination entre les signaux. Lors de notre recherche, on a découvert que le signal capté à distance présente des profils différents, c'est-à-dire des courbes caractéristiques uniques. Cela nous amène à croire que, dans une surface particulière, il y a des profils d'eaux différents qui dépendent non seulement de la concentration de CHL-A mais aussi d'autres facteurs tels que les conditions atmosphériques.

En tenant compte des problèmes mentionnés ci-dessus, on essaie de concevoir non pas un modèle, mais plutôt un système d'évaluation de la concentration de CHL-A qui puisse être l'objet d'un entraînement. Nous aurons alors un système capable d'apprendre en utilisant les données prises sur les lieux (*in situ data*), plus les images satellites; avec cet ensemble de données, on peut procéder à l'entraînement du système. Cela rend le système capable de travailler dans n'importe quelle surface, car il n'est pas nécessaire d'adapter un modèle paramétrique (ou en calculer un autre à nouveau). Nous avons choisi des méthodes d'apprentissage statistique; une fois entraînés, les systèmes d'apprentissage statistique génèrent un modèle qui sera appliqué à l'ensemble des données pour obtenir des résultats, dans notre cas, la concentration de CHL-A.

Afin d'améliorer la précision du système, on a décidé de ne pas appliquer un seul modèle à tout l'ensemble de données. Tel que mentionné, nous avons remarqué que pour une même surface, le signal capté présente des profils caractéristiques. Ces courbes différentes nous donnent des types d'eaux

différentes. Nous avons l'hypothèse qu'en isolant chaque type d'eau et en ayant un modèle neuronal différent pour chaque type d'eau, nous allons améliorer la précision du système.

Nous allons tester la validité de notre hypothèse dans le cadre de cette recherche.

3.3 Méthodologie

La clé de notre approche est de classifier les eaux *a priori*, avant de faire une estimation de la concentration du CHL-A. ***La nouveauté proposée dans le cadre de notre recherche est le fait de classifier les eaux par la forme de sa courbe spectrale caractéristique***. Nous allons exposer dans les paragraphes suivants quelques méthodes différentes des classifications / *clustering* existantes avec leurs limitations ainsi que la solution proposée.

3.3.1 Utilisation de SVM pour la classification des eaux

Pour faire la classification des eaux, nous avons décidé d'utiliser un classificateur intelligent capable de faire de l'apprentissage supervisé. De cette façon, nous pouvons prendre un ensemble de courbes classifiées *a priori* et les donner au classificateur comme un ensemble de données d'entraînement, ou d'« exemples ». En utilisant lesdits « exemples », le classificateur génère un modèle de classification, qui est ensuite utilisé pour faire la classification des autres données.

Il y a plusieurs types d'apprentissage intelligent (« *boosting* », machines à vecteurs de support, réseaux de neurones, méthodes de k plus proches voisins, etc.). Nous avons décidé d'utiliser les machines à vecteurs de support (« *support vector machines* » ou SVM).

SVM, dans sa version originale, est un classificateur linéaire; il fut crée en tant que tel [37]. De plus, et lors de nos développements préliminaires, nous avons comparé SVM et AdaBoost avec des résultats semblables, SVM offrant une meilleure performance et précision dans un contexte de manque de données (c'est-à-dire peu d'échantillons de référence). Finalement, il y avait des contraintes externes comme la disponibilité de librairies. En conclusion, nous avons retenu SVM comme classificateur des eaux.

3.3.2 Utilisation des réseaux de neurones pour l'estimation du CHL-A

Un réseau artificiel de neurones (RAN) est une structure utilisée pour la modélisation d'un grand volume de données non linéaires; il est considéré comme une « boite-noire », étant donné qu'on cherche à prédire les valeurs de sortie en fonction des valeurs d'entrée, sans savoir comment le modèle qui est sous-tendu fonctionne. Un réseau de neurones est très utile pour modéliser un problème particulier : il ne faut que faire l'entraînement du réseau en fonction des valeurs d'entrée et des valeurs de sortie déjà connues. Cela génère un modèle qui sera utilisé plus tard pour obtenir une estimation des valeurs de sortie en fonction des valeurs d'entrée inconnues.

Le choix d'un réseau artificiel de neurones est justifié par le fait qu'un RAN est capable de bien généraliser avec une faible dégradation progressive, ce qui est avantageux dans des situations avec peu de données. Par exemple, Sorayya et al. [25] n'utilisent que 68 enregistrements pour l'entraînement et 38 pour l'évaluation; Wang et al. [38] utilisent 100 enregistrements pour l'entraînement et 100 pour l'évaluation; Xiuli et al.[29] utilisent 151 et 50 enregistrements pour l'entraînement et l'évaluation respectivement.

Les RAN fonctionnent bien dans un contexte de manque de linéarité; d'autres méthodes, telles que des méthodes bayésiennes, font l'hypothèse d'une distribution nettement gaussienne [39]. En ce qui concerne d'autres méthodes d'apprentissage statistique comme SVM, nous n'étions pas en mesure d'établir le paramètre d'écart (« *slack parameter* ») a priori (le paramètre d'écart contrôle le compromis entre un marge large et l'erreur du classificateur). Puis, SVM fut conçu comme un outil de classification, pas un outil de régression (Boser, Vapnik et al. [37]). De plus, il y a des essais dans la littérature faisant la comparaison entre SVM et RANs; dans cette comparaison, les réseaux de neurones montrent une performance supérieure à SVM [20].

Finalement, nous avons des considérations pratiques : les RANs sont déjà utilisés comme de modèles inverses pour faire la correction atmosphérique des images satellites. En utilisant des RAN, nous pouvons intégrer nos systèmes avec des procédures existantes.

En conclusion, les réseaux artificiels de neurones (RAN) sont pleinement justifiés comme outil pour l'estimation du CHL-A. Dans les sections suivantes, nous allons expliquer comment un RAN fonctionne et l'architecture choisie pour faire l'estimation du CHL-A.

3.3.3 Sommaire de la méthodologie

Tel que décrit dans les sections précédentes, on essaie de concevoir pas un modèle, mais plutôt un système d'évaluation de la concentration de CHL-A qui puisse être l'objet d'un entraînement. Nous aurons alors un système capable d'apprendre en utilisant les données prises sur les lieux (in situ data), plus les images satellites. Avec cet ensemble de données, on pourra procéder à l'entraînement du système. D'abord, on fera une classification manuelle des types d'eaux et ensuite, on procédera à entraîner un classificateur SVM pour le rendre capable de classifier les différents types d'eaux. Ensuite, on prendra chaque type d'eau différente et on procédera à l'utiliser comme entrée pour chaque réseau de neurones. Le schéma du processus d'entraînement est montré dans la figure 13 ci-dessous.

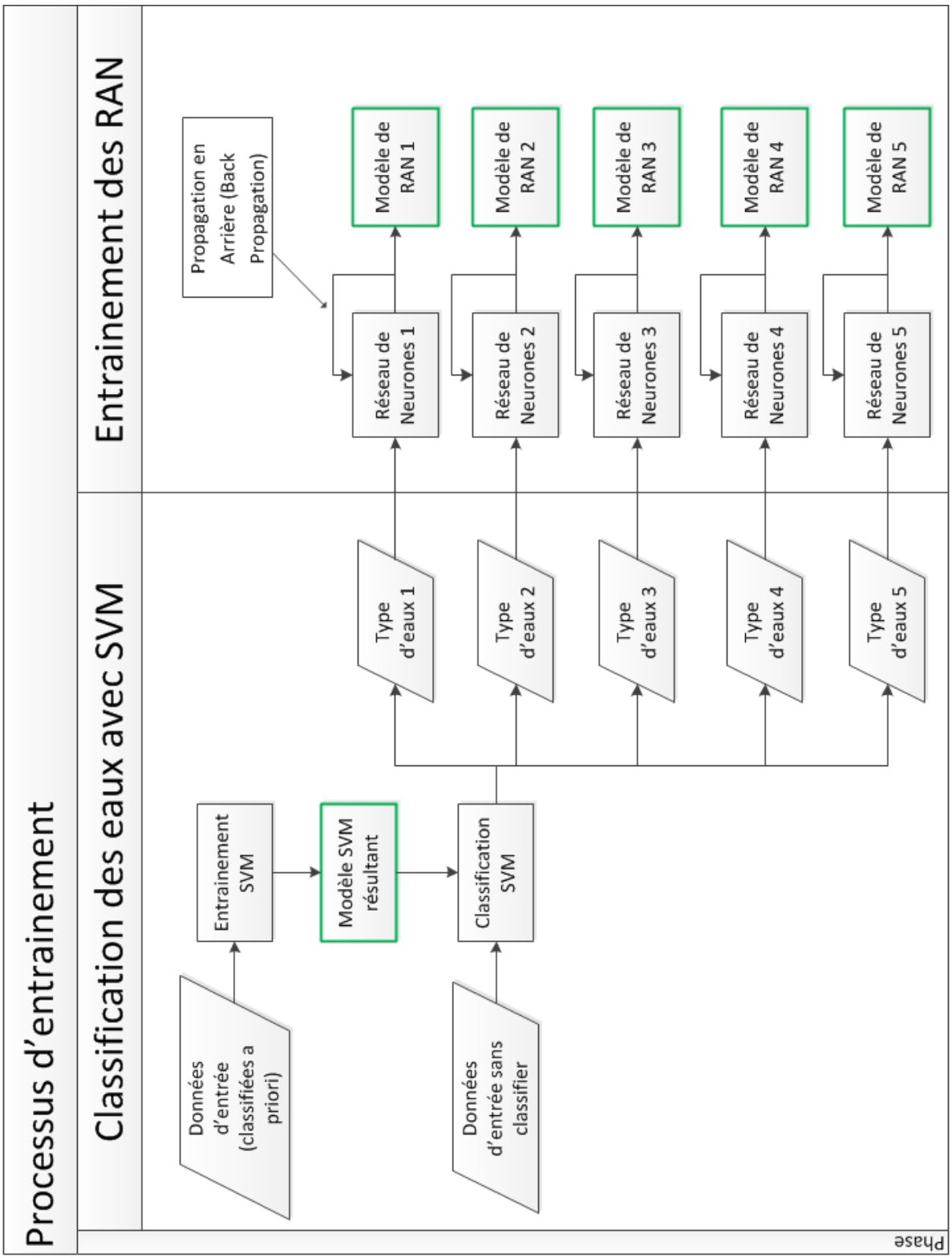

Figure 5: Sommaire du processus d'entrainement

Les modèles ainsi générés serviront à faire les estimations de CHL-A; un modèle unique sert à faire la classification des types d'eau, tandis que cinq modèles différents (un pour chaque type d'eau) sont utilisés par cinq réseaux de neurones pour faire l'estimation du CHL-A. Voici le schéma du processus :

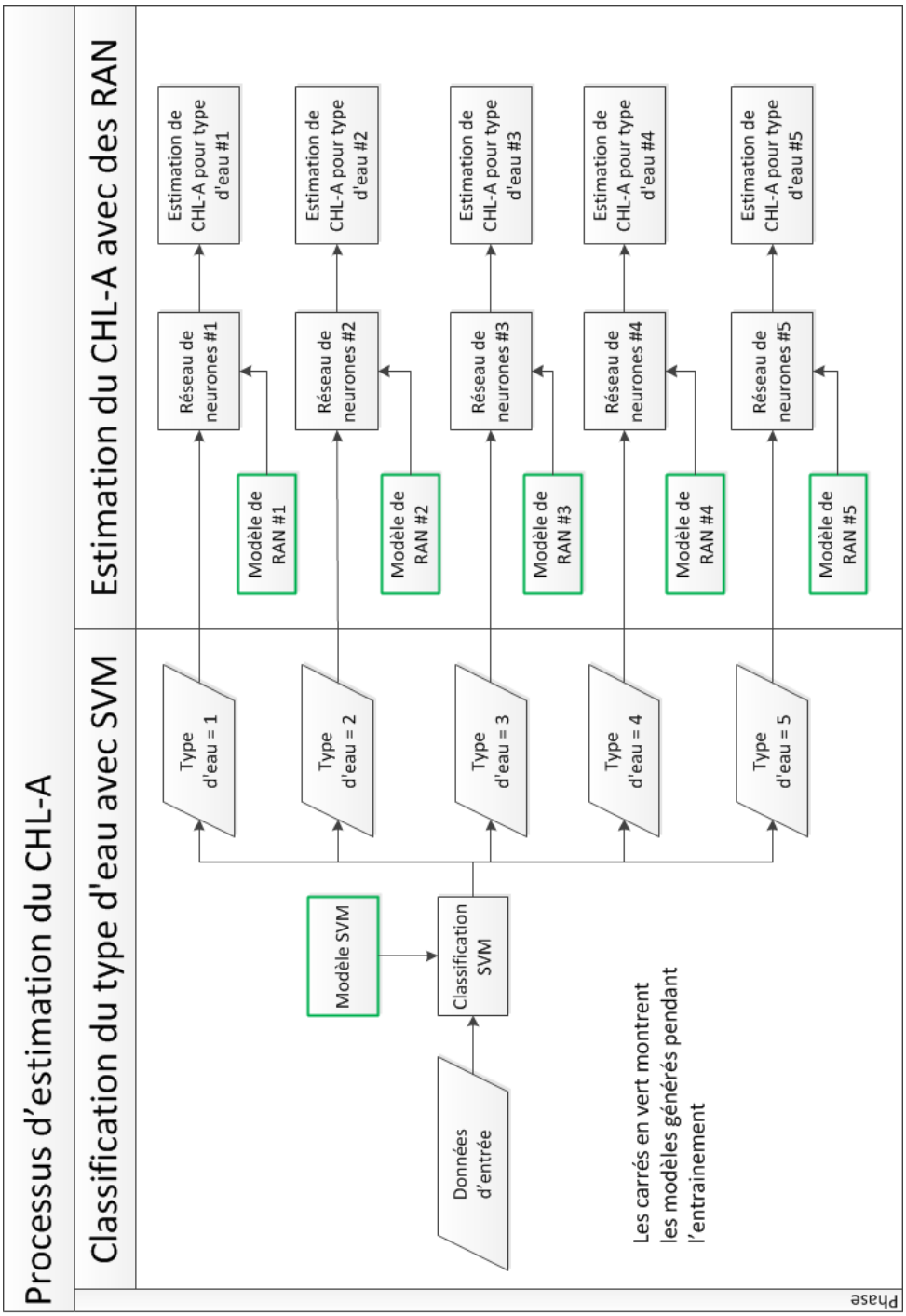

Figure 6 : Sommaire du processus d'estimation du CHL-A

3.4 Méthode d'évaluation des résultats

Pour évaluer nos résultats, nous allons prendre les valeurs du CHL-A estimées à partir du réseau de neurones et nous allons les comparer avec les prises de données in situ. La performance du modèle sera évaluée en utilisant la moyenne quadratique (« Root Mean Square Error », RMSE_log).

Il y a deux façons de définir le RMSE. La façon la plus conventionnelle est de le définir comme suit :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - t_i)^2} \qquad (10)$$

N est le nombre d'éléments dans l'ensemble de données d'entraînement ou validation; e_i est l'erreur entre les valeurs estimées a_i et les valeurs cueillies in situ t_i.

Cette façon de calculer le RMSE est plutôt commune dans la littérature. Cependant, en raison que la distribution de CHL-A est log-normale, les données in situ et les estimés doivent être transformés en logarithmes (base 10) avant de faire la comparaison (Pan et al. [40]). Ceci donne une formule différente pour le RMSE, que nous allons appeler RMSE_log :

$$RMSE\_log = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(a_i) - \log(t_i))^2} \qquad (11)$$

Afin d'être en ligne avec la littérature, nous allons utiliser le RMSE_log pour nos essais, tout au long du processus de recherche. Plus tard, et seulement avec le but de comparer nos résultats avec d'autres auteurs, nous allons utiliser le RMSE.

Dans quelques cas, l'estimation du CHL-A donne des résultats négatifs. Il ne peut cependant avoir une concentration négative de CHL-A; de plus, pour la formule du RMSE_log, log(0) n'est pas défini. Dans ces cas spéciaux, nous avons converti les résultats négatifs en 0.01. Cette valeur donne un taux d'erreur raisonnable sans pourtant affecter l'ensemble des erreurs.

3.5 La classification des eaux

Les caractéristiques de réflexion des eaux peuvent varier grandement dans un même lac. En effet, quelques sous-régions d'un lac peuvent être influencées par le phytoplancton, tandis que d'autres peuvent être affectées par des matières inorganiques en suspend ou même par la contamination. Une même aire peut changer ses propriétés d'un jour à l'autre [41, 42].

Alors, il est important de faire une classification de ces eaux avant d'essayer d'évaluer la concentration de CHL-A. Il faut avoir un bon pont de départ, c'est-à-dire un ensemble de données consistant et cohérent, avant de développer des nouveaux modèles (ou d'appliquer des modèles existants).

Il y a eu quelques approches pour cela. Les auteurs de [43] utilisent un arbre de décision assez rudimentaire qui compare les différences de magnitudes entre les bandes MERIS. Cet arbre est montré ci-dessous :

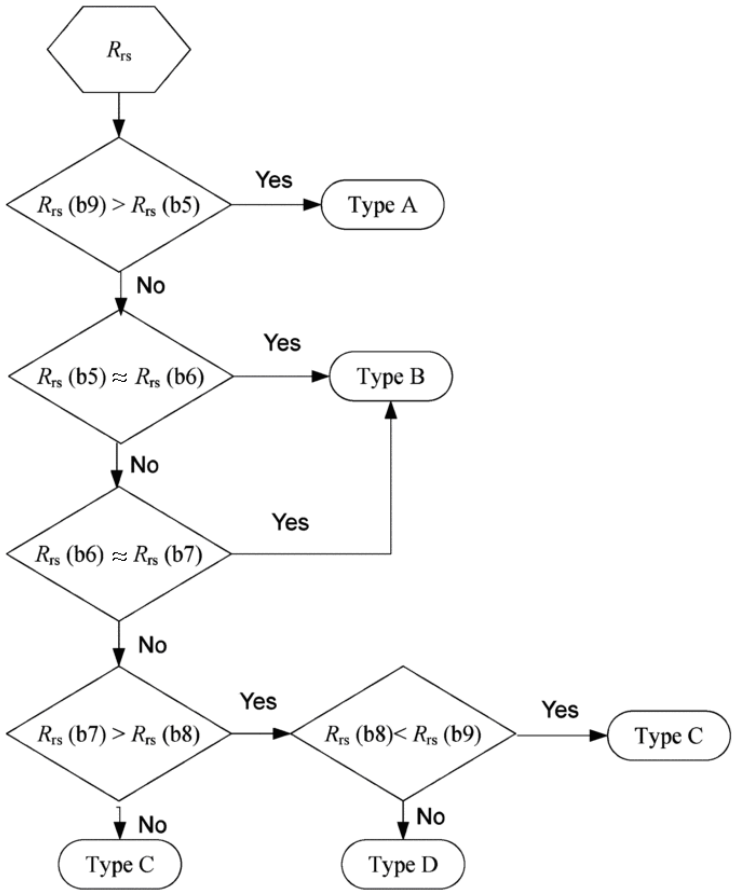

Figure 7: Arbre de décision pour la classification des eaux [43]

Cet arbre fait la classification des eaux dans quatre classes (A, B, C, D), en utilisant les bandes 5,…,9 de MERIS. Ces classes donnent les profils optiques suivants :

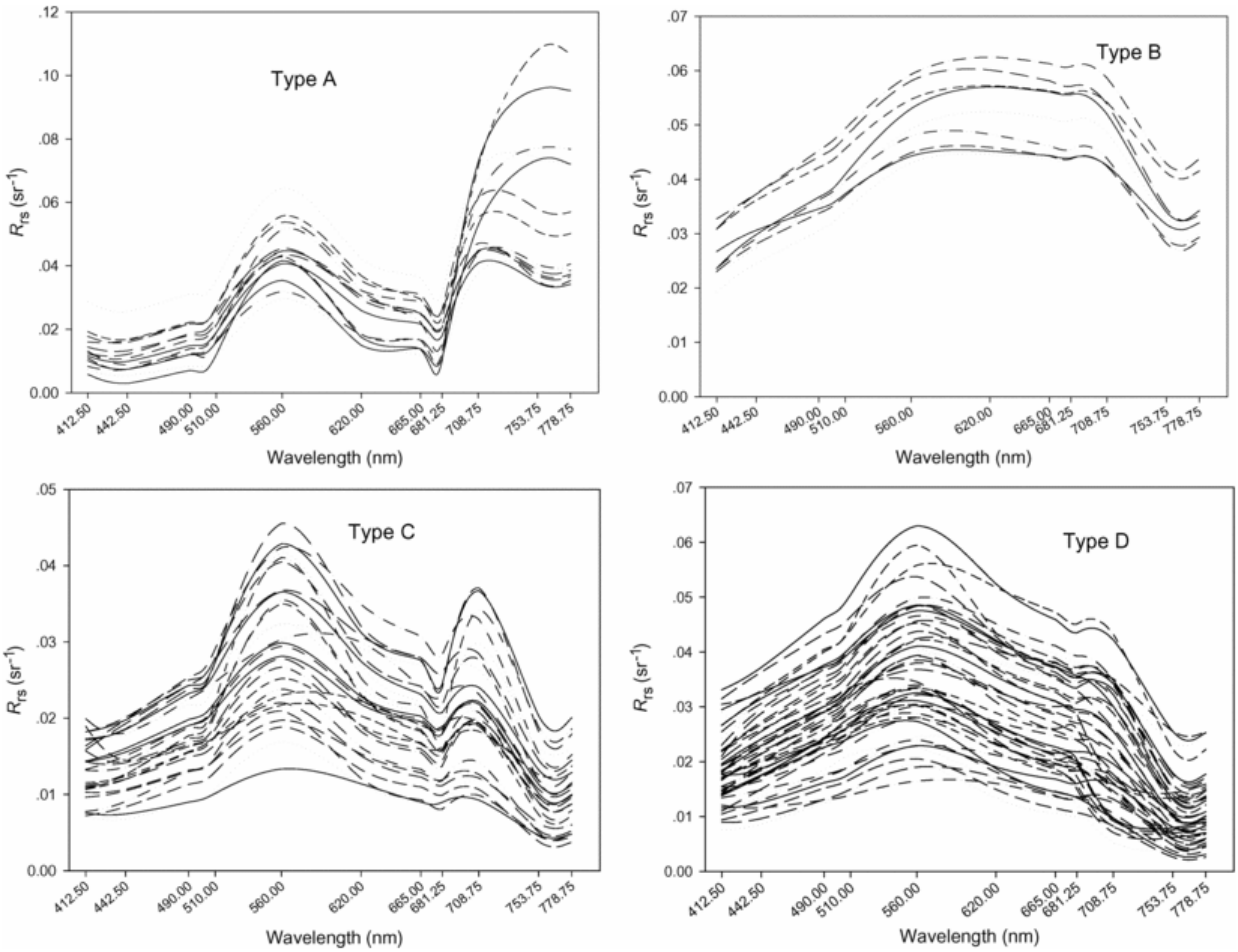

Figure 8: Profil des types d'eaux obtenues en utilisant l'arbre de décision montré dans la Figure 7 [43].

NB Chaque courbe correspond à un échantillon de données particulier, c.à.d. la valeur de réflectance de l'ensemble des bandes.

Les auteurs ont découvert que les types d'eaux A et B n'étaient pas bonnes pour établir la concentration de CHL-A, soit par un manque de rémission ou par une très haute concentration de sédiments. À cause de cela, les modèles d'évaluation de CHL-A furent développés pour les classes C et D seulement.

Ressom et al. [35] utilisent une méthode de *clustering* avec l'algorithme *fuzzy C-means*. Cet algorithme est conçu pour minimiser la fonction objective J_m définie comme suit :

$$J_m = \sum_{k=1}^{c} \sum_{i=1}^{N} (u_{ik})^m |x_i - v_k|^2 \qquad (12)$$

Où u_{ik} est le membre de la *i*-ème observation du *k*-ème *cluster*; $|x_i-v_k|$ est la norme Euclidienne entre les vecteurs x_i et v_k; m est le rapport de pondération qui peut être n'importe quel numéro réel plus grand que 1 (d'habitude établi à 2); *c* est le nombre de *clusters* et *N* est le nombre d'observations.

L'algorithme choisit des *clusters* qui ont une distance minimale entre les points des données et les centres des *clusters* (v_k). Les centres de *clusters* v_k sont ajustés d'une façon itérative jusqu'au moment que le critère d'optimisation soit satisfait.

Avec la méthode énoncée ci-dessus, les auteurs ont trouvé 7 *clusters*. Pour ce faire, ils ont utilisé des échantillons du NOMAD (« *NASA bio-Optical Marine Algorithm Dataset* ») et du IOCCG (« *Ocean Colour Coordinating Group* »). Les auteurs ne mentionnent pas quelles bandes furent utilisées, alors on assume qu'on travaille avec toutes les bandes. Les *clusters* sont montrés ci-dessous :

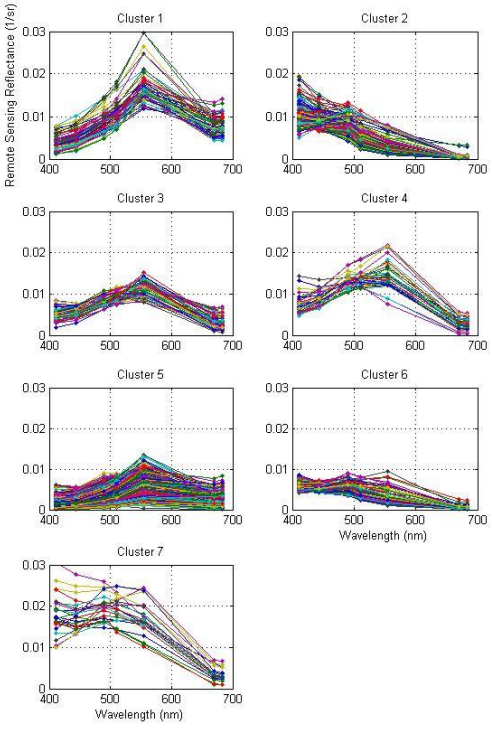

Figure 9: Clusters obtenues à partir des types d'eaux différentes [35]

Une fois les *clusters* créés, les auteurs ont construit un réseau des fonctions de base radiales (« *radial basis function network* »). La raison pour un tel réseau est d'être capable de déterminer l'appartenance d'un vecteur inconnu à un *cluster* spécifique. Ce vecteur est composé par l'ensemble de valeurs d'une courbe en particulier, c'est-à-dire qu'un vecteur peut être interprété comme une des courbes montrées dans la Figure 9. Autrement dit, les auteurs ont construit le réseau des fonctions de base radiales et donné les courbes comme entrée, afin de déterminer l'appartenance des courbes aux clusters.

3.5.1 La classification par la forme

Malgré le fait que les essais de classification de l'eau énoncés ont fait des pas dans la bonne direction, ils ont des problèmes à surmonter. Le premier modèle avec son arbre de décision très rudimentaire est susceptible aux erreurs; aussi, il n'est pas du tout flexible, car il n'est pas capable de s'adapter à des nouveaux types d'eau. Finalement, il fut développé pour MERIS – cependant, on veut un système dit générique et qui puisse s'adapter à différents capteurs / satellites. La deuxième méthode penche vers l'autre coté : elle n'est pas rudimentaire, mais plutôt trop complexe. Il nous faut beaucoup de préparation pour définir les *clusters* et pour faire la bonne classification. Elle nous donne 7 types d'eaux – ce qui nous poussera plus tard à créer 7 modèles différents pour l'estimation du CHL-A. Finalement, cette méthode fut développée pour des eaux océaniques.

Alors, nous allons nous concentrer dans le développement d'une méthode de classification avec les caractéristiques suivantes :

- Simple : Simple à calculer, simple à implanter.
- Flexible : Cette méthode peut être adaptée à plusieurs surfaces d'eau – elle n'est pas limitée à un environnement géographique particulier.
- Nombre de classes limité : On essaie de trouver entre 3-5 classes.

Le choix du nombre de classes est conditionné par plusieurs facteurs. Il nous faut un nombre de classes qui puisse nous permettre de bien classifier les types d'eaux existantes par rapport aux profils de réflexion. Bien que un nombre de clases réduit (3 par exemple) est plus simple du point de vue de l'implantation, nous avons remarqué que quelques types d'eau étaient mélangés dans une même classe, et que leur profil ainsi que la quantité de leurs échantillons dictaient qu'ils pourraient appartenir à une autre classe. De plus, le fait d'avoir regroupé plusieurs classes potentiellement différentes dans une seule avait plus tard un impact négatif dans l'estimation du CHL-A (une réduction de la précision). D'un autre côté, le fait d'avoir un nombre élevé de classes (plus que 5) n'améliorait pas la précision. De plus,

le nombre d'échantillons disponibles étant très limité, la quantité d'échantillons par classe diminuait considérablement.

En tenant compte de ces faits, nous avons trouvé qu'un nombre de classes égal à 5 est optimal pour l'estimation du CHL-A, tout dans un contexte de disponibilité limitée de données.

En faisant la révision soigneuse des données des autres auteurs ainsi que des données disponibles pour notre recherche, nous avons remarqué que les différents types d'eau (c.à.d. de classes) ont des courbes caractéristiques différentes; ces courbes se ressemblent entre elles dans leur forme générale, indépendamment de la magnitude acquise du signal. Les deux courbes montrées dans la Figure 10 sont basées sur les mêmes données et ont un profil similaire. Cependant, la magnitude absolue des données est plus grande dans la courbe verte, c'est-à-dire que la quantité de lumière incidente sur les senseurs du satellite est simplement plus élevée. Pour bien évaluer les profils des courbes, il nous faut alors un processus de normalisation.

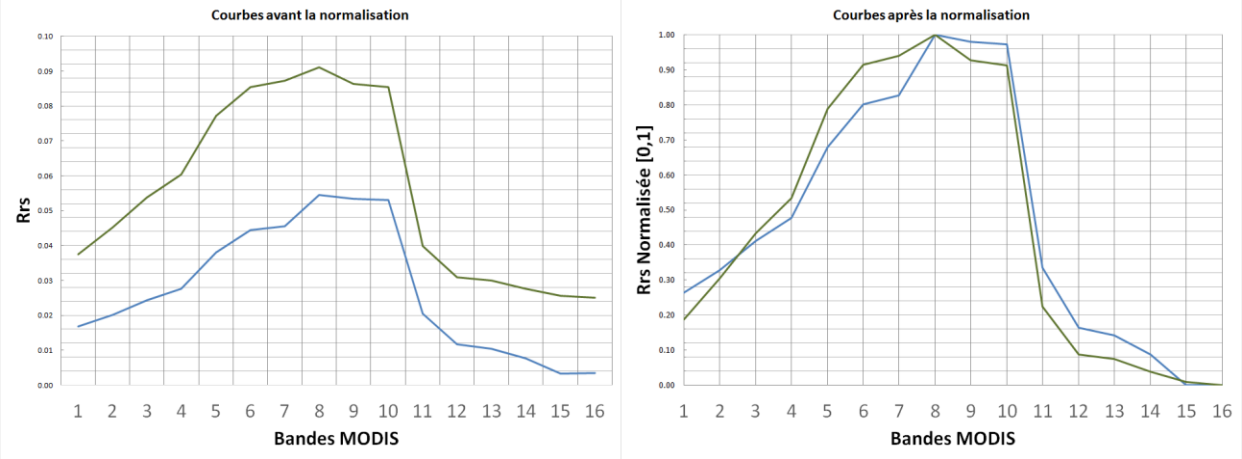

Figure 10: Courbes obtenues avant et après normalisation

Dans la Figure 10, nous pouvons voir deux courbes avant et après normalisation. L'axe des X montre chacune des bandes MODIS, tandis que l'axe des Y montre la magnitude du signal acquis (ce qui, pour notre exemple, est peu important).

Le but de la normalisation est d'approcher les courbes vers une étendue commune de caractéristiques, sans y perdre la relation entre les valeurs de chaque bande. Pour normaliser les courbes, nous utilisons la formule suivante [20]:

$$C' = \frac{C - C_{min}}{C_{max} - C_{min}} \quad (13)$$

C'est la valeur normalisée, C est la valeur réelle des données; C_{min} et C_{max} représentent les valeurs minimales et maximales du vecteur contentant les valeurs discrètes de C.

Le processus de normalisation montré ci-dessus nous permet d'avoir des courbes avec une échelle de magnitude similaire avec des valeurs dans un intervalle [0,1]. Une fois les courbes normalisées, on peut essayer de faire la classification par rapport aux formes caractéristiques. Dans la Figure 10, on peut voir dans les courbes après normalisation que les maximum et les minimums sont dans le même range, et ce rend plus simple la comparaison entre les courbes.

Il y a plusieurs façons de classifier par la forme de la courbe; on peut utiliser les valeurs absolues de chaque bande, calculer la différence entre chaque bande et ainsi de suite. Aussi, quelques points des courbes ne sont pas importants par rapport à l'estimation du CHL-A; il nous faut alors choisir le meilleur ensemble des valeurs pour éliminer le bruit et prendre les valeurs qui sont vraiment importantes pour le calcul du CHL-A.

Les méthodes de calcul tel que les valeurs absolues et les différentiels entre bandes présentent quelques inconvénients. Malgré le processus de normalisation, il peut y avoir des différences importantes entre des courbes appartenant à une même classe. Ce qui nous permet d'obtenir le meilleur résultat est plutôt une division pour établir la relation entre les bandes.

Étant donné qu'il y a 16 bandes, le problème qui se pose alors est de savoir quelles sont les bandes qu'il faut choisir. Pour ce faire, nous avons analysé les bandes *a priori* pour déterminer les types de courbes. Nous avons remarqué des caractéristiques distinctives de chaque type de courbe et nous avons pris les points de chaque courbe (c.-à-d. la bande) qui nous donnaient cette caractéristique particulière.

Pour extraire des caractéristiques, on utilise la relation entre deux bandes. En effet, le fait d'obtenir cette relation (c'est-à-dire une division) nous donne la pente entre deux bandes. Ladite pente est la caractéristique recherchée.

Nous avons extrait les caractéristiques pour identifier les types d'eaux suivantes :

Classe 1: Cette classe montre un petit triangle au-dessus, son sommet étant donné par la bande #8. Les caractéristiques utilisées pour la distinguer sont la relation entre les bandes 7 et 9 et la relation entre les bandes 8 et 9.

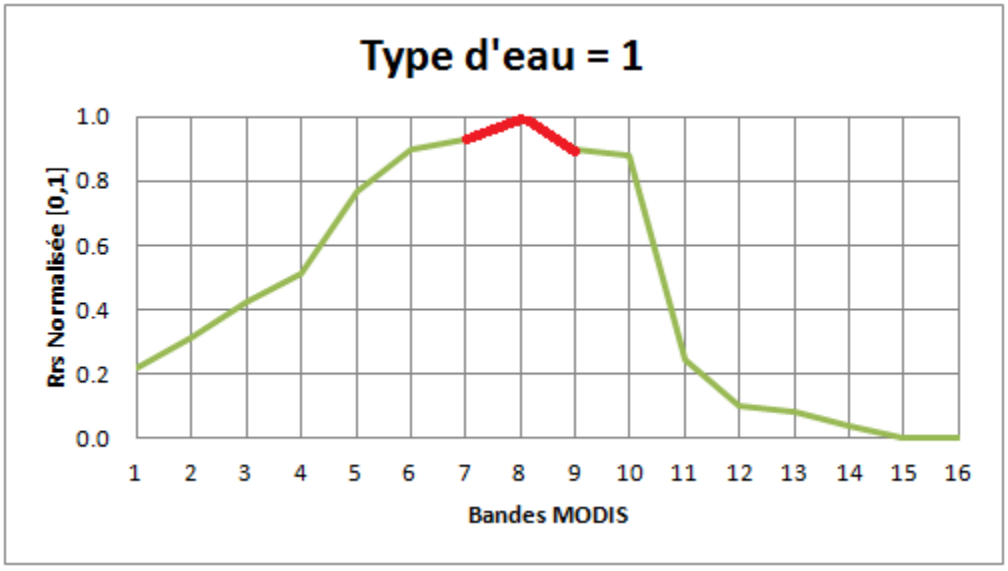

Figure 11: Types d'eaux - Classe 1

N.B. L'axe des X indique les bandes MODIS, tandis que le axe des Y indique la valeur normalisée de la bande (équation (13)).

Classe 2: Cette classe montre une forme de parallélogramme donnée par les bandes 4 à 7 et une pointe vers la fin dans la bande 11. Pour détecter cette courbe, nous proposons une relation de deuxième ordre F, donnée par les bandes 7/8 divisées par 10/11, c'est-à-dire :

$$F = \frac{R_{rs}(\lambda 7) * R_{rs}(\lambda 11)}{R_{rs}(\lambda 8) * R_{rs}(\lambda 10)} \qquad (14)$$

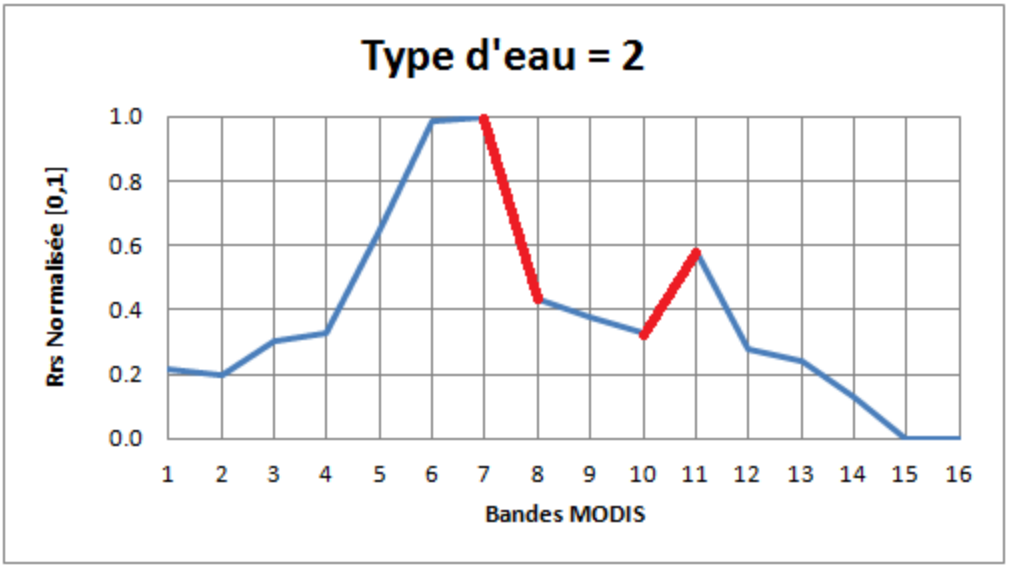

Figure 12: Types d'eaux - Classe 2

Classe 3: Cette classe montre la forme de parallélogramme donnée par les bandes 4 à 8 sans la pointe vers la fin. Pour la détecter, nous utilisons la relation entre les bandes 7 et 8 et la relation entre les bandes 6 et 4.

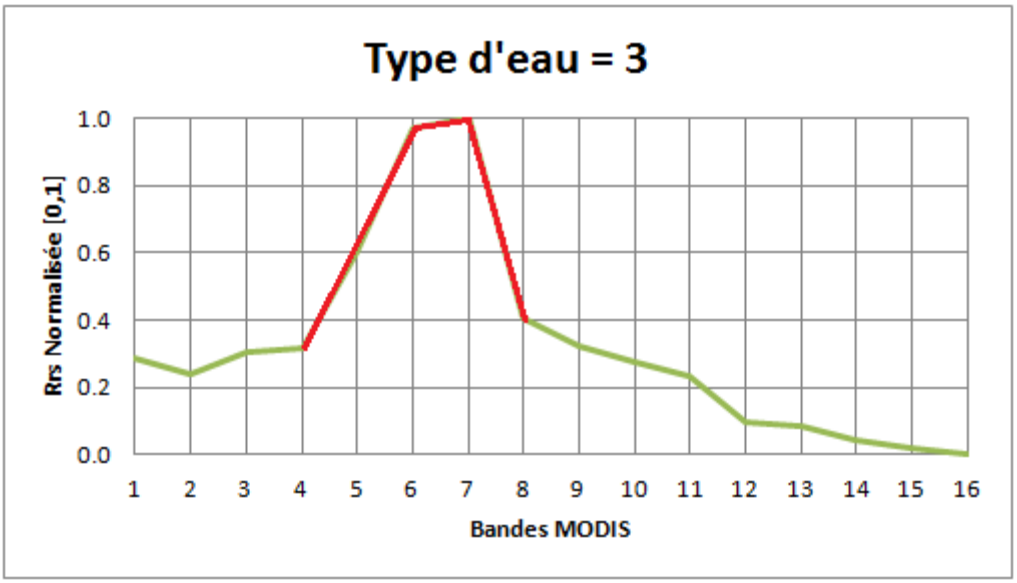

Figure 13: Types d'eaux - Classe 3

Classe 4: Cette classe est très similaire à la classe 3, mais la taille du parallélogramme est inférieure. Nous utilisons aussi la relation entre les bandes 7 et 8 pour la détecter.

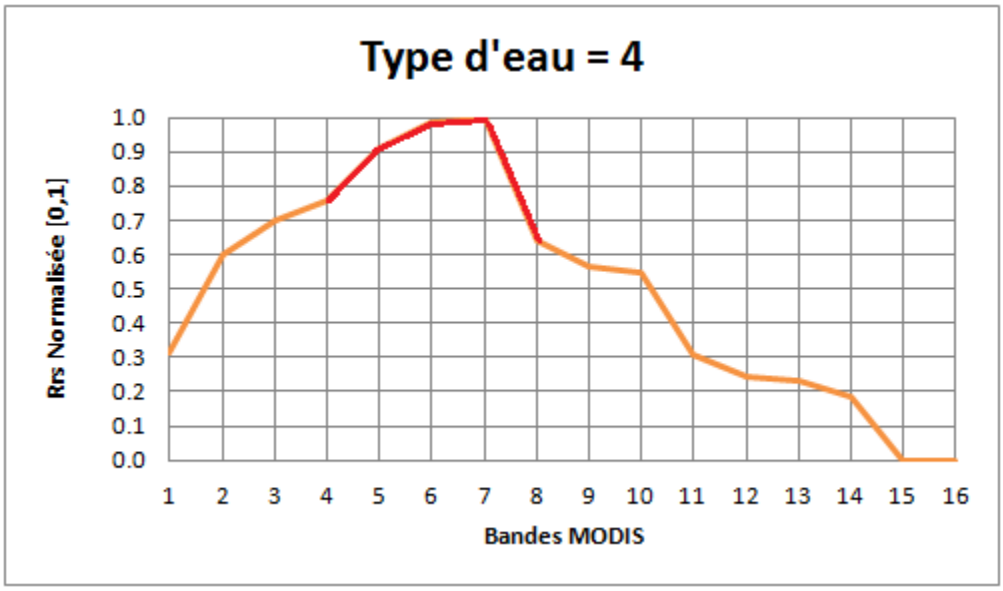

Figure 14: Types d'eaux - Classe 4

Classe 5: Cette classe est encore similaire à celle antérieure, mais la taille de la pointe est si petite que la courbe a l'air d'une surface presque plane. Nous utilisons la relation entre les bandes 7 et 8 pour la détecter.

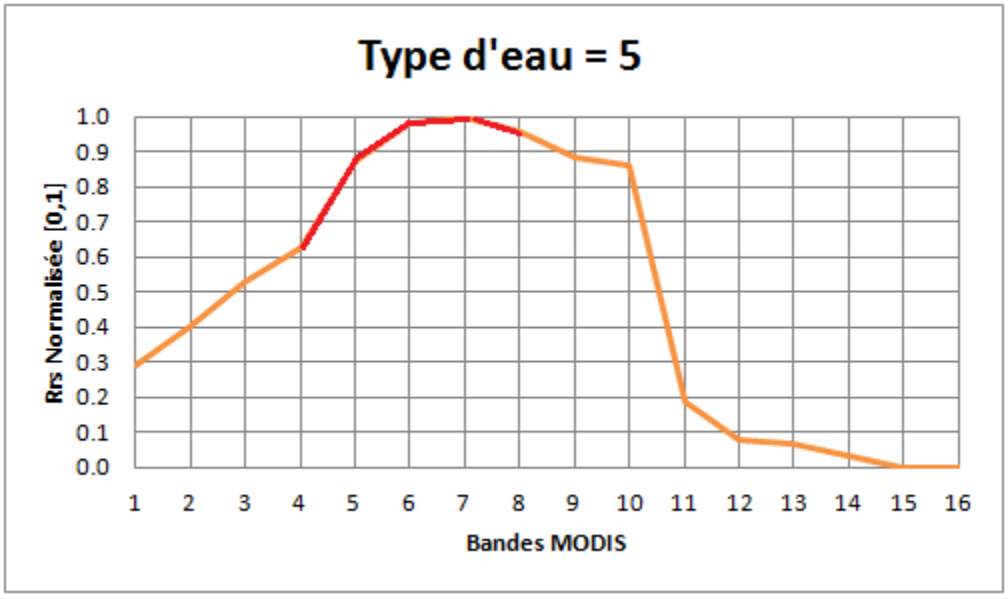

Figure 15: Types d'eaux - Classe 5

En tenant compte des caractéristiques trouvées dans les exemples précédents, nous avons calculé les relations suivantes :

$$
\begin{aligned}
C_1 &= \frac{R_{rs}(B7)}{R_{rs}(B8)} = \frac{R_{rs}(\lambda_{555})}{R_{rs}(\lambda_{645})} \\
C_2 &= \frac{R_{rs}(B6)}{R_{rs}(B4)} = \frac{R_{rs}(\lambda_{551})}{R_{rs}(\lambda_{488})} \\
C_3 &= \frac{R_{rs}(B8)}{R_{rs}(B9)} = \frac{R_{rs}(\lambda_{645})}{R_{rs}(\lambda_{667})} \\
C_4 &= \frac{R_{rs}(B7)}{R_{rs}(B9)} = \frac{R_{rs}(\lambda_{555})}{R_{rs}(\lambda_{667})} \\
C_5 &= \frac{R_{rs}(B7) * R_{rs}(\lambda 11)}{R_{rs}(B8) * R_{rs}(\lambda 10)} = \frac{R_{rs}(\lambda_{555}) * R_{rs}(\lambda_{748})}{R_{rs}(\lambda_{645}) * R_{rs}(\lambda_{678})}
\end{aligned}
\tag{15}
$$

Dans l'équation précédente, C_n représente la relation entre les bandes, R_{rs} est la valeur de réflectance et λ_{xyz} représente une bande centrée dans la fréquence xyz nanomètres, tandis que BX représente le numéro de bande MODIS (par exemple, B7 = λ_{555}).

3.5.2 Utilisation de SVM pour la classification des eaux

Une fois les caractéristiques distinctives obtenues, il nous faut établir un processus pour les regrouper ou les classifier. Nous avons décidé de nous éloigner des modèles paramétriques ou des arbres de décision [36] étant donné que cela n'offre pas beaucoup de flexibilité.

SVM, dans sa version originale, est un classificateur linéaire; il fait la classification en construisant un hyperplan optimal capable de séparer les données en deux catégories. On dit un hyperplan optimal quand la distance de l'hyperplan aux données est maximale (ce qu'on appelle la marge maximale). Les points les plus proches de l'hyperplan sont les vecteurs de support [44, 45]. La relation entre l'hyperplan et les vecteurs de support est que le produit scalaire de chaque vecteur de support et du vecteur normal à l'hyperplan donne une valeur soit égale à 1, soit égale à -1.

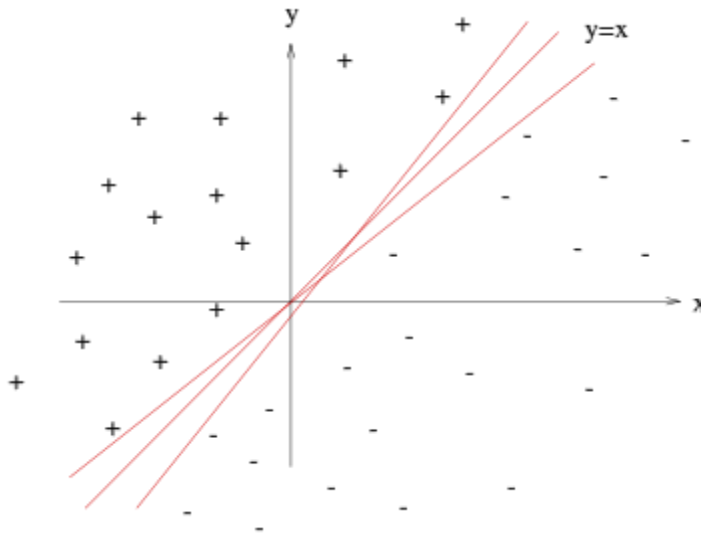

Figure 16: Exemple de plusieurs hyperplans possibles

Dans la Figure 16, nous avons le cas de plusieurs hyperplans possibles; chaque hyperplan établit une frontière entre les deux classes de données. Dans la Figure 17, nous avons l'exemple d'un hyperplan optimal : la distance entre l'hyperplan et les vecteurs de support est maximale. Un hyperplan (n'importe quel) est un ensemble de points x représenté par l'équation suivante :

$$w \bullet x + b = 0 \qquad (16)$$

où w est le vecteur normal à l'hyperplan et b est le déplacement de l'hyperplan par rapport au point d'origine et • est le produit scalaire.

Pour déterminer l'hyperplan optimal, nous devons d'abord trouver deux autres hyperplans de façon qu'ils soient capables de séparer les données et de maximiser leur distance. Ces hyperplans seront alors décrits par les équations :

$$w \bullet x + b = 1 \qquad (17)$$

Et

$$w \bullet x + b = -1 \qquad (18)$$

Nous savons que la distance entre les deux hyperplans est $\frac{2}{\|w\|}$, alors trouver l'hyperplan optimal devient un problème d'optimisation (minimiser ||w||).

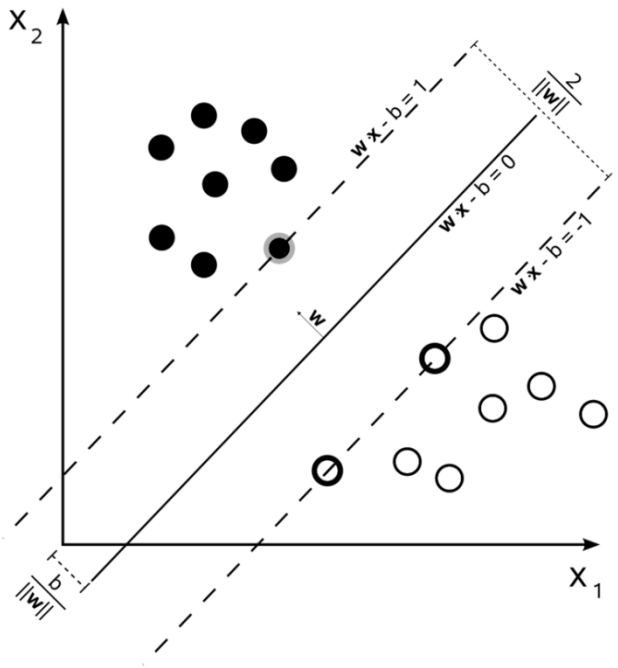

Figure 17: Exemple d'un hyperplan en utilisant la marge maximale

3.5.3 La classification non linéaire

Il arrive souvent que la frontière entre les deux ensembles de données ne soit pas linéaire. L'algorithme original de SVM était, en effet, un classificateur linéaire. Pour résoudre ce problème, on utilise une fonction de noyau non-linéaire (*non-linear kernel*) à la place du produit scalaire. Cela permet à l'algorithme de trouver l'hyperplan de marge maximale dans un espace de caractéristiques transformées. [37]

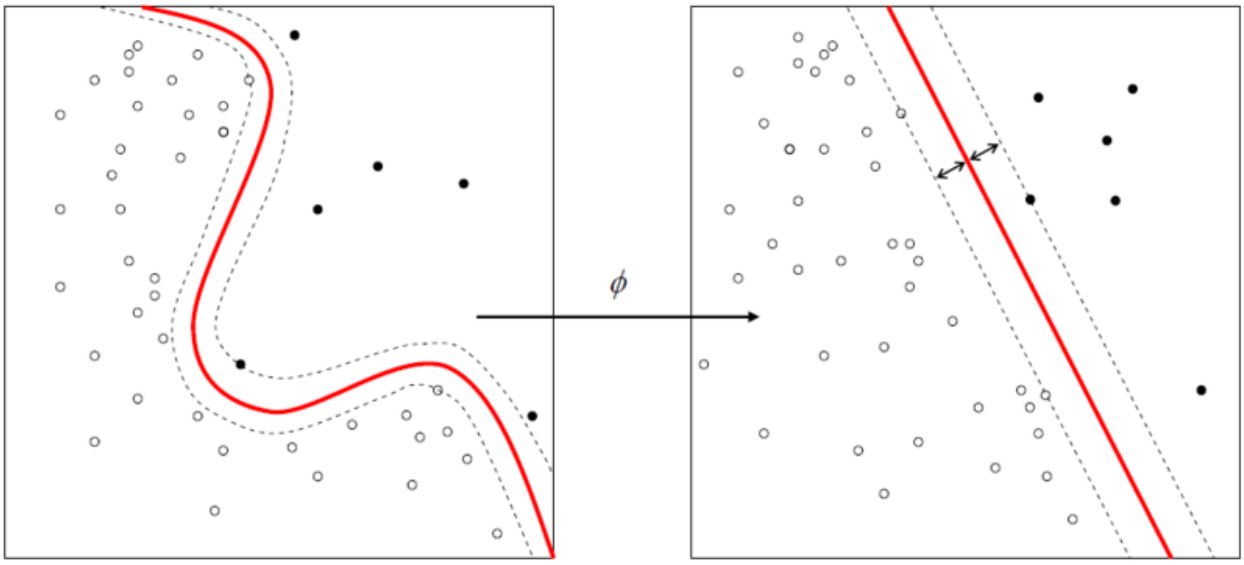

Figure 18: Exemple de classification non-linéaire et transformation de l'espace de caractéristiques.

NB dans la Figure 18 la fonction de noyau non-linéaire est représentée par ϕ.

3.5.4 La classification en plusieurs catégories

Par définition, SVM est un classificateur binaire, c'est-à-dire qu'il est capable de trouver seulement deux catégories. Dans notre cas, il nous faut plus de catégories pour classifier les types d'eaux. Pour ce faire, on utilise des extensions au SVM pour la classification multi-classe. L'approche la plus répandue est de réduire un seul problème multiclasses à plusieurs problèmes binaires. La méthode la plus commune est de construire un classificateur binaire qui fera la distinction entre une de nos classes et tout le reste (« *one-versus-all* »).

3.6 Les réseaux de neurones pour l'estimation du CHL-A

3.6.1 Fonctionnement des réseaux de neurones

Un réseau artificiel de neurones est composé par plusieurs couches; chaque couche est composée soit par les variables d'entrée originales, soit par des variables construites. L'architecture utilisée la plus souvent est une structure à trois couches :

1. La couche d'entrée, où les variables d'entrée sont mises.
2. La couche « cachée », composée d'un ensemble de variables construites
3. La couche de sortie, composée des réponses ou données de sortie.

La figure ci-dessus montre un exemple de réseau de neurones avec h nœuds par couche et l-2 couches cachées [46]:

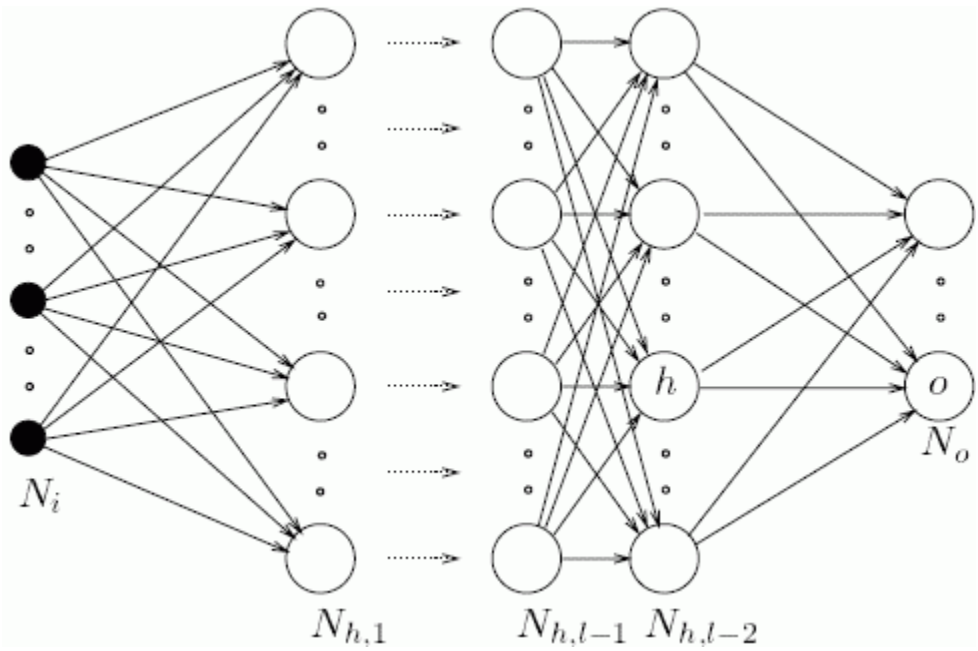

Figure 19: Architecture d'un réseau artificel de neurones

Chaque variable dans chaque couche est appelée un nœud (c.-à-d. un neurone artificiel). Un nœud est une unité élémentaire de traitement d'un réseau de neurones. Il est connecté à des sources d'information en entrée (les variables d'entrée ou d'autres nœuds) et renvoie une information en sortie qui sert comme entrée aux nœuds après lui.

On considère par exemple que x_i (1<=i<=p) sont les informations qui arrivent au nœud (p est le nombre d'éléments). Chacune de ces informations x_i sera plus ou moins valorisée par le biais d'un poids (un coefficient w_i lié à chaque information x_i). Alors, on peut considérer l'entrée du nœud comme la somme pondérée des produits de chaque entrée par son poids associé (x_i * w_i).

Il existe aussi un poids supplémentaire, appelé « coefficient de biais ». Ce poids supplémentaire sera associé à une fonction de transfert. Or, notre fonction d'information qui arrive au nœud peut être exprimée par:

$$\mathrm{d_u} = \left(\sum_{i=1}^{p} \mathrm{w_i} * \mathrm{x_i} \right) - \mathrm{w_0} \qquad (19)$$

C'est en fait la donnée d_u qui sera traitée par le nœud. Dans l'équation (19), *x* est le vecteur contenant les données, *w* est le vecteur contenant les coefficients (poids), w_0 est le coefficient de biais et *p* est le nombre total d'éléments contenus dans les vecteurs *x* et *w*.

La fonction d'activation est une fonction qui donne un nombre réel proche de 1 quand les « bonnes » informations d'entrée sont données et un réel proche de 0, le cas échéant. En conséquence, on définit la fonction d'activation *g* dans un intervalle réel [0,1]. La fonction d'activation peut être exprimée par :

$$\mathrm{a} = \mathrm{g}(\mathrm{d_u}) = g\left(\left(\sum_{i=1}^{p} \mathrm{w_i} * \mathrm{x_i} \right) - \mathrm{w_0} \right) \qquad (20)$$

Dans l'équation (20), *a* est la sortie de la fonction d'activation *g*; l'entrée de cette fonction est la donnée d_u obtenue d'après l'équation (19).

Il y a plusieurs types de fonctions d'activation, les plus utilisées sont :

- La fonction de Heaviside: il s'agit d'une fonction binaire, c'est-à-dire le résultat de cette fonction ne peut être que 0 ou 1.
- La fonction sigmoïde: il s'agit d'une fonction logarithmique, qui offre des valeurs intermédiaires (toujours dans [0,1]), et qui a l'avantage d'être dérivable.

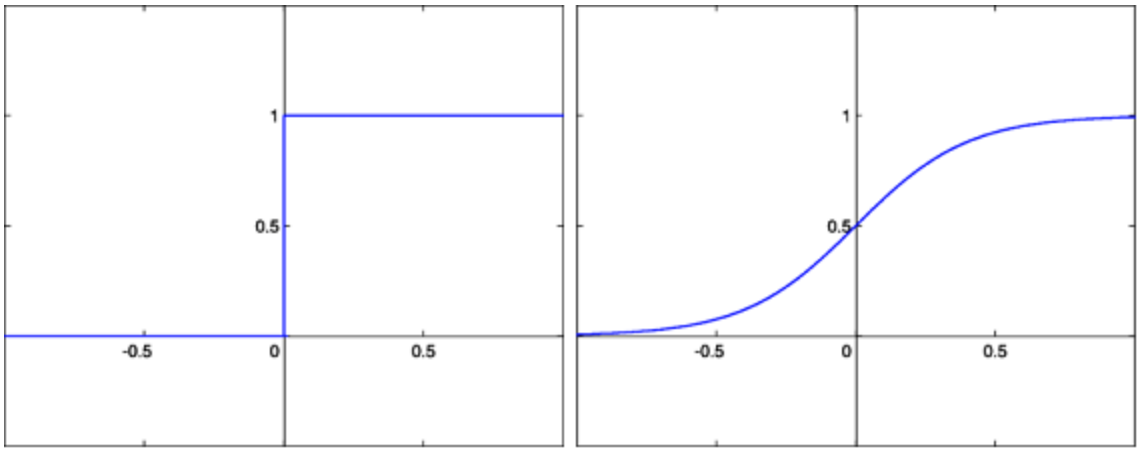

Figure 20: Graphes des fonctions Heaviside et sigmoïde

Il faut remarquer que les deux fonctions ont un seuil; dans les deux cas, le seuil est atteint quand X=0. La valeur du seuil de la fonction Heaviside est 1 tandis que celui de la fonction sigmoïde est en 1/2.

C'est-à-dire que notre seuil est atteint quand d_u=0.

$$d_u = \left(\sum_{i=1}^{p} w_i * x_i\right) - w_0$$

$$d_u = 0 \Leftrightarrow \left(\sum_{i=1}^{p} w_i * x_i\right) - w_0 = 0 \qquad (21)$$

$$\left(\sum_{i=1}^{p} w_i * x_i\right) = w_0$$

Alors, le seuil de la fonction d'activation est atteint quand la somme pondérée des informations d'entrée est égale au coefficient de biais (w_0). Quand $g(d_u)$ >= seuil = g(0), on dit que la neurone est active.

La fonction sigmoïde présente l'avantage d'être dérivable (ce qui va être utile par la suite) ainsi que de donner des valeurs intermédiaires (des réels compris entre 0 et 1) par opposition à la fonction de Heaviside qui, elle, renvoie soit 0 soit 1.

Dans le cas particulier de ce projet, l'implantation fut réalisée avec le logiciel MatLab. Ledit logiciel fournit la fonction de transfert 'tansig' par défaut. Cette fonction est utilisé par des autres chercheurs ([28, 29, 38]). La fonction « tansig » est très similaire à la fonction sigmoïde, la seule différence étant qu'elle est définie dans l'intervalle [-1,1]. Ceci explique l'obtention des estimations négatives de CHL-A.

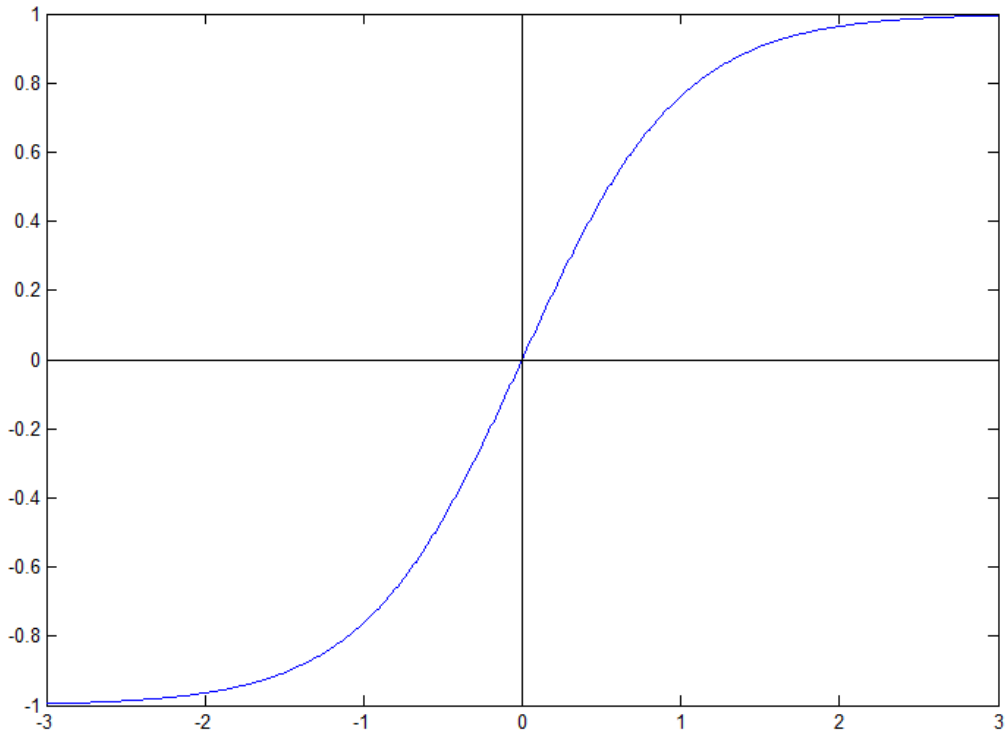

Figure 21: Représentation graphique de la fonction "tansig"

Il y a plusieurs architectures possibles pour un réseau artificiel de neurones, la plus populaire étant celle de la « Propagation en Arrière » (PA) (« *BackPropagation* »). Pour entraîner un réseau de neurones avec PA, il y a deux étapes à suivre : dans la première étape, les entrées sont propagées en avant pour calculer les sorties pour chaque nœud. Chaque sortie est comparée avec le résultat désiré, ce qui donne une valeur d'erreur pour chaque nœud. Dans la deuxième étape, chacune de ces erreurs est renvoyée en arrière (vers les nœuds d'origine) et les poids relatifs de chaque entrée sont corrigés. Ce deux étapes

sont répétées d'une façon itérative jusqu'au moment que la somme carré des erreurs ait une valeur acceptable.

3.6.2 Choix de données d'entrée

Un facteur important est le choix des données qui seront envoyées au réseau de neurones. La qualité de la sortie dépend toujours de la validité et le choix soigneux des données d'entrée. Nous avons les choix suivants :

1. Les relations des bandes de MODIS que nous avons aussi utilisées pour l'entraînement du SVM, tel que montré dans l'équation (15).
2. Données MODIS brutes; toutes les bandes, pas de normalisation.
3. Données MODIS normalisées; toutes les bandes.
4. Données MODIS normalisées, mais pas toutes les bandes – seulement les 7 bandes qui font partie des relations dans l'équation(15) seront incluses ici.

Pour établir l'ensemble de données qui va mieux performer, nous avons pris un échantillon de données appartenant à la classe d'eaux #4, car cet échantillon était le plus nombreux (79 enregistrements). Nous avons construit un RAN avec une seule couche cachée. La couche d'entrée et la couche cachée avaient le même nombre de neurones et cette quantité est établie par le nombre de données d'entré :

Table 3: Nombre de neurones selon le types de données utilisées.

#	Type de données	Nombre de neurones
1	Relations entre bandes - équation (15)	6
2	Données MODIS brutes; toutes les bandes, pas de normalisation.	16
3	Données MODIS normalisées; toutes les bandes.	16
4	Données MODIS normalisées, mais pas toutes les bandes – seulement les 7 bandes qui font partie des relations dans l'équation (15) seront incluses ici.	7

Le réseau de neurones fut configuré avec les fonctions 'tansig 'pour la couche cachée et une fonction de transfert linéaire pour la sortie. La fonction de transfert en arrière (BackPropagation) est 'trainlm' (Levenberg-Marquardt) [28, 29, 38] .

Les échantillons furent envoyés au réseau de neurones pour l'entraînement et les mêmes échantillons furent utilisés pour la validation. Nous avons obtenu les résultats suivants :

Table 4: Comparaison entre résultats pour les différents types de données

#	Type de données	RMSE_log
1	Relations entre bandes - l'équation (15)	1.5589
2	Données MODIS brutes; toutes les bandes, pas de normalisation.	0.3504
3	Données MODIS normalisées; toutes les bandes.	1.0492
4	Données MODIS normalisées, mais pas toutes les bandes – seulement les 7 bandes qui font partie des relations dans l'équation (15) seront incluses ici.	1.0007

Le RMSE_log minimal est ceci du type de données 2 (toutes les bandes, pas de normalisation). Nous allons quand même vérifier la qualité de ces données dans d'autres architectures (plus de couches, plus de neurones, etc.).

3.6.3 Architecture du réseau de neurones

Il faut davantage définir la structure du réseau de neurones, notamment le nombre de couches cachées et le nombre de neurones par couche. D'abord, une révision de la littérature s'impose. Hsun-Hsin et al. [47] utilisent trois couches (couche d'entrée, couche cachée, couche de sortie) avec six neurones, car ils ont six données d'entrée et une seule à la sortie. Ils ont donc fini avec une structure 6-6-1. Liu et al. [20] utilisent aussi trois couches, avec un nombre variable de neurones dans la couche cachée (entre 3 et 15). Les auteurs montrent un tableau avec le résultat de leurs expériences et ont déterminé que le nombre optimal de neurones est de 10. Wang et al. [38] utilisent trois couches, avec 19 neurones dans la couche cachée. Dans une étude similaire, Wang et al. [26] utilisent encore trois couches, mais avec 10 neurones. Ying-Ying et al. [28] utilisent 18, 20 et 25 neurones pour la couche cachée, sans pourtant justifier le choix du nombre de neurones.

Huajun et al. [30] font une étude très intéressante en faisant la comparaison entre des réseaux de trois et 4 couches cachées, avec un nombre variable de neurones dans chacune. Le nombre de neurones d'entrée est toujours cinq (car ils ont cinq données à l'entrée), et le nombre de neurones de sortie est toujours un (une seule donnée à la sortie – l'estimation du CHL-A). Dans le cas d'une seule couche cachée, ils ont utilisé entre 3 et 7 neurones. Dans le cas de deux couches cachées, la première couche a toujours quatre neurones, tandis que la deuxième a entre 1 et 4 neurones. Les auteurs ont choisi une structure 5-4-1-1.

Une évaluation intéressante est celle de Song et al. [48]. Dans cet article, les auteurs soutiennent qu'une seule couche cachée améliore l'efficacité computationnelle et que le nombre de neurones dans la couche cachée est le même que dans la couche d'entrée, en considérant « qu'un nombre élevé de neurones dans la couche cachée peuvent impacter négativement le temps de calcul, tandis qu'un nombre réduit de neurones [c.-à-d. inférieur à ceux de la couche d'entrée] peut provoquer une oscillation dans le processus d'apprentissage ».

Une approche similaire apparaît dans l'article de Sorayya et al. [25]. Ce qui est plus intéressant dans cet article est le choix du nombre de couches cachées. Les auteurs n'utilisent qu'une couche et affirment qu'il a été prouvé par [49],[50] et [51] qu'une seule couche cachée peut approximer n'importe quelle fonction continue. De plus, les auteurs citent [52] et [53] selon qui « l'utilisation de deux couches cachées est justifiée seulement dans les application les plus ésotériques (sic) ».

En tenant compte des considérations présentées par d'autres auteurs, nous avons décidé de faire les essais suivants :

- Utiliser une seule couche cachée avec un nombre de variable de neurones (entre 1 et 20).
- Utiliser deux couches cachées : la première avec un nombre de neurones entre 1 et 20, la deuxième entre 1 et 5.

La couche d'entrée aura le même nombre de neurones que les données d'entrée, et la couche de sortie aura toujours un seul neurone, car il n'y a qu'une seule donnée de sortie (l'estimation du CHL-A).

Pour le test, nous avons utilisé le même échantillon que dans l'essai précédent (79 enregistrements). Le réseau de neurones fut configuré avec les fonctions 'tansig' pour les couches cachées et une fonction de transfert linéaire pour la sortie. La fonction de transfert en arrière (BackPropagation) est 'trainlm' (Levenberg-Marquardt).

Pour une seule couche cachée, voici les résultats qui se trouvent dans la table ci-dessous. On peut remarquer que pour une seule couche cachée, le minimum RMSE_log est obtenu quand le nombre de neurones est de 2 et les données d'entrée sont toutes les bandes, sans normalisation.

Table 5: Résultats pour une seule couche cachée, tous les types de données

# de neurones	#1 Relations entre les bandes	#2 Données sans normalisation, toutes les bandes	#3 Données normalisées, toutes les bandes	#4 Données normalisées, 7 bandes
1	0.2978	0.5079	0.3065	0.2847
2	2.1940	**0.2690**	0.2722	0.5966
3	0.4605	0.3100	0.6292	0.7660
4	0.3504	0.2782	0.5706	0.2732
5	0.7500	0.2805	0.8205	0.2918
6	0.8472	0.5628	1.1371	0.3633
7	0.4358	0.4796	1.0262	0.4275
8	0.5406	0.6595	1.8200	1.2136
9	0.4426	1.3776	1.1046	0.3467
10	0.6968	0.9345	1.0255	0.5986
11	0.6065	0.7349	1.1096	0.7566
12	0.7838	0.6807	0.8899	0.9786
13	0.8250	0.7183	0.6036	0.9581
14	0.7209	0.6713	1.7535	0.8790
15	1.3857	0.7102	2.3451	0.5194
16	0.5422	0.3504	1.0492	1.2789
17	0.3854	0.5206	1.3790	1.1124
18	1.1016	0.3115	1.0117	0.7098
19	0.8756	0.9335	0.8452	0.9884
20	1.0959	1.0159	1.0042	1.3349

Ensuite, nous avons procédé à faire l'essai avec deux couches cachées (la première avec un nombre de neurones entre 1 et 20, la deuxième entre 1 et 5). Nous avons effectué les essais pour tous les types de données (normalisées, sans normaliser, etc.). Dans les tables qui suivent, les colonnes montrent le nombre de neurones pour la deuxième couche cachée, tandis que les rangées montrent le nombre de neurones pour la première couche cachée.

Table 6: Résultats pour deux couches cachées, types de données 1 (relations entre les bandes).

Nombre de neurones dans la première couche cachée	Nombre de neurones dans la deuxième couche cachée					
		1	2	3	4	5
	1	0.2956	0.3163	0.3689	0.4897	0.6588
	2	0.3822	0.3623	0.2659	0.3724	0.3294
	3	0.3436	0.2849	0.2894	0.5690	0.4835
	4	0.2908	0.2960	0.4229	0.2812	0.2982
	5	0.3584	0.2706	0.2961	0.4515	0.3040
	6	0.3528	0.2596	0.7370	0.2642	0.5718
	7	0.2798	0.3830	0.6623	0.2773	0.3062
	8	0.3759	0.4416	1.4852	0.2943	0.9283
	9	0.3419	0.2629	0.2675	0.6939	0.4374
	10	1.0861	0.3346	0.2915	0.5345	0.3009
	11	1.8814	0.2717	0.3913	0.2524	0.2366
	12	0.2497	0.2559	0.3219	0.3137	0.4566
	13	0.2712	0.3023	0.2972	0.4003	0.3076
	14	0.7771	0.3678	0.4313	0.6563	0.9684
	15	0.3530	0.3527	0.4185	0.2554	0.3797
	16	0.3250	0.9022	0.9208	0.4082	0.2795
	17	0.2665	0.3648	0.3019	0.2765	0.5921
	18	0.3677	0.3062	0.2769	0.3308	0.5538
	19	0.3730	0.3685	0.5575	0.5366	0.4609
	20	0.3581	0.3491	0.3094	0.9420	0.3511

Table 7: Résultats pour deux couches cachées, types de données 2 (données sans normalisation, toutes les bandes)

		Nombre de neurones dans la deuxième couche cachée				
Nombre de neurones dans la première couche cachée		1	2	3	4	5
	1	0.3317	0.3187	0.3070	0.2571	0.4942
	2	0.3931	0.3472	0.3186	0.2497	0.3480
	3	0.3023	0.3322	0.3141	0.2993	0.8494
	4	0.3001	0.3388	0.2651	0.3543	0.2820
	5	0.3536	0.3565	0.4684	0.3268	0.3140
	6	0.2495	0.3949	0.2951	0.6372	0.3455
	7	0.3423	0.3689	0.2786	0.3281	1.1832
	8	0.3299	0.3566	0.2815	0.5143	0.3117
	9	0.3168	0.3025	0.3023	0.6551	0.3377
	10	0.3587	0.3292	0.2884	0.3770	0.6838
	11	0.3534	0.3511	0.4103	0.3071	0.2874
	12	0.3818	0.3229	0.3412	0.5852	0.6317
	13	0.3083	0.3440	0.3240	1.4220	1.1821
	14	0.3966	0.2864	0.6648	0.5946	0.7771
	15	0.2772	0.2732	0.3567	0.4189	0.3638
	16	0.2641	0.3312	0.4458	0.5955	0.2569
	17	0.2802	0.2656	0.2347	0.2978	0.3391
	18	0.3204	0.3479	0.2956	0.6526	0.3106
	19	0.3102	0.3032	0.3056	0.7117	0.6324
	20	0.3422	0.2527	0.2972	0.2850	1.2218

Table 8: Résultats pour deux couches cachées, types de données 3 (données normalisées, toutes les bandes)

Nombre de neurones dans la première couche cachée	Nombre de neurones dans la deuxième couche cachée					
		1	2	3	4	5
	1	0.4474	0.4036	0.3306	0.6060	0.3620
	2	0.2719	0.2533	0.3938	0.2617	0.2944
	3	0.3778	0.2918	0.2842	0.2557	0.7413
	4	0.3451	0.2609	0.2824	0.3401	0.2703
	5	0.3645	0.2503	0.3368	1.3830	0.2683
	6	0.3290	0.3398	0.2774	0.2813	0.6422
	7	0.3897	0.2330	0.5815	0.9367	0.5087
	8	0.3241	0.6010	0.5455	0.2859	0.5758
	9	0.3146	0.3912	0.3113	0.2920	0.2717
	10	0.3252	0.7297	0.2825	0.2490	0.3465
	11	0.3704	0.2819	0.2424	0.7797	0.7465
	12	0.3460	0.5516	0.4102	0.2647	0.4465
	13	0.4263	0.3301	0.2959	0.3414	1.0627
	14	0.3777	0.3106	0.2842	0.9511	0.8447
	15	0.2884	0.3060	0.5624	0.2885	0.2863
	16	0.3321	2.6031	0.2810	0.2557	0.5632
	17	0.2967	0.3810	1.4452	0.4582	0.9834
	18	0.3148	0.3045	0.2438	0.4641	0.3369
	19	0.3175	0.5504	0.4448	0.3133	0.3759
	20	0.2873	0.3844	0.2823	0.4360	0.3485

Table 9: Résultats pour deux couches cachées, types de données 4 (données normalisées, 7 bandes)

Nombre de neurones dans la première couche cachée	Nombre de neurones dans la deuxième couche cachée					
		1	2	3	4	5
	1	0.3855	0.3489	0.5014	0.3467	0.6510
	2	0.3129	0.3202	0.3531	0.2752	0.4326
	3	0.3461	0.3391	0.4015	0.4817	0.7735
	4	0.2840	0.3731	0.3115	0.3845	0.8459
	5	0.3038	0.3334	0.3504	0.3039	0.8590
	6	0.3822	0.3000	0.2850	0.7393	0.3417
	7	0.2796	0.3371	0.7457	0.2699	0.3592
	8	0.3822	0.3105	0.3236	0.4151	1.2510
	9	0.4898	0.2759	0.2618	0.9093	0.3184
	10	0.3655	0.3279	0.4085	0.3158	0.3204
	11	0.3598	0.3361	0.3103	0.5409	1.6464
	12	0.2795	1.0607	0.5453	0.6438	1.6075
	13	0.3540	1.2726	0.2658	0.7234	0.6433
	14	0.2726	0.8884	0.3874	1.2252	0.4033
	15	0.6774	0.2653	0.5715	0.3200	1.0081
	16	0.3334	0.3000	0.5381	0.5912	0.6867
	17	0.3595	0.9425	0.3692	0.3553	1.4358
	18	0.3822	0.2950	0.3616	0.4280	1.4734
	19	0.3948	0.3147	0.4274	0.3616	0.3247
	20	0.3276	0.3310	0.3024	1.5699	0.2313

On peut voir que le nombre optimal est de 20 neurones dans la première couche cachée et de 5 neurones dans la deuxième couche cachée, pour le type de données 4 (données normalisées, seulement les 7 bandes utilisées pour calculer les relations). Aussi, une comparaison entre les deux architectures s'impose :

Table 10: Comparaison entre les résultats obtenus par les différentes architectures de réseaux de neurones.

Architecture	Type de données	RMSE_log
1 Couche Cachée, 9 neurones	#4	0.2690
2 Couches cachées, 11-5 neurones	#1	0.2366
2 Couches cachées, 17-3 neurones	#2	0.2347
2 Couches cachées, 7-2 neurones	#3	0.2330
2 Couches cachées, 20-5 neurones	**#4**	**0.2313**

Il faut toujours remarquer que les résultats précédents sont obtenus à partir d'un seul type d'eau, et avec le but d'établir la meilleure architecture du réseau de neurones.

On peut voir que les résultats des deux dernières architectures (7-2 et 20-5) sont très proches entre eux, avec une complexité réduite dans le cas de l'architecture 7-2. Cependant, le but de la recherche n'est pas de trouver la meilleur méthode d'estimation du CHL-A, sinon de démontrer qu'on l'on obtenir des meilleurs résultats en faisant la classification des eaux avant le processus d'estimation. Il se peut, effectivement, qu'on puisse trouver des architectures avec une performance supérieure. Par conséquence, nous avons choisi un réseau de neurones avec une structure 20-5. Cette architecture sera utilisée pour les essais d'estimation du CHL-A.

Chapitre 4 - Résultats

4.1 Types d'eaux identifiés

D'abord, après avoir effectué le prétraitement de données (c.-à-d. la normalisation des données) nous avons identifié plusieurs types d'eaux – toujours par rapport à leurs courbes caractéristiques. Nous avons identifié les types d'eaux suivantes :

Classe 1: Cette classe montre un petit triangle au-dessus, son sommet étant donné par la bande #8. Les caractéristiques utilisées pour la distinguer sont la relation entre les bandes 7 et 9 et la relation entre les bandes 8 et 9.

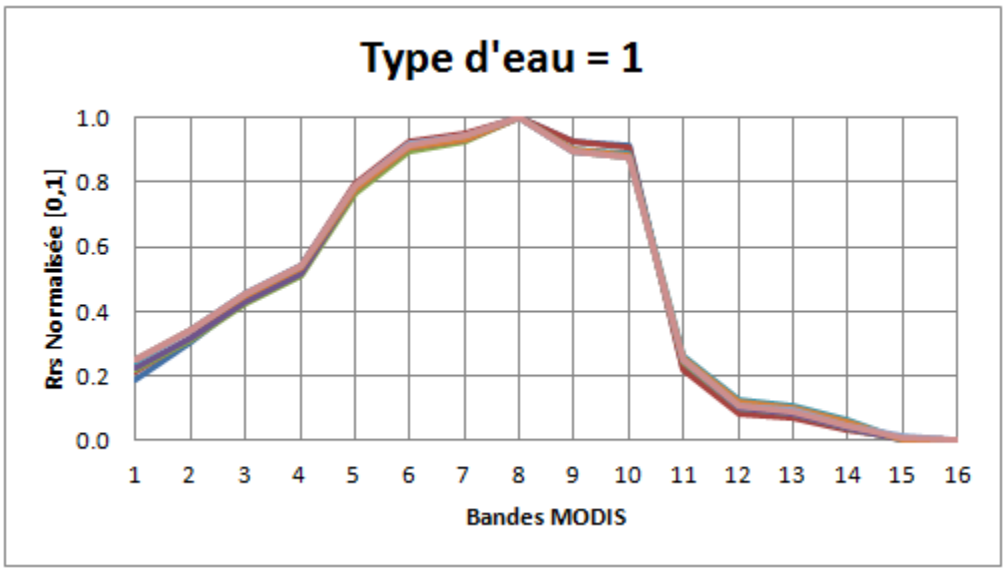

Figure 22: Types d'eaux - Classe 1

N.B. L'axe des X indique les bandes MODIS, tandis que le axe des Y (non visible) indique la valeur normalisée de la bande (équation (13)). Le but des figures montrant les courbes est de présenter la forme de la courbe, la valeur de l'axe Y étant peu importante.

Classe 2: Cette classe montre une forme de parallélogramme donnée par les bandes 4 à 7 et une pointe vers la fin dans la bande 11. Pour détecter cette courbe, nous proposons une relation de deuxième ordre F, donnée par les bandes 7/8 divisées par 10/11, c'est-à-dire :

$$F = \frac{R_{rs}(\lambda 7) * R_{rs}(\lambda 11)}{R_{rs}(\lambda 8) * R_{rs}(\lambda 10)} \quad (22)$$

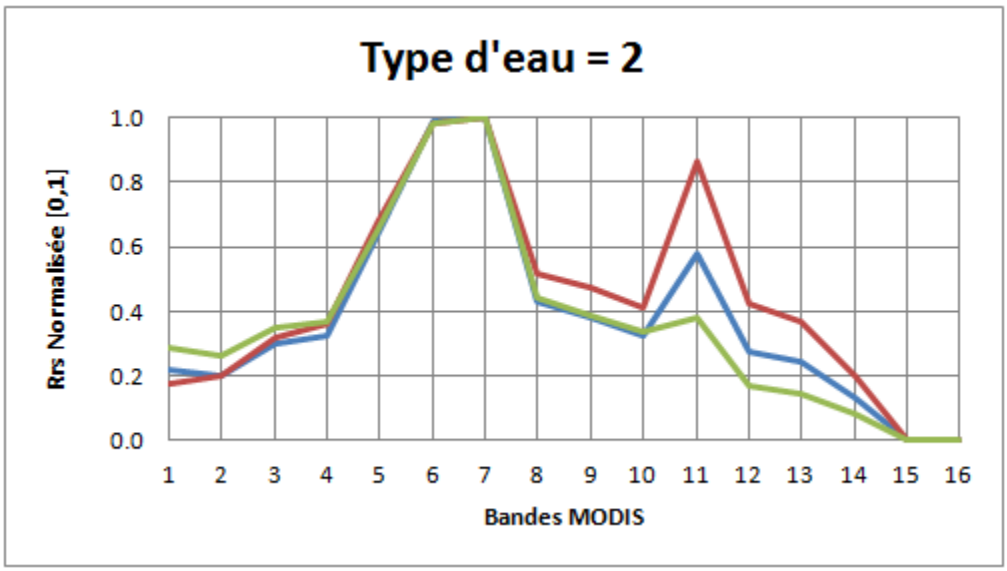

Figure 23: Types d'eaux - Classe 2

Classe 3: Cette classe montre la forme de parallélogramme donnée par les bandes 4 à 8 sans la pointe vers la fin. Pour la détecter, nous utilisons la relation entre les bandes 7 et 8 et la relation entre les bandes 6 et 4.

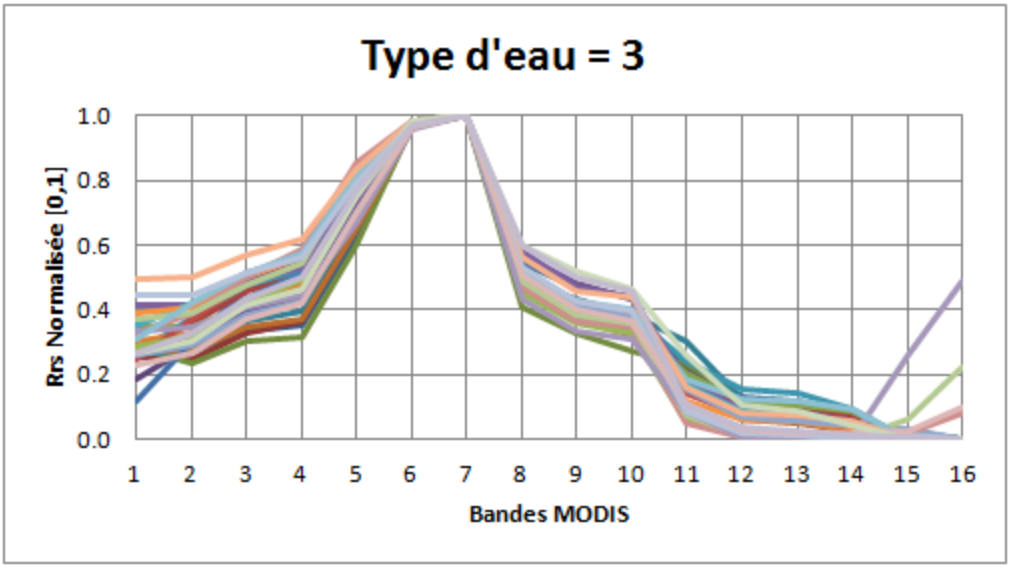

Figure 24: Types d'eaux - Classe 3

Classe 4: Cette classe est très similaire à la classe 3, mais la taille du parallélogramme est inférieure. Nous utilisons aussi la relation entre les bandes 7 et 8 pour la détecter.

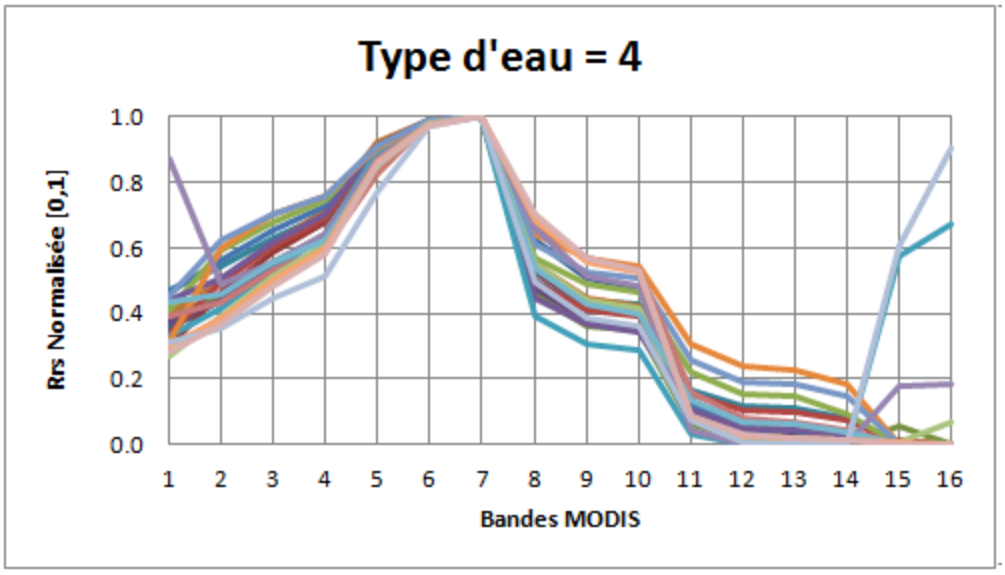

Figure 25: Types d'eaux - Classe 4

Classe 5: Cette classe est encore similaire à celle antérieure, mais la taille de la pointe est si petite que la courbe a l'air d'une surface presque plane. Nous utilisons la relation entre les bandes 7 et 8 pour la détecter.

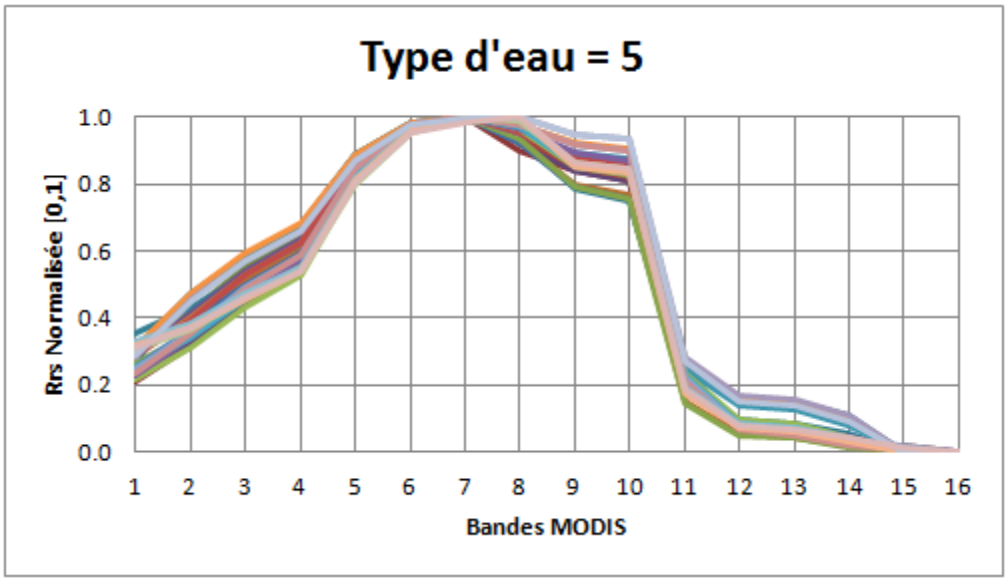

Figure 26: Types d'eaux - Classe 5

4.2 Le processus de classification

Nous avons fait une classification manuelle de tous les types d'eaux et nous avons mis des étiquettes dans chaque type. Ensuite, nous avons procédé à faire la classification en utilisant SVM. Nous avons divisé la source des données en deux : les données d'août 2002 jusqu'à août 2003 furent utilisées pour l'entraînement du SVM tandis que les données de juin 2004 à août 2004 furent utilisées pour la validation du modèle.

L'ensemble d'entraînement consistait à 106 éléments tandis que l'ensemble de validation avait 38 éléments. Les caractéristiques suivantes furent utilisées pour l'entraînement SVM :

- $R_{rs}(\lambda 7) / R_{rs}(\lambda 8)$: relation entre les bandes 555nm / 645nm.
- $R_{rs}(\lambda 6) / R_{rs}(\lambda 4)$: relation entre les bandes 551nm / 488nm.
- $R_{rs}(\lambda 8) / R_{rs}(\lambda 9)$: relation entre les bandes 645nm / 667nm.
- $R_{rs}(\lambda 7) / R_{rs}(\lambda 9)$: relation entre les bandes 555nm / 667nm.
- $R_{rs}(\lambda 7) * R_{rs}(\lambda 11) / R_{rs}(\lambda 8) * R_{rs}(\lambda 10)$: relation de deuxième ordre (555/645) / (678/748)

La classification fut faite en utilisant MATLAB et la librairie LIBSVM [54]. Nous avons choisi cette librairie par le fait qu'elle supporte la classification SVM multi classe. Nous avons utilisé un noyau polynomial K du type :

$$K = (\gamma * u' * v + coef)^d \qquad (23)$$

N.B. le premier paramètre ϒ c'est un gamma, pas un y grecque.

Nous avons utilisé une valeur de ϒ=0.2 (c.-à-d. 1/nombre de caractéristiques), un degré *d*=4 et une valeur de coefficient coef=5.

Avec les valeurs mentionnées ci-dessus, nous avons obtenu une précision de 97.17% au moment de la classification des données d'entraînement et une précision de 92.10% quand on faisait la classification des données de validation. Ces résultats sont très encourageants; cependant, nous croyons qu'ils peuvent être améliorés en ajustant les caractéristiques envoyées au classificateur.

4.3 L'utilisation des réseaux de neurones pour l'estimation de CHL-A

Pour faire l'estimation du CHL-A, nous avons fait les opérations suivantes

- Nous avons classifié les types d'eau, obtenant 5 types d'eaux différentes.
- Nous avons intentionnellement exclu le type d'eau 2, car il n'y avait que 3 échantillons.
- Nous avons entraîné un réseau de neurones pour chaque type d'eau (quatre réseaux en total). L'architecture du réseau étant 7-20-5-1, les types de données envoyées en entrée étant les bandes utilisées pour le calcul par la forme, données normalisées.
- Après l'entraînement et la simulation, nous avons obtenu les résultats suivants :

Table 11: Résultats pour une deux couches cachées, types de données 4 (données normalisées, 7 bandes)

Type d'eau	RMSE_log	# d'échantillons
Type d'eau #1	0.3695	11
Type d'eau #2	s/o	3
Type d'eau #3	0.1930	26
Type d'eau #4	0.2313	79
Type d'eau #5	0.1205	29
Tous les types d'eau	0.2309	145

Nous avons décidé de faire aussi une comparaison entre les estimés du CHL-A sans y faire une séparation entre les types d'eau. Nous avons utilisé la même architecture pour le réseau de neurones et les mêmes données d'entrée. Voici les résultats :

Table 12: Comparaison entre résultats avec et sans classification des eaux

	RMSE_log
RMSE_log global, avec la classification des eaux	0.2309
RMSE_log sans la classification des eaux	0.5386

Dans cette estimation, si l'on isole les estimés pour le type d'eau #4 (Qui, avec la classification des eaux, donne un RMSE_log de 0.2313), son RMSE_log est de 0.3772; c'est-à-dire que même dans les conditions les plus défavorables, la classification des eaux AVANT l'estimation du CHL-A fonctionne mieux qu'une estimation directe. ***Ceci est la preuve de l'effectivité de notre approche.***

Nous avons aussi comparé notre approche avec d'autres méthodes présentes dans la littérature, tel que montré dans la table ci-dessous :

Table 13: Comparaison entre différents méthodes

Modèle	RMSE	RMSE_log	Référence
Notre méthode	9.8518	0.2309	Cette mémoire
OC2v4		0.4490	O'Reilly et al. 1998 [14]
OC4v4		0.4830	O'Reilly et al. 2000 [13]
OC2v4_D'Ortenzio		0.2890	D'Ortenzio et al. 2002 (cité par Pan et al. 2010 [40]).
OC2_TP		0.2450	Pan et al. 2010 [40]
OC4_TP		0.2450	Pan et al. 2010[40]
OC3M		0.4000	O'Reilly et al. 2000[13]
OC3M_TP		0.2880	Pan et al. 2010[40]
RAN*	21.3668		Huajun et al. 2010 [30]
RAN		0.29 (clear water) 0.37 (turbid water)	Slabakova et al. 2011 [22]
RAN*	14.1035		Liu et al. 2009 [20]

* Ces méthodes n'utilisent pas la télédétection, mais plutôt des paramètres tels que l'oxygène dissous, azote ammoniacal, pH, etc.

Chapitre 5 - Conclusion

Dans le cadre de ce mémoire, nous proposons un nouveau modèle pour l'estimation du Chlorophylle-A axé sur le principe fondamental que, dans une même surface, l'eau peut se présenter avec des caractéristiques différentes. Pour chacun de ces « états » de l'eau, nous utilisons un modèle différent d'évaluation.

La nouveauté de notre approche consiste à faire une classification des eaux en fonction de sa courbe spectrale caractéristique. Plus précisément, nous essayons de faire une classification par la forme de la courbe en prenant des caractéristiques saillantes de la dite courbe. Plusieurs caractéristiques sont évaluées afin d'obtenir le meilleur ensemble de données pour la classification. Cette classification est achevée par des méthodes d'apprentissage statistique, plus précisément, SVM.

Un fois les types d'eau classifiés, nous faisons une estimation du niveau du CHL-A en utilisant des réseaux artificiels de neurones (RANs). Cette méthode d'approximation non-linéaire nous a permis d'obtenir des meilleurs résultats dans certains types d'eau, notamment les types d'eau 4 et 5; Ceci est justifié par la quantité d'échantillons présents dans chaque classe.

Globalement, nous avons obtenu de meilleur résultats que comparés aux résultats sans classification de l'eau a priori et des résultats légèrement supérieurs à ceux mentionnés dans la littérature dans des situations de la classification de l'eau par un arbre décisionnel ainsi que ceux obtenus par l'application de méthodes paramétriques. Les résultats présentés dans ce mémoire démontrent la viabilité de notre approche, avec un RMSE_log de 0.2309, légèrement supérieur à RMSE_log de 0.2880 trouvé dans la littérature [40].

Nos tests ont été réalisés avec les données du lac Winnipeg – un lac d'eaux peu profondes situé aux latitudes septentrionales. Pour généraliser nos résultats, on devrait effectuer des tests avec plus de données et dans d'autres environnements.

Chapitre 6 - Bibliographie

[1] Z. Xiaoshen, L. Wei, L. Wenling, and W. Jing, "Methods of Chlorophyll Concentration Retrieval," in *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, 2008, pp. 76-80.

[2] Centre canadien de télédétection and Ressources naturelles Canada. (June 21, 2013). *Notions fondamentales de télédétection*. Available: http://www.rncan.gc.ca/sciences-terre/limite-geographique/teledetection/fondements/1100

[3] C. Simand. (2007, 24 Juin 2013). *La radiographie II. Qu'est-ce qu'un rayon X ?* Available: http://culturesciences.chimie.ens.fr/content/la-radiographie-ii-quest-ce-quun-rayon-x-comment-en-produire-quel-mecanisme-permet-dobtenir-une-radiographie-1197

[4] C. R. McClain, G. C. Feldman, and S. B. Hooker, "An overview of the SeaWiFS project and strategies for producing a climate research quality global ocean bio-optical time series," *Deep Sea Research Part II: Topical Studies in Oceanography,* vol. 51, pp. 5-42, 1// 2004.

[5] Y. Dingtian, P. Delu, Z. Xiaoyu, B. Yan, H. Xianqiang, W. Difeng*, et al.*, "Detection of algal bloom with in situ and MODIS in Lake TaiHu, China," presented at the Proceedings of the SPIE., 2005.

[6] S. Wouthuyzen, C. K. Tan, J. Ishizaka, T. Son, V. Ransi, S. Tarigan*, et al.*, "Monitoring of Algal Blooms and Massive Fish Kill in the Jakarta Bay, Indonesia using Satellite Imageries," *Proceedings of The First Joint PI symposium of ALOS Data Nodes for ALOS Science Program,* 2007.

[7] J. H. W. Lee, I. J. Hodgkiss, K. T. M. Wong, and I. H. Y. Lam, "Real time observations of coastal algal blooms by an early warning system," *Estuarine, Coastal and Shelf Science,* vol. 65, pp. 172-190, 10// 2005.

[8] European Space Agency (ESA). (2013, June 7th, 2013). *MERIS Design*.

[9] National Aeronautics and Space Administration (NASA). (2013, June 7, 2013). *MODIS Web*.

[10] A. Morel and L. Prieur, "Analysis of variations in ocean color," *Limnol. Oceanogr,* vol. 22, pp. 709-722, Jul. 1977.

[11] W. J. Moses, A. A. Gitelson, S. Berdnikov, and V. Povazhnyy, "Satellite Estimation of Chlorophyll-a Concentration Using the Red and NIR Bands of MERIS – The Azov Sea Case Study," *Geoscience and Remote Sensing Letters, IEEE,* vol. 6, pp. 845-849, October 2009 2009.

[12] K. Carder, Fchen, J. Cannizzaro, J. Campbel, and B. Mitchel, "Performance of the MODIS semi-analytical ocean color algorithm for chlorophyll-a," *Advanced Space Research,* vol. 33, pp. 1152-1150, Jan. 2004.

[13] J. E. O'Reilly, S. Maritorena, D. Siegel, M. C. O'brien, D. Toole, B. G. Mitchell*, et al.*, "Ocean color chlorophyll a algorithms for SeaWiFS, OC2, and OC4: Version 4," in *Hooker S. B. and E. R. Firestone (Eds.), SeaWiFS Postlaunch Technical Report Series, SeaWiFS Post¬launch Calibration and Validation Analyses* vol. 11, ed. Greenbelt, Maryland: NASA Goddard Space Flight Center, 1-51., 2000.

[14] J. E. O'Reilly, S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver*, et al.*, "color chlorophyll algorithms for SeaWiFS," *J. Geophys. Res,* vol. 103, pp. 24937- 24953, 1998.

[15] J. Gower, S. King, G. Borstad, and L. Brown, "Detection of intense plankton blooms using the 709nm band of the MERIS imaging spectrometer," *International Journal of Remote Sensing,* vol. 26, pp. 2005-2012, 2005.

[16] T. T. Wynne, R. P. Stumpf, M. C. Tomlinson, R. A. Warner, P. A. Tester, J. Dyble*, et al.*, "Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes," *International Journal of Remote Sensing,* vol. 29, pp. 3665-3672, 2008.

[17] T. T. Wynne, R. P. Stumpf, M. C. Tomlinson, and J. Dyble, "Characterizing a cyanobacterial bloom in Western Lake Erie using satellite imagery and meteorological data," *Limnol. Oceanogr,* vol. 55, pp. 2025-2036, 2010.

[18] C. Binding, T. Greenberg, B. Bukata, S. Watson, S. Rastin, and J. Gould. (2012, June 24, 2013). *Monitoring algal blooms using the MERIS Maximum Chlorophyll Index*. Available: http://www.ssec.wisc.edu/meetings/ciw/Workshop_Presentations/Wednesday_6_20_2012/2_Algorithm_Approaches/ignite_algorithm_application/8_Binding_NASA%20workshop.pdf

[19] J. Gower and S. King, "New results from a global survey using MERIS MCI," I. o. O. Sciences, Ed., ed. Sidney, BC, Canada.: Fisheries and Oceans Canada, , 2008.

[20] J. Liu, Y. Zhang, and X. Qian, "Modeling Chlorophyll-A in Taihu Lake with Machine Learning Models," in *Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on*, 2009, pp. 1-6.

[21] P. Cipollini, G. Corsini, M. Diani, and R. Grasso, "Retrieval of sea water optically active parameters from hyperspectral data by means of generalized radial basis function neural networks," *IEEE Transactions on Geoscience and Remote Sensing, ,* vol. 39, pp. 1508-1524, 2001.

[22] V. Slabakova, N. Slabakova, O. Hristova, and B. Dzhurova, "Assessment of MERIS ocean color products using in situ data collected in the Northwestern Black Sea," in *Recent Advances in Space Technologies (RAST), 2011 5th International Conference on*, 2011, pp. 132-136.

[23] J. Chen, M. Zhang, T. Cui, and Z. Wen, "A Review of Some Important Technical Problems in Respect of Satellite Remote Sensing of Chlorophyll-a Concentration in Coastal Waters," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of,* vol. 6, pp. 2275-2289, 2013.

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

[25] M. Sorayya, "A Comparison between Neural Network Based and Fuzzy Logic Models for Chlorophll-a Estimation," in *Second International Conference on Computer Engineering and Applications (ICCEA)*, 2010, pp. 340-343.

[26] X. Wang, J. Dai, Z. Liu, X. Zhao, S. Dong, Z. Zhao*, et al.*, "The lake water bloom intelligent prediction method and water quality remote monitoring system," in *Natural Computation (ICNC), 2010 Sixth International Conference on*, 2010, pp. 3443-3446.

[27] Y. Chou-Ping, H. Chi-Ying, W. Yu-Min, and H. Wen-Ping, "Using artificial neural network for reservoir water quality analysis in Taiwan," in *Consumer Electronics, Communications and Networks (CECNet), 2011 International Conference on*, 2011, pp. 1883-1886.

[28] Z. Ying-ying, A. M. Vorontcov, S. Ji-chang, L. Yan, and H. Guang-li, "Rolling prediction of single water quality parameter based on neural network," in *Natural Computation (ICNC), 2012 Eighth International Conference on*, 2012, pp. 350-353.

[29] L. Xiuli, H. Honghui, D. Ming, and Q. Zhanhui, "Chlorophyll-a predicting based on artificial neural network for marine cage fish farming area in dapeng cove in Daya Bay, South China Sea," in *Natural Computation (ICNC), 2012 Eighth International Conference on*, 2012, pp. 203-206.

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

[31] L. Jin-Suo, H. Ting-Lin, and W. Chun-yan, "Data Mining on Source Water Quality (Tianjin, China) for Forecasting Algae Bloom Based on Artificial Neural Network (ANN)," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, 2009, pp. 191-195.

[32] W. Liu, X. Zheng, X. Ao, and J. Wang, "Research on Inversion Methods of Chlorophyll Concentrations in Bohai Sea," in *Computer Science and Information Engineering, 2009 WRI World Congress on*, 2009, pp. 135-138.

[33] W. Tai-Sheng, T. Chih-Hung, C. Li, and T. Yu-Chu, "Applying Artificial Neural Networks and Remote Sensing to Estimate Chlorophyll-a Concentration in Water Body," in *Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium on*, 2008, pp. 540-544.

[34] H. M. C. Ribeiro, A. C. Almeida, B. R. P. Rocha, and A. V. Krusche, "Water Quality Monitoring in Large Reservoirs Using Remote Sensing and Neural Networks," *Latin America Transactions, IEEE (Revista IEEE America Latina),* vol. 6, pp. 419-423, 2008.

[35] H. W. Ressom, K. Turner, and M. T. Musavi, "Estimation of Ocean Water Chlorophyll-a Concentration Using Fuzzy C-means Clustering and Artificial Neural Networks," presented at the International Joint Conference on Neural Networks, Vancouver, BC, 2006.

[36] Y. Li, Q. Wang, C. Wu, S. Zhao, X. Xu, Y. Wang*, et al.*, "Estimation of Chlorophyll a Concentration Using the NIR/Red Bands procedure in inland turbid water," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 50, pp. 988-997, March 2012.

[37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," presented at the Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA, 1992.

[38] X.-y. Wang, J.-p. Xu, Z.-w. Liu, J. Dai, and S.-p. Zhu, "An Intelligent System on Water Quality Remote Monitor and Water Bloom Prediction," in *Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on*, 2010, pp. 521-524.

[39] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*: Springer-Verlag New York, Inc., 2006.

[40] Y. Pan, D. Tang, and D. Weng, "Evaluation of the SeaWiFS and MODIS Chlorophyll a Algorithms Used for the Northern South China Sea during the Summer Season," *Terrestrial, Atmospheric, and Oceanic Sciences,* vol. 21, pp. 997-1005, December 2010 2010.

[41] D. Y. Sun, Y. M. Li, C. C. Le, S. Q. Gong, L. Wu, and C. C. Huang, "Scattering characteristics of Taihu Lake and its relationship models with suspended particle concentration," *Environmental Science,* vol. 28, pp. 2688-2694, December 2007.

[42] A. Morel, B. Gentili, M. Chami, and J. Ras, "Bio-optical properties of high chlorophyll Case 1 waters and of yellow-substance-dominated Case 2 waters," *Deep-Sea Res. Part 1—Oceanogr. Res. Papers,* vol. 53, pp. 1439-1459, September 2006.

[43] Y. Li, Q. Wang, C. Wu, S. Zhao, X. Xu, Y. Wang*, et al.*, "Estimation of Chlorophyll a Concentration Using NIR/Red Bands of MERIS and Classification Procedure in Inland Turbid Water," *IEEE Transactions on Geoscience and Remote Sensing* vol. 50, pp. 988-997, 2012.

[44] P. Harrington, *Machine Learning in Action*. Greenwich, CT, USA: Manning Publications Co, 2012.

[45] E. S. Olivas, J. D. M. Guerrero, M. M. Sober, J. R. M. Benedito, and A. J. S. Lopez, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques* Hershey, PA: Information Science Reference, 2009.

[46] M. Firat. (2007, 12 juin 2013). *Image Recognition with Neural Networks*. Available: http://www.codeproject.com/Articles/19323/Image-Recognition-with-Neural-Networks

[47] H. Hsun-Hsin, C. Li, K. Chang-Huan, Y. Hui-Chung, and W. Tai-Sheng, "Applying Multi-temporal Satellite Imageries to Estimate Chlorophyll-a Concentration in Feitsui Reservoir Using ANNs," in *Artificial Intelligence, 2009. JCAI '09. International Joint Conference on*, 2009, pp. 345-348.

[48] K. Song, L. Li, S. Li, L. Tedesco, H. Duan, Z. Li*, et al.*, "Using Partial Least Squares-Artificial Neural Network for Inversion of Inland Water Chlorophyll-a," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. PP, pp. 1-1, 2013.

[49] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks 0893-6080,* vol. 2, pp. 359-366, 1989.
[50] G. Cybendo, *Approximations by superpositions of a sigmoidal function*. NY: Springer-Verlag, 1989.
[51] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks,* vol. 2, pp. 183-192, // 1989.
[52] J. de Villiers and E. Barnard, "Backpropagation neural nets with one and two hidden layers," *Neural Networks, IEEE Transactions on,* vol. 4, pp. 136-141, 1993.
[53] T. Masters, *Practical Neural Network Recipes in C++*: Academic Press, 1993.
[54] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, pp. 27:1--27:27, 2011.